Universidad de Ciencias Económicas y Sociales (UCES)

Maestría en Investigación de Mercados, Medios y Opinión (MIMMO)

Propuesta de un procedimiento superador que en un momento específico determine el éxito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina

Autor: Lic. Ramón Enrique de Windt Carbuccia

Tutor: Mg. Gonzalo Peña

INDICE

1. Introducción ……. 8

2. Objetivos ……. 12

2.1. Objetivos generales ……. 12

2.2. Objetivos específicos ……. 12

3. Generalidades del Procedimiento de Medición Face2.0 Existente en la Argentina y del Procedimiento Superador …… 13

4. Comparación y viabilidad de los criterios de aplicación tanto del Procedimiento de Medición Face2.0 Existente en la Argentina como del Procedimiento Superador ……... 15

4.1. Exito de la e-campaña vía Facebook ……... 16

4.1.1. Intención de voto ……... 19

4.1.2. Electorado ……... 20

4.1.3. Influencia …… 21

4.2. Diseño de investigación ……..... 22

4.3. Uso de muestra o población ……... 25

4.4. Muestra ……... 26

4.5. Unidad de análisis …….. 27

4.6. Población de estudio ……..... 30

4.7. Tamaño mínimo muestral …….. 31

4.8. Tamaño de la muestra ……... 34

4.9. Tipo de muestra …… 36

4.10. Selección de los elementos muestrales …… 40

4.11. Técnicas de análisis de datos ……... 41

5. Viabilidad del Procedimiento de Medición Face2.0 Existente en la Argentina y del Procedimiento Superador ……. 52
6. Criterios de aplicación de las técnicas de análisis de datos del Procedimiento Superador ……..... 54
7. Análisis de regresión lineal …….. 54
7.1. Análisis de regresión lineal simple …… 55
7.1.1. Ecuación de regresión estimada en regresión lineal simple ... 57
7.1.2. Método de los mínimos cuadrados en regresión lineal simple …….. 59
7.1.3. Medidas de variación en regresión lineal simple …… 60
7.1.4. Coeficiente de determinación en regresión lineal simple ….. 62
7.1.5. Coeficiente de correlación en regresión lineal simple …… 63
7.1.6. Análisis residual en regresión lineal simple ……..... 67
7.1.7. Detección de observaciones atípicas en regresión lineal simple ……. 67
7.1.7.1. Residuales estudentizados eliminados y observaciones atípicas en regresión lineal simple ……...…… 70
7.1.8. Observaciones influyentes en regresión lineal simple ……..... 71
7.1.8.1. Uso de la medida de la distancia de Cook para identificar observaciones influyentes en regresión lineal simple ……. 74
7.1.9. Suposiciones del modelo de regresión lineal simple ….…… 76
7.1.9.1. Evaluación de la linealidad en regresión lineal simple ……... 77

7.1.9.2. Evaluación de la independencia en regresión lineal simple ……...… 80

7.1.9.3. Evaluación de la normalidad en regresión lineal simple ……..……. 86

7.1.9.4. Evaluación de la igualdad de varianza en regresión lineal simple ……. 89

7.1.10. Prueba de significancia de la relación en regresión lineal simple …… 91

7.1.10.1. Estimación de σ^2 y error de estimación en regresión lineal simple ……...…… 92

7.1.10.2. Prueba *t* en regresión lineal simple …….……... 93

7.1.10.3. Determinación de la significancia a partir de la estimación del intervalo de confianza en regresión lineal simple ……..……. 96

7.1.10.4. Prueba *F* en regresión lineal simple …….…….. 97

7.1.10.5. Algunas advertencias acerca de la interpretación de las pruebas de significancia en regresión lineal simple ……..……..... 99

7.1.11. Uso de la ecuación de regresión estimada para estimaciones y predicciones en regresión lineal simple ….. 100

7.1.11.1. Estimación puntual a partir de la ecuación de regresión estimada en regresión lineal simple …….……... 100

7.1.11.2. Estimación por intervalo a partir de la ecuación de regresión estimada en

regresión lineal simple ……...…... 101

7.1.11.2.1. Intervalo de confianza para el valor medio de *y* en regresión lineal simple .. 101

7.1.11.2.2. Intervalo de predicción para un solo valor de *y* en regresión lineal simple 103

7.2. Análisis de regresión múltiple …….. 105

7.2.1. Modelo de regresión y ecuación de regresión en regresión múltiple …….. 105

7.2.2. Ecuación de regresión múltiple estimada …….. 107

7.2.3. Método de mínimos cuadrados en regresión múltiple …….. 108

7.2.4. Modelo lineal general ……. 112

7.2.5. Interacción ……..…… 113

7.2.6. Coeficiente de determinación múltiple ……..…. 116

7.2.7. Coeficiente de correlación múltiple ……..….. 119

7.2.8. Detección de observaciones atípicas en regresión múltiple …….. 120

7.2.9. Observaciones influyentes en regresión múltiple ……..…. 121

7.2.9.1. Distancia de Cook e identificación de observaciones influyentes …...…… 122

7.2.10. Supuestos del modelo de regresión múltiple ……... 122

7.2.10.1. Incumplimiento de los supuestos de regresión y uso de transformaciones …… 125

7.2.11. Multicolinealidad ……..…… 144

7.2.12. Prueba de significancia en regresión múltiple …….…… 151

7.2.12.1. Prueba *F* en regresión múltiple ……. 152

7.2.12.2. Prueba *t* en regresión múltiple …….. 154

7.2.13. Determinación de cuándo agregar o eliminar variables en regresión múltiple …….……... 157

7.2.13.1. Uso del valor-*p* en regresión múltiple …….. 163

7.2.13.2. Análisis de un problema mayor en regresión múltiple ……. 163

7.2.13.2.1. Regresión por pasos …….……... 169

7.2.13.2.2. Selección hacia adelante ..…….……... 171

7.2.13.2.3. Eliminación hacia atrás ……...…… 171

7.2.13.2.4. Regresión de los mejores subconjuntos …… 172

7.2.13.2.5. Elección final …….……..… 174

7.2.13.3. Variables confusoras ……....……. 175

7.2.14. Estimación del intervalo de confianza en regresión múltiple …… 176

7.2.14.1. Uso de la ecuación de regresión estimada para estimaciones y predicciones en regresión múltiple ……..….... 176

7.2.15. Coeficientes de regresión parcial estandarizados en regresión múltiple ……..... 177

7.2.16. Variables cualitativas independientes ……..…….……. 181

7.2.16.1. Interpretación de los parámetros ante la presencia de variables cualitativas independientes …….. 181

7.2.16.2. Variables cualitativas más complejas ..…..…… 185

7.3. Ejemplos de análisis de regresión en SPSS ……..……. 187

7.3.1. Caso de regresión múltiple en SPSS con análisis de supuestos y eliminación de observaciones influyentes …….. 187

7.3.2. Caso de regresión múltiple en SPSS con análisis de componentes principales como solución de la multicolinealidad …… 211

7.3.3. Caso de regresión múltiple en SPSS con determinación de intervalos de confianza y predicción …….……. 226

7.3.4. Caso de regresión múltiple con presencia de efecto cuadrático en modelo de regresión simple inicial …….. 229

7.3.5. Caso de regresión múltiple con variables categóricas ……... 240

8. Propuesta de un procedimiento superador que en un momento específico determine el éxito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina ……..…… 245

9. Conclusión …….. 247

10. Anexos …….. 249

11. Bibliografía ……... 256

1. Introducción

Las redes sociales están establecidas en la cultura universal de hoy día. Por un lado debido a la presencia global de redes sociales ampliamente seguidas[1] y por otro lado al tiempo que le destinan sus usuarios. Así, según iRedes (2013) el *Mapa iRedes (versión III)*[2] señala que existen 101 redes sociales[3] con más de 10 millones de usuarios alrededor del mundo. En tanto, comScore (2012) indica en noviembre de 2012 que los países que integran el top 10 en cantidad de horas dedicadas a usar redes sociales al mes registran un uso igual o superior a 7,5 horas promedio. Destacan especialmente los casos de Argentina y Brasil que según comScore (2012) representan respectivamente el primer y segundo puesto de esa clasificación. Brasil de acuerdo a comScore (2012) destina al mes 9,7 horas promedio. Mientras, la Argentina es el líder mundial con promedio de 9,8 horas invertidas al mes en redes sociales. Según comScore (2012) dicho estudio se realizó con mayores de 15 años que se conectaron desde una computadora de escritorio en su casa o trabajo. Por esa razón conforme a Crettaz, J. (2013) Zuzenberg (2013)[4] señala que si se considera el consumo desde móviles el tiempo que los argentinos invierten en dichas redes sería superior.

Entre las diversas redes sociales existentes iRedes (2013) destaca a Facebook como la principal del mundo con más de 1 billón (1.060 millones) de usuarios. De acuerdo a Cosenza, V. (2013) esta red social también lidera el mercado argentino. Al respecto Crettaz (2013) afirma

[1] En este estudio cada red social con 10 millones de usuarios o más se considera ampliamente seguida.
[2] La cursiva es nuestra.
[3] La cantidad sería mayor al contar las redes sociales con menos de 10 millones de usuarios en el mundo.
[4] Alejandro Zuzenberg es el director de Facebook Argentina.

> [Según] (…) Zuzenberg (…) algo más de 20 millones de argentinos, la mitad de la población, usan (…) [esa] red social para publicar, compartir o dar un "like" (valorar positivamente) textos, fotos o videos propios o de sus contactos. Eso equivale al 65% de los 31,1 millones de usuarios de Internet en el país.

No obstante, en la Argentina el uso masivo de Facebook no es novedad. Esta herramienta virtual no solo ha sido utilizada años atrás por particulares sino también por políticos fuera y dentro del marco de campañas electorales. Así, en las elecciones legislativas de 2009 los candidatos argentinos utilizaron esta tecnología como nunca antes en una campaña electoral. De acuerdo a Erbin, A. (2009) Facebook se constituyó en una de las redes sociales protagonistas en las estrategias de los aspirantes al voto argentino. Los casos más destacados conforme a Erbin (2009) fueron los de Francisco de Narváez, Margarita Stolbizer y Martín Sabatella. Según Erbin (2009) aunque dichos candidatos no apelaron al recurso de interacción con la población que ofrece la web 2.0, usaron activamente sus cuentas en Facebook. Tras celebrarse dichas elecciones el uso de Facebook se popularizó en la política argentina. A partir del segundo semestre de 2009 el interés político en esta herramienta se hizo más marcado y se consolidó en 2010, año que representó la antesala de las campañas electorales de 2011. En 2011 se realizó dos elecciones. Por una parte las elecciones para Jefe de Gobierno de la Ciudad Autónoma de Buenos Aires y por otra parte las elecciones presidenciales. Respecto a los años anteriores en ninguna de las dos se verificó un cambio en el modo de uso. Conforme a Nadal, H. (2011) Erbin (s. f.) indica que nuevamente no se explotó la capacidad interactiva que brinda la web 2.0. Sin embargo en esta oportunidad destaca una mayor participación de los distintos candidatos. Según Todo Noticias [TN] (2011) en ambas elecciones todos los candidatos, excepto el candidato a Jefe de Gobierno del Nuevo MAS

llamado César Rojas, tuvieron presencia en Facebook. TN (2011) indica que en las elecciones a Jefe de Gobierno contaron con una cuenta Facebook oficial los candidatos Mauricio Macri (Propuesta Republicana - PRO), Daniel Filmus (Frente Para la Victoria - FPV), Fernando Pino Solanas (Proyecto Sur), Jorge Telerman (Frente Progresista Más Buenos Aires), Silvana Giudici (Unión Cívica Radical – UCR), Ricardo López Murphy (Partido Autonomista Nacional - PAN), María Eugenia Estenssoro (Coalición Cívica – CC), Javier Castrilli (Acción Ciudadana - AC), Luis Zamora (Autodeterminación y Libertad - AyL), Alejandro Biondini (Alternativa Social - AS), Enrique Piragini (Alianza Frente de los Ciudadanos - AFC), Jorge Todesca (Movimiento de Integración y Desarrollo – MID), y Myriam Bregman (Frente de Izquierda y de los trabajadores - FIT). Igualmente TN (2011) señala que en las elecciones presidenciales también tuvieron una cuenta Facebook oficial todos los candidatos excepto Jorge Altamira (Frente de Izquierda y de los Trabajadores - FIT). En ese sentido según TN (2011) estuvieron en Facebook Eduardo Duhalde (Frente Popular - FP), Cristina Fernández de Kirchner (Frente Para la Victoria - FPV), Ricardo Alfonsín (Unión por el Desarrollo Social – UDeSo), Hermes Binner (Frente Amplio Progresista – FAP), y Alberto Rodríguez Saá (Compromiso Federal – CF), Elisa Carrió (Coalición Cívica – CC). En tanto, en 2013 se celebró las elecciones legislativas a nivel provincial y nacional. En dicho escenario, una vez más, la interacción que ofrece Facebook no fue explotada. No obstante, según Agencia Nacional de Noticias [Télam] (2013) se mantuvo la amplia participación de los candidatos electorales. Como ejemplo según Télam (2013) destaca el caso de los candidatos a Diputado Nacional en la fase general de la elección pues 91 de los 122 candidatos contó con presencia en Facebook.

Como se observa, en la Argentina se puede destacar dos situaciones fundamentales en torno a Facebook. Por un lado, el dominio de esa red social en el mercado argentino. Por otro lado, la presencia generalizada de actores políticos en la misma especialmente en tiempos de campaña electoral. Debido a ambas situaciones Facebook representa una herramienta electoral a investigar.

En la Argentina existe un tipo de procedimiento al que en este estudio se denomina *Procedimiento de Medición Face2.0*[5]. Se trata de todo procedimiento que afirme o sugiera implícitamente que puede determinar en una elección el éxito de la e-campaña vía Facebook, también entendida como campaña a través de esa red social. Por un lado, esto puede ser midiendo el éxito de la e-campaña vía Facebook de un candidato electoral en específico. Por otro lado, puede ser determinando el éxito de una e-campaña vía Facebook en sentido general, no asociada a un candidato en particular. En este último caso se hará referencia al éxito de la e-campaña vía Facebook de un candidato y en el caso de la primera postura se hará alusión al éxito de la e-campaña vía Facebook de un candidato determinado. Según el estado del arte en la Argentina hay un solo Procedimiento de Medición Face2.0 el cual se denomina en esta investigación *Procedimiento de Medición Face2.0 Existente en la Argentina*[6]. Dado a que dicho procedimiento plantea implícitamente ser capaz de realizar la medición del citado éxito, es oportuno evaluar si puede determinarlo. Asimismo, más allá de los resultados de la evaluación, conviene establecer un procedimiento que supere al actualmente vigente.

[5] La cursiva es nuestra.
[6] Ibidem.

2. Objetivos

2.1. Objetivos generales

1. Determinar la viabilidad del Procedimiento de Medición Face2.0 Existente en la Argentina para medir el éxito de la e-campaña vía Facebook de un candidato electoral en particular.

2. Proponer un procedimiento superador que determine en un momento dado el éxito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina.

2.2. Objetivos específicos

1.1 Definir los criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina.

1.2 Señalar el aval científico de los criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina.

1.3 Indicar la pertinencia de los criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina.

1.4 Establecer la viabilidad del Procedimiento de Medición Face2.0 Existente en la Argentina según el aval científico y la pertinencia de sus criterios de aplicación.

2.1. Identificar las diferencias entre los criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina y los del Procedimiento Superador.

2.2. Detallar la manera en que el Procedimiento Superador debe emplearse.

3. Generalidades del Procedimiento de Medición Face2.0 Existente en la Argentina y del Procedimiento Superador

José Fernández-Ardáiz, Presidente de la Consultora Integral en Comunicación Aplicada (CICoA), es autor del *Indice de medición 2.0*[7]. Conforme a Fernández-Ardáiz, J. (2010) dicho índice mide el éxito de una e-campaña, o campaña por internet, de un candidato electoral en específico[8]. Según Fernández-Ardáiz (2010) la medición se realiza a partir de herramientas virtuales como Facebook, Twitter, entre otras, en las que dicho político participe. Dentro del marco de esta investigación en el caso puntual de Facebook la medición se hace a través del Procedimiento de Medición Face2.0 Existente en la Argentina. Para eso dicho procedimiento cuenta con los denominados *Criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina*[9]. Los mismos se han constituido en este estudio en base a varios términos. A saber:

- Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina

[7] La cursiva es nuestra.
[8] De acuerdo a Fernández-Ardáiz (2010) el Indice de medición 2.0 puede medir el éxito de una e-campaña de cualquier político, candidato o partido político. No obstante en esta investigación solo se asumirá la e-campaña de un candidato electoral para facilitar la comparabilidad con el procedimiento superador que se plantea en la misma.
[9] La cursiva es nuestra.

- Diseño de investigación del Procedimiento de Medición Face2.0 Existente en la Argentina
- Uso de muestra o población en el Procedimiento de Medición Face2.0 Existente en la Argentina
- Muestra en el Procedimiento de Medición Face2.0 Existente en la Argentina
- Unidad de análisis del Procedimiento de Medición Face2.0 Existente en la Argentina
- Población de estudio del Procedimiento de Medición Face2.0 Existente en la Argentina
- Tipo de muestra del Procedimiento de Medición Face2.0 Existente en la Argentina
- Selección de los elementos muestrales del Procedimiento de Medición Face2.0 Existente en la Argentina
- Tamaño mínimo muestral del Procedimiento de Medición Face2.0 Existente en la Argentina
- Tamaño de la muestra del Procedimiento de Medición Face2.0 Existente en la Argentina
- Técnicas de análisis de datos del Procedimiento de Medición Face2.0 Existente en la Argentina

En tanto, con la aspiración de realizar la mencionada medición de una mejor manera este estudio propone un procedimiento superador, llamado de ahora en más

Procedimiento Superador[10]. A esos fines, el mismo establece los llamados en esta investigación *Criterios de aplicación del Procedimiento Superador*[11]. A saber son:

- Exito de la e-campaña vía Facebook en el Procedimiento Superador
- Diseño de investigación del Procedimiento Superador
- Uso de muestra o población en el Procedimiento Superador
- Muestra en el Procedimiento Superador
- Unidad de análisis del Procedimiento Superador
- Población de estudio del Procedimiento Superador
- Tipo de muestra del Procedimiento Superador
- Selección de los elementos muestrales del Procedimiento Superador
- Tamaño mínimo muestral del Procedimiento Superador
- Tamaño de la muestra del Procedimiento Superador
- Técnicas de análisis de datos del Procedimiento Superador

Como se aprecia los criterios de aplicación son los mismos en ambos procedimientos sin embargo los contenidos de cada uno varían entre sí. En ese sentido a continuación se detallan dichas diferencias y se analiza la viabilidad de los referidos criterios.

4. Comparación y viabilidad de los criterios de aplicación tanto del Procedimiento de Medición Face2.0 Existente en la Argentina como del Procedimiento Superador

[10] La cursiva es nuestra.
[11] Ibidem.

4.1. Exito de la e-campaña vía Facebook

El término *Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina*[12] se desprende del concepto del Indice de medición 2.0 llamado *Exito de la e-campaña*[13]. Este último conforme a Fernández-Ardáiz (2010) cuenta con dos acepciones. La primera según Fernández-Ardáiz (2010) es el funcionamiento de la e-campaña en las diferentes herramientas virtuales de un candidato determinado. Desde esa perspectiva se deduce que la e-campaña sería exitosa en función del rendimiento que la misma logre en las distintas herramientas virtuales de un candidato puntual. En tanto, de acuerdo a Fernández-Ardáiz (2010) la segunda acepción es la cantidad de personas que se sume a trabajar física y virtualmente en el proyecto político en cuestión. Considerando estas dos definiciones del Exito de la e-campaña se deduce que el Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina puede interpretarse restringiendo el alcance de dichas acepciones al contexto de Facebook. Por un lado, el Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina se define conceptualmente como el funcionamiento de la referida e-campaña en la cuenta Facebook oficial/personal de un candidato dado. Por otro lado, el Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina se define conceptualmente como la adición de una cantidad de personas que trabajen física y virtualmente en el proyecto político de un candidato en particular.

[12] La cursiva es nuestra.
[13] Ibidem.

De esa manera la viabilidad del criterio *Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina*[14] se aborda conforme a sus dos acepciones. Por una parte figura la acepción, denominada en este estudio, Funcionamiento de la e-campaña en la cuenta Facebook oficial/personal de un candidato electoral determinado. La misma no consta de aval científico ya que define el Exito de la e-campaña vía Facebook en base a variables y estándares de dichas variables establecidos por el propio criterio sin poder demostrar la razón por la que esas variables y sus respectivos estándares explican el éxito de la referida e-campaña. Tampoco es pertinente pues es incoherente explicar el mencionado éxito en base a la medición aislada del funcionamiento de la e-campaña en la cuenta Facebook oficial/personal de un candidato electoral específico sin demostrar una relación de dependencia. Por otra parte se encuentra la acepción llamada, en esta investigación, Cantidad de personas que se incorpore a trabajar física y virtualmente en el proyecto político de un candidato electoral dado. La misma no cuenta con aval científico pues define al éxito de la e-campaña vía Facebook en función de una variable sin ser capaz de comprobar por qué dicha variable explica el éxito en cuestión. Igualmente, más allá de que esta acepción no especifica el número de personas que deben sumarse al proyecto político del candidato para considerar la e-campaña vía Facebook exitosa, se considera que la misma no es pertinente. Eso se debe a que la adición de un determinado número de individuos al trabajo del proyecto de un candidato en particular representa un indicador poco fiable de que la e-campaña vía Facebook de dicho político sea exitosa. Mientras, el *Exito de la e-campaña vía Facebook en el Procedimiento Superador*[15] se basa en la relación entre dos factores. Por un lado, la principal cuenta Facebook de los candidatos a Diputado que encabecen la lista de sus

[14] La cursiva es nuestra.
[15] Ibidem.

partidos en unas elecciones legislativas generales nacionales, en lo adelante también denominados *Candidatos a Diputado*[16]. En otras palabras la (s) variable (s) de la referida principal cuenta Facebook que sea (n) utilizada (s) en el análisis como representación de dicha cuenta. Por otro lado, la intención de voto de los candidatos a Diputado en cuestión ya que si en un momento dado de una campaña electoral se releva apropiadamente dicha intención se puede estimar los resultados electorales que se obtendrían si la elección se celebrara ese día. En ese sentido la experiencia en campañas electorales a nivel mundial demuestra que cuando se releva adecuadamente la intención de voto a pocos días de los comicios se logra una acertada predicción de los resultados electorales finales. Así, al demostrar que la principal cuenta Facebook de un candidato a Diputado, representada en el análisis por una o varias variables de esa cuenta, influye[17] sobre la intención de voto se establece que la e-campaña vía Facebook del mismo es exitosa. De esa manera el Exito de la e-campaña vía Facebook en el Procedimiento Superador se define conceptualmente como la influencia de la principal cuenta Facebook de un candidato a Diputado sobre la intención de voto. En ese sentido se concluye que el criterio en cuestión goza de aval científico pues define el Exito de la e-campaña vía Facebook en base a una relación de dependencia estadísticamente comprobable. Igualmente se concluye que es pertinente pues es coherente determinar el Exito de la e-campaña vía Facebook en función de la influencia de la principal cuenta Facebook de un candidato a Diputado sobre la intención de voto.

[16] La cursiva es nuestra.
[17] Ver definición del término Influencia en el apartado Influencia.

En tanto, ya que las recién detalladas definiciones conceptuales están relacionadas con los términos intención de voto, electorado e influencia dichas terminologías se abordan a continuación.

4.1.1. Intención de voto

En el Procedimiento de Medición Face2.0 Existente en la Argentina no se contempla la intención de voto pero tampoco es necesario que se considere. En cambio, en el Procedimiento Superador la referida intención de voto sí es tenida en cuenta. *Diccionario de Ciencias Jurídicas, Políticas y Sociales*[18] establece: "[Intención es un] propósito de conducta" (Osorio, 2007, p. 527). Igualmente *Diccionario de Ciencias Jurídicas, Políticas y Sociales*[19] indica: "[Voto representa] (…) en los comicios el parecer que se manifiesta (…) por medio de papeletas [o vía electrónica] (…) para elegir a alguna persona (…) para determinados cargos (…)" (Osorio, 2007, p. 1027). De esa manera *Intención de voto*[20] implica en este estudio *la opinión*[21] que emite el electorado respecto a su propósito de votar a un candidato para que ocupe un puesto específico. Por ende en el Procedimiento Superador la misma equivale a la opinión que declare el electorado sobre su propósito de votar para Diputado a un candidato que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina.

[18] La cursiva es nuestra.
[19] Ibidem.
[20] Ibidem.
[21] Ibidem.

4.1.2. Electorado

Diccionario de Ciencias Jurídicas, Políticas y Sociales[22] afirma: "[Electorado es el] conjunto de los electores de un país o circunscripción" (Osorio, 2007, p. 379). Al respecto, *Diccionario de Ciencias Jurídicas, Políticas y Sociales*[23] agrega: "[Los electores son las personas] con derecho de sufragio" (Osorio, 2007, p. 379). En procedimientos como los analizados en este estudio la variable *Electorado*[24], bajo la definición recién especificada, debe formar parte de la descripción del Exito de la e-campaña vía Facebook. En ese sentido está supuesta a figurar, explícita o implícitamente, en la relación de dependencia que sustenta al mencionado éxito. Esto no sucede en el Procedimiento de Medición Face2.0 Existente en la Argentina pues Fernández-Ardáiz (2010) no basa en una relación de dependencia ninguna de las dos acepciones con que define el éxito en cuestión. En consecuencia, la variable *Electorado*[25] no se presenta explícita ni implícitamente en la relación de dependencia requerida. Contrario pasa con el Procedimiento Superador en el cual el electorado participa implícitamente en la relación de dependencia que fundamenta al Exito de la e-campaña vía Facebook. Es decir, la relación entre la principal cuenta Facebook de un candidato a Diputado y la intención de voto. La participación implícita se debe a que la variable dependiente en ese contexto está constituida por la intención de voto del electorado, entendiendo *Electorado*[26] como los individuos aptos para votar en unas elecciones legislativas generales nacionales de la Argentina.

[22] La cursiva es nuestra.
[23] Ibidem.
[24] Ibidem.
[25] Ibidem.
[26] Ibidem.

4.1.3. Influencia

A los fines de este estudio el concepto *Influencia*[27] se define como la capacidad de una o varias variables de alterar a otra. Igualmente en el ámbito del Procedimiento Superador dicho concepto representa la capacidad de la principal cuenta Facebook de un candidato a Diputado de variar la intención de voto del electorado en unas elecciones legislativas generales nacionales de la Argentina. El estado del arte carece de una clasificación de los tipos de influencia en el marco de las redes sociales. Por ende, en el Procedimiento Superador, se ha establecido diferentes tipos de influencia según las personas que se influya y la naturaleza de la relación de dependencia. De acuerdo a las personas que se influya la influencia puede ser *Directa*[28], *Indirecta*[29] y *Mixta*[30]. La primera, la influencia directa, se presenta cuando la principal cuenta Facebook de un candidato a Diputado influye en los electores miembros de dicha cuenta. La segunda, la influencia indirecta, se aprecia cuando la referida cuenta influye en los electores que no son miembro de la misma. La tercera, la influencia mixta, se observa cuando la mencionada cuenta influye tanto en los electores que son miembro de la cuenta en cuestión como en los que no lo son. Según la naturaleza de la relación de dependencia la influencia puede ser *Lineal*[31] o *Intrínsecamente lineal*[32]. La influencia lineal se verifica cuando un modelo de regresión lineal cumple con los supuestos de dicho tipo de regresión. En dicho modelo la (s) variable (s) independiente (s) es (son) la (s) variable (s) que representa (n) a la principal cuenta Facebook de un candidato a Diputado y la variable dependiente es la intención de voto. Los mencionados supuestos se abordan en los apartados *Supuestos del modelo*

[27] La cursiva es nuestra.
[28] Ibidem.
[29] Ibidem.
[30] Ibidem.
[31] Ibidem.
[32] Ibidem.

en regresión lineal simple[33] y *Supuestos del modelo en regresión múltiple*[34]. En tanto, la influencia intrínsecamente lineal se presenta en un modelo de regresión no lineal que luego de ser transformado a lineal satisface los mencionados supuestos. En ese modelo también la (s) variable (s) independiente (s) es (son) la (s) variable (s) que representa (n) a la principal cuenta Facebook de un candidato a Diputado y la variable dependiente es la intención de voto. Cabe destacar que el estado del arte no muestra evidencia sobre relaciones no lineales no linealizables entre la intención de voto y otras variables. En tanto, Fernández-Ardáiz (2010) no considera la variable Influencia como parte del Procedimiento de Medición Face2.0 Existente en la Argentina lo cual constituye un tema central. Esto así, ya que en el contexto de este tipo de procedimiento se precisa investigar si una o varias variables influyen en otra a fin de determinar si la e-campaña es exitosa. Sin embargo, en el Procedimiento Superador se asumirá como *Influencia*[35] únicamente la influencia mixta en su modalidad lineal o intrínsecamente lineal.

4.2. Diseño de investigación

"Diseño [es el] plan o estrategia que se desarrolla para obtener la información que se requiere en una investigación" (Hernández Sampieri, Fernández Collado, y Baptista Lucio, 2010, p. 120). "En la literatura sobre la investigación cuantitativa es posible encontrar diferentes clasificaciones de los diseños. En [el marco de este estudio se adopta la clasificación que agrupa los diseños en] investigación experimental e investigación no experimental" (Hernández Sampieri y otros, 2010, p. 121). "[En

[33] La cursiva es nuestra.
[34] Ibidem.
[35] Ibidem.

relación a la investigación experimental] el término experimento tiene dos acepciones, una general y otra particular" (Hernández Sampieri y otros, 2010, p. 121). Hernández Sampieri y otros (2010) establecen que según Babbie (2009) la general consiste en realizar una acción y luego observar las consecuencias. "(…) [La particular implica] un estudio (…) [donde] se manipulan intencionalmente una o más variables independientes (supuestas causas-antecedentes) para analizar las consecuencias que la manipulación tiene sobre una o más variables dependientes (supuestos efectos-consecuentes) dentro de una situación de control (…)" (Hernández Sampieri y otros, 2010, p. 121). En cuanto a la investigación no experimental Hernández Sampieri y otros (2010) establecen: "(…) [La misma] podría definirse como [aquella] que se realiza sin manipular deliberadamente variables. (…) Lo que [se hace] (…) es observar fenómenos tal como se dan en su contexto natural para posteriormente analizarlos" (p. 149).

En tanto, "los diseños no experimentales se pueden clasificar en longitudinales y transeccionales" (Hernández Sampieri y otros, 2010, p. 151). "(…) [Los] diseños longitudinales [implican] estudios que recaban datos en diferentes puntos del tiempo para realizar inferencias acerca de la evolución, sus causas y (…) efectos" (Hernández Sampieri y otros, 2010, p. 158). "(…) [En cambio,] los diseños de investigación transeccional o transversal recolectan datos en un solo momento, en un tiempo único" (Hernández Sampieri y otros, 2010, p. 151). "A su vez, los diseños transeccionales se dividen en tres: exploratorios, descriptivos y correlacionales-causales" (Hernández Sampieri y otros, 2010, p. 152). "El propósito de los diseños transeccionales exploratorios es comenzar a conocer una variable o un conjunto de variables, (…) un evento, una situación, [a través de] (…) una exploración inicial en

un momento específico" (Hernández Sampieri y otros, 2010, p. 152). "Por lo general se aplican a problemas de investigación nuevos o poco conocidos" (Hernández Sampieri y otros, 2010, p. 152). Igualmente "los diseños transeccionales descriptivos tienen como objetivo (…) ubicar en una o diversas variables a un grupo de personas u otros seres vivos, objetos, situaciones, contextos, fenómenos, comunidades, y así proporcionar su descripción" (Hernández Sampieri y otros, 2010, pp. 152-153). "(…) [Mientras, los diseños transeccionales correlacionales-causales] describen relaciones entre dos o más categorías, conceptos o variables en un momento determinado. A veces, únicamente en términos correlacionales, otras en función de la relación causa-efecto (causales)" (Hernández Sampieri y otros, 2010, p. 154). "Cuando [dichos diseños] se limitan a relaciones no causales se fundamentan en planteamientos e hipótesis correlacionales, del mismo modo cuando buscan evaluar vinculaciones causales se basan en planteamientos e hipótesis causales" (Hernández Sampieri y otros, 2010, p. 155). De esa manera se observa que el Procedimiento de Medición Face2.0 Existente en la Argentina se basa en un diseño transeccional descriptivo pues en el caso de las dos acepciones del Exito de la e-campaña vía Facebook se limita a medir variables en un momento específico y a partir de dichos resultados establece si la e-campaña fue exitosa. Así, se concluye que el criterio *Diseño de investigación del Procedimiento de Medición Face2.0 Existente en la Argentina*[36] consta de aval científico ya que es un tipo de diseño reconocido en la metodología de investigación. Igualmente se concluye que dicho diseño no es pertinente pues se restringe a medir determinadas variables en un momento dado sin analizar la presencia de relación correlacional-causal entre ellas. En cambio, el Procedimiento Superador se fundamenta en un diseño de investigación transeccional correlacional-causal que

[36] La cursiva es nuestra.

permite determinar si una o varias variables influyen en otra lo cual define si la e-campaña ha sido exitosa. En ese sentido se concluye que en el Procedimiento Superador el criterio en cuestión goza de aval científico y es pertinente. Eso se debe a que el mismo responde a un tipo de diseño científicamente aprobado que habilita el análisis de una relación correlacional-causal entre variables.

4.3. Uso de muestra o población

El Procedimiento de Medición Face2.0 Existente en la Argentina implica la realización de un estudio poblacional. En ese sentido se debe tener en cuenta las especificaciones desarrolladas a continuación en los apartados *Unidad de análisis*[37] y *Población de estudio*[38]. De esa forma se concluye que en el Procedimiento de Medición Face2.0 Existente en la Argentina el criterio *Uso de muestra o población*[39] cuenta con aval científico pues la modalidad de investigación poblacional es reconocida en el contexto de la ciencia. Igualmente se concluye que el mencionado criterio es pertinente pues la metodología de la investigación habilita al investigador a estudiar a una población en tanto se garantice el cumplimiento de las especificaciones de dicha modalidad. En tanto, en el Procedimiento Superador solo se realiza el tipo de estudio muestral. Eso se debe a que dicho procedimiento se basa en la utilización del programa estadístico Statistical Package for the Social Sciences (SPSS, por sus siglas en inglés) el cual según Universidad de Antioquía (s. f.) solo realiza estimaciones de las medidas de variación. Así, el análisis de regresión lineal en SPSS debe fundamentarse en datos muestrales a fin de estimar las referidas medidas de variación y, consecuentemente, lograr el análisis deseado. Por ende, al basarse en una muestra,

[37] La cursiva es nuestra.
[38] Ibidem.
[39] Ibidem.

en el Procedimiento Superador se considera las especificaciones abordadas a continuación en los apartados *Unidad de análisis*[40], *Población de estudio*[41], *Tamaño mínimo muestral*[42], *Tamaño de la muestra* [43], *Tipo de muestra*[44] y *Selección de los elementos muestrales*[45]. En ese sentido el criterio *Uso de muestra o población*[46] en el Procedimiento Superador cuenta con aval científico pues la modalidad de investigación muestral es reconocida en el contexto de la ciencia. Igualmente se concluye que el mencionado criterio es pertinente pues la metodología de la investigación habilita al investigador a utilizar una muestra en tanto se garantice el cumplimiento de las especificaciones de dicha modalidad.

4.4. Muestra

"Para el proceso cuantitativo la muestra es un subgrupo de la población de interés sobre la cual se recolectarán los datos, y que tiene que definirse o delimitarse de antemano con precisión" (Hernández Sampieri y otros, 2010, p. 173). Fernández-Ardáiz (2010) no detalla una definición de muestra en la descripción que hace del Indice 2.0. Sin embargo no es necesario que lo haga pues no utiliza la modalidad de estudio muestral en el Procedimiento de Medición Face2.0 Existente en la Argentina. En ese sentido no se establece una conclusión respecto a la viabilidad del criterio *Muestra*[47] en dicho procedimiento. En tanto, el Procedimiento Superador opera bajo la definición de muestra de Hernández Sampieri y otros (2010). Por ende se concluye que el criterio en cuestión consta de aval científico. Igualmente se concluye que es

[40] La cursiva es nuestra.
[41] Ibidem.
[42] Ibidem.
[43] Ibidem.
[44] Ibidem.
[45] Ibidem.
[46] Ibidem.
[47] Ibidem.

pertinente pues plantea disminuir la cantidad de casos a analizar en función de criterios científicos.

4.5. Unidad de análisis

Según Hernández Sampieri y otros (2010) a las unidades de análisis también se les conoce como casos o elementos. Conforme a Hernández Sampieri y otros (2010) dichos casos representan a los participantes, objetos, sucesos, u otros, que serán relevados en el estudio. Así, la unidad de análisis del Procedimiento de Medición Face2.0 Existente en la Argentina es la misma en el caso de las dos acepciones del Exito de la e-campaña vía Facebook. En ambos casos consiste en la cuenta Facebook oficial/personal del candidato electoral que se desee investigar en un momento determinado. Esto sugiere que Fernández-Ardáiz (2010) se basa en la descripción de unidad de análisis que plantea Hernández Sampieri (2010). Por ende en el contexto de las dos acepciones del Exito de la e-campaña vía Facebook el criterio *Unidad de análisis*[48] consta de aval científico. No obstante, siendo la unidad de análisis del Procedimiento de Medición Face2.0 Existente en la Argentina la cuenta Facebook oficial/personal de un candidato puntual se concluye que dicho criterio no es pertinente. Eso se debe a que esa perspectiva no considera el caso de los candidatos que puedan tener varias cuentas oficiales/personales. En cambio, la unidad de análisis del Procedimiento Superador es la principal cuenta Facebook de los candidatos a Diputado que encabezan la lista de sus respectivos partidos en unas elecciones legislativas generales nacionales de la Argentina. Así, en el escenario del Procedimiento Superador se concluye que el criterio *Unidad de análisis*[49] goza de

[48] La cursiva es nuestra.
[49] Ibidem.

aval científico dado a que responde a la definición aportada por Hernández Sampieri y otros (2010). Igualmente se concluye que es pertinente. Eso se debe a que solo puede existir una cuenta Facebook principal de los candidatos a Diputado que encabezan la lista de sus respectivos partidos en unas elecciones legislativas generales nacionales de la Argentina. Es decir, solo puede haber una cuenta Facebook en la que dichos candidatos demuestren mayor participación o, en caso contrario, donde se verifique en mayor grado la relación *Cantidad de contactos o seguidores - Participación de personas*[50]. De esa forma es exhaustiva la descripción de unidad de análisis que adopta el Procedimiento Superador.

En este contexto conviene señalar que tanto en el Procedimiento de Medición Face2.0 Existente en la Argentina como en el Procedimiento Superador la unidad de análisis está compuesta por dos factores en común denominados *Candidato*[51] y *Cuenta Facebook*[52]. Ya que según el procedimiento ambos factores difieren entre sí es oportuno comparar por un lado las terminologías *Candidato electoral*[53] y *Candidato a Diputado*[54], y por otro lado los conceptos *Cuenta Facebook oficial/personal*[55] y *Principal cuenta Facebook*[56].

En cuanto a la primera comparación se destaca lo siguiente. *Diccionario de Ciencias Jurídicas, Políticas y Sociales*[57] señala: "[Un candidato es una] persona que pretende alguna dignidad, honor o cargo" (Osorio, 2007, p. 151). Considerando cualquiera de las dos acepciones del *Exito de la e-campaña vía Facebook en el Procedimiento de*

[50] La cursiva es nuestra.
[51] Ibidem.
[52] Ibidem.
[53] Ibidem.
[54] Ibidem.
[55] Ibidem.
[56] Ibidem.
[57] Ibidem.

Medición Face2.0 Existente en la Argentina[58] un candidato electoral puede ser cualquier persona que dispute una elección electoral. En cambio, el Procedimiento Superador se enfoca solo en la campaña correspondiente a unas elecciones legislativas generales nacionales. En ese sentido el término *Candidato a Diputado*[59] representa a toda persona que dispute un cargo como Diputado y encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina. Respecto a la segunda comparación se resalta lo siguiente. En el marco de ambas acepciones del *Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina*[60] Fernández-Ardáiz, J., y Doria, A. (2011) indican que debe elegirse la cuenta que el candidato declare como oficial, la suya personal, señalando a las demás cuentas restantes, si existieran, como apócrifas. Dicha cuenta puede tratarse de un *Perfil privado*[61] o de un *Perfil público*[62]. En un perfil privado solo se admite el acceso a un total de 5.000 amigos o contactos. En un perfil público, también llamado *Página*[63] o *Fanpage*[64], se permite que un número ilimitado de personas tenga acceso a la información que se presenta en el mismo. Si los visitantes desean pueden cliquear sobre el botón *Me gusta esto*[65] y convertirse en fans o seguidores. Mientras, en el Procedimiento Superador se utiliza el término *Principal cuenta Facebook*[66]. De esa manera se hace referencia a la principal cuenta Facebook de los candidatos a Diputado dentro de las varias cuentas que los mismos puedan tener. En este estudio la referida cuenta principal también consta de las características recién atribuidas a una cuenta Facebook.

[58] La cursiva es nuestra.
[59] Ibidem.
[60] Ibidem.
[61] Ibidem.
[62] Ibidem.
[63] Ibidem.
[64] Ibidem.
[65] Ibidem.
[66] Ibidem.

4.6. Población de estudio

La población o universo se refiere al conjunto total de elementos de estudio que cumplen con las características de la unidad de análisis. En ese sentido de acuerdo a Hernández Sampieri y otros (2010) luego de definir la unidad de análisis se delimita la población de interés. "Las poblaciones deben situarse claramente en torno a sus características de contenido, de lugar y en el tiempo" (Hernández Sampieri y otros, 2010, p. 174).

En el Procedimiento de Medición Face2.0 Existente en la Argentina no se brinda una definición de población. Sin embargo se observa que sus acepciones, Funcionamiento de la e-campaña en la cuenta Facebook oficial/personal de un candidato electoral específico y Cantidad de personas que se incorpore a trabajar física y virtualmente en el proyecto político de un candidato electoral dado, restringen su población a un único candidato electoral. Por ende se concluye que el criterio *Población de estudio en el Procedimiento de Medición Face2.0 Existente en la Argentina*[67] goza de aval científico pues se asume que adopta la definición de Población descrita por Hernández Sampieri y otros (2010). Igualmente se concluye que es pertinente pues contempla a un solo candidato electoral al aplicar el procedimiento en cuestión. En tanto, en el marco del Procedimiento Superador la población estará compuesta por la principal cuenta Facebook de los candidatos a Diputado que encabecen la lista de su partido en un momento dado de unas elecciones legislativas generales nacionales de la Argentina. Así, se concluye que el criterio *Población de estudio en el Procedimiento Superador*[68] consta de aval científico pues define *Población*[69] según

[67] La cursiva es nuestra.
[68] Ibidem.

Hernández Sampieri y otros (2010). Igualmente se concluye que el referido criterio es pertinente ya que considera posibles casos a analizar a todos los que cumplan con las características de interés previamente establecidas.

4.7. Tamaño mínimo muestral

Conforme a Soper, D. (2006) para lograr estimaciones confiables a través del modelo de regresión se requiere determinar el tamaño mínimo que debe tener la muestra que se analizará. Soper (2006) indica que dicho tamaño mínimo se puede obtener a partir de los factores: Tamaño de efecto, potencia de la prueba, nivel de significancia y número de variables independientes que se pretenden considerar en el análisis.

Según Morales Vallejo, P. (2012) el tamaño de efecto consiste en la magnitud o importancia de la diferencia entre la media de dos grupos. Así, la diferencia de las medias puede resultar significativa pero no necesariamente grande o importante. Conforme a Morales Vallejo (2012) la fórmula básica del tamaño de efecto, d, se calcula dividiendo la diferencia de las medias de dos grupos entre la desviación estándar de uno de dichos grupos. Para interpretar el resultado Morales Vallejo (2012) señala que es común valerse de las orientaciones propuestas por Cohen quien propone lo siguiente:

- Efecto pequeño → En torno a d = 0,20.
- Efecto moderado → En torno a d = 0,50.
- Efecto grande → En torno a d = 0,80.

[69] La cursiva es nuestra.

En ese sentido cuando se desea calcular el tamaño mínimo muestral, el tamaño del efecto debe estimarse utilizando el criterio de 0,20 (efecto pequeño) a modo de garantizar un mayor número de casos.

En tanto, Bono, R., y Arnau, J. (1995) afirman que cuando se aplica una prueba estadística existe la posibilidad de cometer un error de Tipo II, o sea, de aceptar la H_0 cuando es falsa. Dicha probabilidad se denota 1-ß. Así, Bono y Arnau (1995) señalan

> (…) La potencia de una prueba es el complemento de la probabilidad de un error Tipo II (1-ß). (…) Cohen (1992) propone por convención una potencia de 0,80 (o sea, ß = 0,20). Un valor sustancialmente inferior a 0,80 implicaría un gran riesgo de incurrir en un error de Tipo II y un valor superior exigiría una muestra muy grande, fuera de los recursos del investigador (p. 194).

Igualmente Bono y Arnau (1995) establecen que también existe la posibilidad de cometer un error de Tipo I, es decir de rechazar la H_0 cuando es verdadera. Dicha probabilidad se denota 1- α. Bono y Arnau (1995) indican que se puede utilizar un valor de significancia de 0,05 a los fines de determinar el tamaño de la muestra.

En tanto, la cantidad de variables independientes dependerá de cuántas variables se pretende analizar en el modelo.

Así, para determinar el tamaño mínimo de la muestra se puede emplear los valores arriba descritos para tamaño de efecto, potencia de la prueba y nivel de significancia, en adición al número de variables independientes que se busque incluir en el modelo. Se puede fácilmente ingresar dichos valores a una calculadora virtual del tamaño de muestra para regresión y obtener el tamaño mínimo adecuado que se permite seleccionar. Se recomienda realizar tal cálculo utilizando la calculadora virtual que se

presenta en la página web de Daniel Soper a la cual se accede a través del enlace: http://www.danielsoper.com/statcalc3/calc.aspx?id=1. Cabe destacar que The University of Texas at Austin (s. f.) indica que se debe procurar una relación de aproximadamente 15 a 20 casos por variable independiente. Esto se cumple al calcular el referido tamaño mínimo en base a los valores recomendados, especialmente cuando se considera 1 solo predictor pues en ese escenario se plantea estudiar como mínimo 41 casos. En tanto, según Wilderdom (s. f.) Francis (2007) establece que idealmente la muestra debe tener un tamaño mínimo de al menos 100 casos. No obstante, dicha postura no es inflexible. Así, si no se puede cumplir con estudiar por lo menos 100 casos se puede recurrir al cálculo del tamaño mínimo muestral que se establece al inicio de este apartado.

De esta manera, se observa que en el caso del Procedimiento de Medición Face2.0 Existente en la Argentina no corresponde calcular un tamaño mínimo de muestra pues el mismo no contempla análisis de regresión alguno. Así, respecto a dicho procedimiento no se emite conclusión respecto a la viabilidad del criterio *Tamaño mínimo muestral del Procedimiento de Medición Face2.0 Existente en la Argentina*[70]. Por el contrario, en el Procedimiento Superador sí. Por ende en cuanto a la viabilidad del criterio *Tamaño mínimo muestral para realizar el análisis de regresión en el Procedimiento Superador*[71] se concluye que el mismo consta de aval científico. Eso se debe a que dicho criterio se basa en lineamientos científicamente reconocidos. Igualmente se concluye que el referido criterio es pertinente dado a que cuando se realiza un análisis de regresión corresponde determinar el tamaño mínimo de la muestra según las especificaciones descritas en este aparatado.

[70] La cursiva es nuestra.
[71] Ibidem.

4.8. Tamaño de la muestra

De acuerdo a Pita Fernández (1996) cuando se desea estimar una media y la población es finita, o sea que se conoce el total de la misma, la fórmula para calcular el tamaño de muestra es:

$$n = \frac{N * Z_{\alpha}^2 * S^2}{d^2 * (N-1) + Z_{\alpha}^2 * S^2}$$

Pita Fernández (1996) indica que en dicha fórmula:

- N = Cantidad de elementos de la población.

- Z = Nivel de confianza. Los valores más comunes son 95% que equivale a z = 1,96 y 99% que es equivalente a z = 2,58. Esos valores Z críticos se obtienen de las tablas del área de la curva normal.

- S^2 = Varianza de la variable principal. En el marco del Procedimiento Superador si la misma es desconocida se puede considerar obtenerla de estudios previos o de lo contrario se utiliza S^2 = 0,25 con lo que se maximiza el tamaño muestral.

- d = Precisión deseada para el estudio. Es común utilizar 0,03 (equivalente a 3%) ó 0,05 (igual a 5%).

Fernández-Ardáiz (2010) no indica en la descripción del Indice 2.0 cómo calcular el tamaño de la muestra. Sin embargo tampoco es necesario que lo haga ya que el Procedimiento de Medición Face2.0 Existente en la Argentina implica un censo. Por eso no se establece una conclusión en torno a la viabilidad del criterio *Tamaño de la muestra del Procedimiento de Medición Face2.0 Existente en la Argentina*[72]. No obstante, la fórmula presentada al inicio de este apartado es aceptada en el contexto de la ciencia y es la utilizada en el Procedimiento Superador. Por ende en cuanto a la viabilidad del criterio *Tamaño de la muestra en el Procedimiento Superador*[73] se concluye que el mismo goza de aval científico. Asimismo, se concluye que el referido criterio es pertinente ya que un estudio como el presente cumple las especificaciones para emplear la fórmula que según Pita Fernández (1996) plantea. Al respecto, el Procedimiento Superador:

- Se basa en una población finita compuesta por la principal cuenta Facebook de los candidatos a Diputado que encabecen la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina.

- Consta de una variable principal cuantitativa la cual se denomina Exito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina (el porcentaje de la intención de voto que es explicado por el modelo que se construya).

[72] La cursiva es nuestra.
[73] Ibidem.

4.9. Tipo de muestra

Según Hernández Sampieri y otros (2010) las muestras se pueden clasificar en dos grupos: las muestras no probabilísticas y las muestras probabilísticas. "En las muestras no probabilísticas la elección de los elementos no depende de la probabilidad (…) sino que depende del proceso de toma de decisiones de un investigador o de un grupo de investigadores" (Hernández Sampieri y otros, 2010, p. 176). "[En cambio] en las muestras probabilísticas todos los elementos de la población tienen la misma probabilidad de ser escogidos (…) por medio de una selección aleatoria (…) de las unidades de análisis" (Hernández Sampieri y otros, 2010, p. 176). Conforme a Hernández Sampieri y otros (2010) las muestras probabilísticas son clave en los diseños de investigación transeccionales, tanto descriptivos como correlacionales-causales, en los que se desea estimar ciertas variables en la población. Según Hernández Sampieri y otros (2010) dichas variables se miden y analizan con pruebas estadísticas en una muestra que se supone que es probabilística. De acuerdo a Hernández Sampieri y otros (2010) para lograr que una muestra sea probabilística se necesita calcular un tamaño de muestra representativo de la población y seleccionar los casos, también denominados elementos muestrales, de modo que en principio todos tengan la misma posibilidad de ser elegidos. En el Procedimiento Superador se emplea el tipo de muestra denominado *Muestra probabilística estratificada*[74]. Conforme a Hernández Sampieri y otros (2010) se trata una muestra probabilística en la que se considerará segmentos o grupos de la población, también llamados estratos. Por ejemplo, Hernández Sampieri y otros

[74] La cursiva es nuestra.

(2010) señalan que estratos de la variable grado de nivel de estudios pueden ser preescolar, primaria, secundaria, bachillerato, universidad (o equivalente) y posgrado.

Según Hernández Sampieri y otros (2010)

> Kish (1995) afirma que, en un número determinado de elementos muestrales $n = \sum nh$, la varianza de la media muestral (…) [de y] puede reducirse al mínimo si el tamaño de la muestra para cada estrato es proporcional a la desviación estándar dentro del estrato.
>
> Esto es,
>
> $$\sum fh = n/N = ksh$$
>
> En donde la muestra n será igual a la suma de los elementos muestrales nh. Es decir, el tamaño de n y la varianza de (…) [la media de y] pueden minimizarse si calculamos "submuestras" proporcionales a la desviación estándar de cada elemento en un determinado estrato. Esto es:
>
> $$fh = nh/Nh = ksh$$
>
> En donde nh y Nh son muestra y población de cada estrato, y sh es la desviación estándar de cada elemento en un determinado estrato. Entonces tenemos que:
>
> $$ksh = n/N$$
>
> (…) De esta manera el total de la subpoblación se multiplicará por esta fracción constante para obtener el tamaño de la muestra para el estrato. Al sustituirse tenemos que (p. 181):
>
> $$(Nh)\,(fh) = nh$$

Para representar la aplicación de ese tipo de muestra en un ejemplo Hernández

Sampieri y otros (2010) proponen

> (…) Supongamos que pretendemos realizar un estudio con directores de recursos humanos para determinar su ideología y política respecto a cómo tratan a los colaboradores de sus empresas. Imaginamos que nuestro universo es de 1.176 organizaciones con directores de recursos humanos. (…) Determinamos que el tamaño de la muestra necesaria para representar a la población sería de n = 298 directivos. Pero supongamos que la situación se complica y que debemos estratificar esta n con la finalidad de que los elementos muestrales o las unidades de análisis posean un determinado atributo. En nuestro ejemplo este atributo podría ser el giro de la empresa. (…) ¿Qué muestra necesitaremos para cada estrato? (p. 181)

$$ksh = n/N = 298/1.776 = 0{,}2534$$

La tabla 1 muestra el total de casos que se debería relevar según cada estrato. En la misma, por ejemplo, se observa que conforme a Hernández Sampieri y otros (2010) el total de directores de empresas extractivas es igual a 53 directivos. Según Hernández Sampieri y otros (2010), dado a que la fracción constante es de 0,2534, la cantidad de casos que corresponde relevar en ese estrato es de 13 directores.

Tabla 1. Casos a relevar por estrato en el ejemplo de los directivos de recursos humanos[75]

Estrato por giro	**Directores recursos humanos del giro**	**Total población (*fh*) = 0,2534 (*Nh*) (*fh*) = *nh***	**Muestra**
1	Extractivo y siderúrgico	53	13
2	Metal-mecánicas	109	28
3	Alimentos, bebidas y tabaco	215	55
4	Papel y artes gráficas	87	22
5	Textiles	98	25
6	Eléctricas y electrodomésticas	110	28
7	Automotriz	81	20
8	Químico-farmacéutica	221	56
9	Otras empresas de transformación	151	38
10	Comerciales	51	13

Fernández-Ardáiz (2010) en la descripción que hace del Indice 2.0 no señala que apela a un tipo de muestra dado pues no trabaja con población. En ese sentido no se establece una conclusión respecto a la viabilidad del criterio *Tipo de muestra del Procedimiento de Medición Face2.0 Existente en la Argentina*[76]. En cambio, en el Procedimiento Superador se opera en base a una muestra probabilística estratificada. Así, queda a criterio de quien dirige la investigación definir los estratos en función de los subgrupos que considere importante para el estudio. Por ende en cuanto a la viabilidad del criterio *Tipo de muestra del Procedimiento Superador*[77] se concluye que el mismo consta de aval científico ya que recurre a un tipo de muestra científicamente aprobado. Igualmente se concluye que el mencionado criterio es

[75] Adaptado de *Metodología de la investigación*, (p. 182), por R. Hernández Sampieri, C. Fernández Collado y P. Baptista Lucio, 2010, Perú: McGraw-Hill. Derechos reservados 2010 por McGraw-Hill/Interamericana Editores, S.A. de C.V.
[76] La cursiva es nuestra.
[77] Ibidem.

pertinente debido a que la población de estudios como el presente se puede segmentar en subgrupos a fin de obtener una muestra probabilística.

4.10. Selección de los elementos muestrales

Respecto a la *Selección sistemática de elementos muestrales*[78] Hernández Sampieri y otros (2010) establecen

> Este procedimiento (…) implica elegir dentro de una población *N* un número de elementos a partir de un intervalo *K*. Este último (*K*) es un intervalo que se va a determinar por el tamaño de la población y el tamaño de la muestra. De manera que tenemos que $K = N/n$, en donde *K* = un intervalo de selección sistemática, *N* = la población y *n* = la muestra.
>
> Ilustremos los conceptos anteriores con un ejemplo. Supongamos que se quiere hacer un estudio que pretende medir la calidad de la atención en los servicios proporcionados por los médicos y las enfermeras de un hospital. Para tal efecto consideremos que los investigadores consiguen grabaciones de todos los servicios efectuados durante un período determinado. Supongamos que se hayan filmado 1.548 servicios (*N*). (…) Determinamos que se necesitan 307,9 (308) servicios para evaluar (con un error máximo de 5%, nivel de confianza de 95% y un porcentaje estimado de 50% para la muestra).
>
> Si necesitamos una muestra de *n* = 308 episodios de servicio filmados, se utiliza para la selección el intervalo *K*, donde:
>
> $$K = N/n = 1.548/308 = 5{,}0259, \text{ redondeado} = 5$$
>
> El intervalo $1/K = 5$ indica que cada quinto servicio $1/K$ se seleccionará hasta completar *n* = 308. La selección sistemática de elementos muestrales $1/K$ se puede utilizar al elegir los elementos de *n* para cada estrato (…).

[78] La cursiva es nuestra.

> (…) Siguiendo nuestro ejemplo, no comenzaremos a elegir de los 1.548 episodios, el 1, 6, 11, 16…, sino que procuramos que el inicio sea determinado por el azar. Así, en este caso, podemos tirar unos dados y si en sus caras muestran 1, 6, 9, iniciaremos en el servicio 169, y seguiremos 174, 179, 184, 189… 1/*K*… y volveremos a empezar por los primeros si es necesario (pp. 184 - 185).

Fernández-Ardáiz (2010) en la explicación que hace del Indice 2.0 no se refiere al modo de selección de los elementos muestrales ya que opera con población. En ese sentido no se establece una conclusión respecto al criterio *Selección de los elementos muestrales en el Procedimiento de Medición Face2.0 Existente en la Argentina*[79]. En cambio, en el Procedimiento Superador los elementos de la muestra se eligen a partir de la *Selección sistemática de elementos muestrales*[80] según las especificaciones realizadas por Hernández Sampieri y otros (2010). Por ende se concluye que el criterio *Selección de los elementos muestrales en el Procedimiento Superador*[81] goza de aval científico. Asimismo, se concluye que el mencionado criterio es pertinente en estudios como el presente se puede elegir los elementos muestrales en base a una selección sistemática sin inconvenientes.

4.11. Técnica de análisis de datos

Como se mencionó en el apartado *Exito de la e-campaña vía Facebook*[82] el término *Exito de la e-campaña vía Facebook en el Procedimiento de Medición Face2.0 Existente en la Argentina*[83] tiene dos acepciones. Así, la denominada en este estudio *Técnica de análisis de datos del Procedimiento de Medición Face2.0 Existente en la*

[79] La cursiva es nuestra.
[80] Ibidem.
[81] Ibidem.
[82] Ibidem.
[83] Ibidem.

Argentina[84] no se restringe a una sola técnica. Puede variar según la acepción que se desee investigar del mencionado éxito. Como dicho término proviene del concepto *Exito de la e-campaña*[85] propio del Indice de medición 2.0 conviene señalar las variantes en el contexto de ese índice. Al terminar de abordar lo relativo a tales variantes se resume dicha información en los términos equivalentes y propios del *Procedimiento de Medición Face2.0 Existente en la Argentina*[86]. En ese sentido Fernández-Ardáiz (2010) indica que la técnica de análisis de datos utilizada para medir el funcionamiento/rendimiento de la e-campaña consiste en un indicador. En esta investigación el mismo se denomina *Indicador de funcionamiento de la e-campaña*[87]. Dicha técnica de análisis de datos cuenta con los denominados *Criterios de aplicación del indicador de funcionamiento de la e-campaña*[88]. Los mismos se componen por los criterios llamados en este estudio *Elección inicial de las variables indicadoras*[89] y *Ponderación de las variables indicadoras*[90]. En cuanto al primer criterio, aunque no indica cómo, Fernández-Ardáiz (2010) dice que en cada una de las herramientas en las que participe el candidato que se desea investigar debe determinarse cuáles variables han de medirse para conocer el Exito de la e-campaña. En referencia al segundo criterio Fernández-Ardáiz (2010) señala que las variables a estudiar deben ponderarse en función de un sistema de ponderación. A pesar de que Fernández-Ardáiz (2010) tampoco detalla dicho sistema indica que el mismo genera como resultado un solo número entre todas las herramientas. Un intento para explicar esa medición se observa en los estudios *Indice 2.0 Diputados 2010*[91] e *Indice*

[84] La cursiva es nuestra.
[85] Ibidem.
[86] Ibidem.
[87] Ibidem.
[88] Ibidem.
[89] Ibidem.
[90] Ibidem.
[91] Ibidem.

Senadores Nacionales 2011[92]. Conforme a Fernández-Ardáiz, J., y Doria, A. (2010) y Fernández-Ardáiz y Doria (2011) se emplea un sistema que pondera las distintas variables estudiadas con valores que oscilan entre 1 y 4. Según Fernández-Ardáiz y Doria (2010) y Fernández-Ardáiz y Doria (2011) en el mismo el puntaje 1 equivale a las respuestas más negativas y 4 a las más positivas. De acuerdo a Fernández-Ardáiz y Doria (2011) el mencionado valor final se obtiene dividiendo la sumatoria de los puntajes de cada pregunta entre la cantidad total de preguntas, logrando el promedio. Fernández-Ardáiz y Doria (2010) indican que ese índice 2.0 final debe oscilar en un determinado rango para que una e-campaña se considere exitosa. Según Fernández-Ardáiz y Doria (2011) dicho rango estaría constituido por valores comprendidos entre 3 y 4. Sin embargo el mencionado sistema de ponderación queda poco claro al considerar cómo se ponderaría las distintas temáticas que Fernández-Ardáiz y Doria (2011) propone analizar. En ese sentido cabe destacar que la batería de temas que Fernández-Ardáiz y Doria (2011) indagan en la investigación *Indice Senadores Nacionales 2011*[93] es distinta a la del estudio *Indice 2.0 Diputados 2010*[94]. Fernández-Ardáiz y Doria (2011) no aclaran a qué se debe la diferencia pero no mencionan que el cuestionario utilizado en *Indice Senadores Nacionales 2011*[95] reemplace al del *Indice 2.0 Diputados 2010*[96]. No obstante, ya que la versión utilizada en el *Indice Senadores Nacionales 2011*[97] es la más reciente de dicho cuestionario se asume que actualmente es la versión oficial. Al respecto, Fernández-Ardáiz y Doria (2011) señalan que se empleó un cuestionario cerrado y con opciones. Ajustando las preguntas del mismo al marco del Procedimiento de Medición Face2.0 Existente en la

[92] La cursiva es nuestra.
[93] Ibidem.
[94] Ibidem.
[95] Ibidem.
[96] Ibidem.
[97] Ibidem.

Argentina se observa que en cada cuenta Facebook analizada interesaría saber las siguientes temáticas:

1) Si el Senador tiene cuenta Facebook

2) Si el Senador responde las consultas de los ciudadanos
   (se envió una consulta por mensaje privado de Facebook)

3) Cuántos seguidores o amigos tiene el Senador

4) Si el Senador publica información

5) Con qué periodicidad el Senador publica información

6) Si el Senador utiliza material audiovisual

7) En qué modo el Senador utiliza el material audiovisual

En ese sentido hay dos aspectos a resaltar. Por un lado, ya que Fernández-Ardáiz y Doria (2011) no presentan el cuestionario utilizado se desconoce cómo se construyó el conjunto de respuestas posibles. Por otro lado, se aprecia que las temáticas 3, 5 y 7 podrían mostrar cuatro respuestas pudiendo cada una de las mismas ser representada por un valor entre 1 y 4. No obstante, las opciones de respuesta de las temáticas 1, 2, 4 y 6 son Sí o No. A pesar de eso Fernández-Ardáiz y Doria (2011) no especifican qué valor de la escala 1 a 4 le correspondería a dichas respuestas. Por ende se

concluye que se desconoce cómo aplicar exhaustivamente el sistema de ponderación del Indice de medición 2.0. Mientras, el resumen equivalente aplicado al Procedimiento de Medición Face2.0 Existente en la Argentina es el siguiente. En este último la técnica de análisis de datos empleada para medir el Funcionamiento de la e-campaña en la cuenta Facebook oficial/personal de un candidato electoral puntual se llama *Indicador de funcionamiento de la e-campaña vía Facebook*[98]. La misma está compuesta por los denominados *Criterios de aplicación del indicador de funcionamiento de la e-campaña vía Facebook*[99]. Dichos criterios se denominan respectivamente *Elección inicial de las variables indicadoras en el Procedimiento de Medición Face2.0 Existente en la Argentina*[100] y *Ponderación de las variables indicadoras en el Procedimiento de Medición Face2.0 Existente en la Argentina*[101]. Las mismas situaciones descritas en el marco del Indice 2.0 aplican al contexto del Procedimiento de Medición Face2.0 Existente en la Argentina.

En cambio, la técnica de análisis de datos utilizada en el contexto de la acepción Cantidad de personas que se incorpore a trabajar física y virtualmente en el proyecto político de un candidato electoral específico se denomina en este estudio *Conteo de voluntades sumadas al proyecto político*[102]. Fernández-Ardáiz (2010) hace una mención poco específica de dicha técnica de análisis de datos. Así, Fernández-Ardáiz (2010) indica: "Una aceitada y correcta gestión de los contactos redundará en un preciso análisis de la cantidad de voluntades que se sumaron al proyecto político que estamos comunicando". En ese contexto se desconoce por un lado cuáles son los

[98] La cursiva es nuestra.
[99] Ibidem.
[100] Ibidem.
[101] Ibidem.
[102] Ibidem.

Criterios de aplicación del Conteo de voluntades sumadas al proyecto político[103] y por otro lado cómo implementar el referido conteo.

En tanto, las conclusiones respecto a la viabilidad del criterio *Técnica de análisis de datos del Procedimiento de Medición Face2.0 Existente en la Argentina*[104] se emiten en función de las técnicas correspondientes a cada acepción del Exito de la e-campaña vía Facebook. En el caso de la acepción Funcionamiento de la e-campaña en la cuenta Facebook oficial/personal de un candidato electoral puntual se concluye en base a los llamados Criterios de aplicación del indicador de funcionamiento de la e-campaña vía Facebook. Por un lado, la Elección inicial de las variables indicadoras en el Procedimiento de Medición Face2.0 Existente en la Argentina. Por otro lado, la Ponderación de las variables indicadoras en el Procedimiento de Medición Face2.0 Existente en la Argentina. En cuanto al primero se concluye que el mismo es científico pues en el marco de la ciencia se admite elegir bajo cualquier criterio las variables a analizar. Igualmente se concluye que el mencionado criterio es pertinente ya que permite al encargado del estudio seleccionar las variables que entienda necesarias de una manera inobjetable por la ciencia. Respecto al segundo se concluye que el aval científico del mismo es reservado ya que el sistema de ponderación no se detalla exhaustivamente y quedan dudas sobre la rigurosidad científica de su aplicación. Sin embargo el referido criterio no es pertinente pues el Exito de la e-campaña vía Facebook debe determinarse en función de la relación de dependencia entre variables, objetivo que no puede lograrse en base a un sistema de ponderación. En tanto, se concluye que la técnica de análisis de datos llamada Indicador del

[103] La cursiva es nuestra.
[104] Ibidem.

funcionamiento de la e-campaña vía Facebook cuenta con un aval científico reservado y no es pertinente.

En el caso de la acepción Cantidad de personas que se incorpore a trabajar física y virtualmente en el proyecto político de un candidato electoral en particular la situación es distinta. Aunque se desconoce cuáles son los criterios de aplicación de la técnica denominada *Conteo de voluntades sumadas al proyecto político*[105] se concluye que la misma goza de aval científico. Eso se debe a que dicha técnica de análisis de datos se basa en la operación matemática denominada *Suma*[106]. No obstante, se concluye que el mencionado criterio no es pertinente. Esto se debe a que el Exito de la e-campaña vía Facebook se debe establecer en función de la relación de dependencia entre variables y no a partir de la adición de personas al trabajo del proyecto político de un candidato dado.

En cambio, las técnicas de análisis de datos del Procedimiento Superador son *Entrevista en profundidad*[107], *Regresión lineal*[108] y *Regresión intrínsecamente lineal*[109]. En ese sentido respecto a la entrevista en profundidad Varguillas Carmona, y Ribot de Flores (2007) afirman

> Es una técnica para recopilar información sobre conocimientos, creencias, rituales de una persona (…). Se caracteriza por una conversación personal (…), no estructurada, en la que se persigue que el entrevistado exprese de forma libre sus opiniones, actitudes, o preferencias sobre el tema objeto de estudio. De esta manera se concibe como una interacción social entre dos personas (…), [en la que

[105] La cursiva es nuestra.
[106] Ibidem.
[107] Ibidem.
[108] Ibidem.
[109] Ibidem.

el entrevistado] (…) va a explicar su visión del tema (…) y [el entrevistador] (…) va a tratar de comprender o interpretar esa explicación.

En cuanto a la regresión lineal, Universidad Complutense de Madrid (s. f.) la define como una técnica que determina si una o varias variables predice (n) a otra. "[En ese sentido] (…) a la variable que se [intenta] predecir se le llama *Variable dependiente*[110] [o variable respuesta]. A la variable (…) que se [usa] para predecir (…) la variable dependiente se [le] llama [*Variable independiente*[111]]" (Anderson, Sweeney, y Williams, 2008, p. 545). En el presente estudio se aborda dos formas de regresión lineal, es decir, la *Regresión lineal simple*[112] y la *Regresión múltiple*[113]. En el primer caso, una sola variable independiente métrica predice a otra dependiente igualmente métrica. En el segundo caso, conforme a Pérez López, C. (2005) se analiza la relación entre una variable dependiente métrica y varias variables independientes también métricas a fin de predecir la citada variable dependiente. Del mismo modo "La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas" (Pérez López, 2005, p. 89). Esto es recodificar las variables independientes de manera binaria e incorporarlas al análisis como si fueran variables numéricas. De esa manera el Procedimiento Superador se presta para determinar en la principal cuenta Facebook de un candidato a Diputado Nacional que encabece la lista de su partido cuáles son las variables métricas y no métricas que influyen en la intención de voto en un momento específico de una campaña legislativa general nacional de la Argentina. Cabe destacar que es posible que un modelo lineal en ocasiones no sea apto para satisfacer los supuestos de regresión los cuales se aborda en el apartado

[110] La cursiva es nuestra.
[111] Ibidem.
[112] Ibidem.
[113] Ibidem.

Supuestos del modelo en regresión lineal simple[114] y *Supuestos del modelo en regresión múltiple*[115]. Sin embargo se puede recurrir a la aplicación de una *Regresión intrínsecamente lineal*[116] para intentar solucionar esa situación a través de transformaciones realizadas a la variable dependiente, a la (s) variable (s) independiente (s), o a una combinación de ambos tipos de variable. En el apartado *Incumplimiento de los supuestos de regresión y uso de transformaciones*[117] se profundiza sobre las referidas transformaciones.

En tanto, las técnicas de análisis de datos correspondiente al Procedimiento Superador cuentan con criterios de aplicación. En el caso de la entrevista en profundidad se aprecia el criterio denominado en esta investigación *Tipo de entrevista*[118]. Como mencionan Varguillas Carmona y otros (2007) en la descripción de la entrevista en profundidad, la misma debe ser no estructurada de forma que el entrevistado se exprese libremente sobre el tema tratado. En cambio, en el caso de los criterios de aplicación de la regresión lineal e intrínsecamente lineal destacan varios criterios de aplicación. Los mismos se detallan más adelante en los apartados *Análisis de regresión lineal simple*[119] y *Análisis de regresión múltiple*[120].

En cuanto a la viabilidad del criterio *Técnicas de análisis de datos del Procedimiento Superador*[121] se concluye que las tres técnicas gozan de aval científico. Esto se debe a que tanto la entrevista en profundidad, la regresión lineal y la regresión intrínsecamente lineal constituyen técnicas científicamente reconocidas. Igualmente

[114] La cursiva es nuestra.
[115] Ibidem.
[116] Ibidem.
[117] Ibidem.
[118] Ibidem.
[119] Ibidem.
[120] Ibidem.
[121] Ibidem.

se concluye que el mencionado criterio es pertinente pues la entrevista en profundidad ofrece la oportunidad de considerar variables que expertos consideren relevantes y tanto la regresión lineal como la intrínsecamente lineal permiten determinar si las variables independientes que se analice influyen en la intención de voto con lo cual se determina si la e-campaña es exitosa.

A partir de las comparaciones realizadas se presenta la tabla 2 a modo de resumen.

Tabla 2. Resumen de la viabilidad de los criterios de aplicación de ambos procedimientos[122]

Procedimiento Superador		**Criterio de aplicación**	**Procedimiento de Medición Face2.0 Existente en la Argentina**	
Viabilidad			**Viabilidad**	
Pertinencia	**Aval científico**		**Aval científico**	**Pertinencia**
Sí	Sí	Exito de la e-campaña vía Facebook	No	No
Sí	Sí	Diseño de investigación	Sí	No
Sí	Sí	Uso de muestra o población en ambos procedimientos	Sí	Sí
Sí	Sí	Muestra	-	-
Sí	Sí	Unidad de análisis	Sí	No
Sí	Sí	Población de estudio	Sí	Sí
Sí	Sí	Tamaño mínimo muestral	-	-
Sí	Sí	Tamaño de la muestra	-	-
Sí	Sí	Tipo de muestra	-	-
Sí	Sí	Selección de los elementos muestrales	-	-
-	-	Técnica de análisis de datos “Indicador de funcionamiento de la e-campaña vía Facebook”	Reservado	No
-	-	Técnica de análisis de datos “Conteo de voluntades sumadas al proyecto político”	Sí	No
Sí	Sí	Técnica de análisis de datos “Entrevista en profundidad”	-	-
Sí	Sí	Técnica de análisis de datos “Regresión lineal”	-	-
Sí	Sí	Técnica de análisis de datos “Regresión intrínsecamente lineal”	-	-

[122] Elaboración propia.

5. Viabilidad del Procedimiento de Medición Face2.0 Existente en la Argentina y del Procedimiento Superador

Como se observa en el apartado anterior[123] quedan múltiples dudas respecto a los criterios de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina. En la explicación que Fernández-Ardáiz (2010) hace del Indice 2.0 no se encuentra respuesta a las mismas. Tampoco en el estudio que realizó Fernández-Ardáiz y Doria (2010) titulado Indice 2.0 Diputados 2010 ni en el realizado por Fernández-Ardáiz y Doria (2011) denominado Indice Senadores Nacionales 2011. En ese sentido en septiembre de 2011 coordiné con José Fernández-Ardáiz una entrevista a través de Skype con el fin de aclarar las referidas inquietudes. Sin embargo no pude entrevistarlo a pesar de haber intentado el contacto repetidas ocasiones desde septiembre hasta diciembre de 2011. Intenté contactarlo tanto a través de su correo electrónico como por medio de su cuenta Facebook. Tampoco funcionó, aunque me consta que revisa y utiliza su cuenta Facebook frecuentemente debido a los comentarios que publica en la misma. Le garanticé que se trataba de una entrevista breve y que podíamos realizarla por Skype o por teléfono ya sea celular o fijo. A pesar de que en una oportunidad me envió un mail excusándose por haber demorado unos meses en responder uno de los mensajes que le dejé, hasta Agosto de 2014 aún lo esperaba para entrevistarlo. Ante ese contexto se procede a evaluar la viabilidad del Procedimiento de Medición Face2.0 Existente en la Argentina.

De esa manera a partir de la tabla 2 se establece que el criterio de aplicación del Procedimiento de Medición Face2.0 Existente en la Argentina denominado Exito de

[123] El mencionado apartado anterior se titula Comparación y viabilidad de los criterios de aplicación tanto del Procedimiento de Medición Face2.0 Existente en la Argentina como del Procedimiento Superador.

la e-campaña vía Facebook no goza de aval científico. Igualmente respecto a dicho procedimiento se indica que la técnica de análisis de datos llamada *Indicador de funcionamiento de la e-campaña vía Facebook*[124] cuenta con un aval científico reservado. Asimismo, se señala que no son pertinentes los siguientes criterios de aplicación del procedimiento en cuestión:

- Exito de la e-campaña vía Facebook
- Diseño de investigación
- Unidad de análisis
- Técnica de análisis de datos (en ninguna de las acepciones del procedimiento)

Por ende se concluye que el Procedimiento de Medición Face2.0 Existente en la Argentina no es viable. En cambio, ya que todos los criterios de aplicación del Procedimiento Superador constan de aval científico y son pertinentes se concluye que dicho procedimiento es viable.

[124] La cursiva es nuestra.

6. Criterios de aplicación de las técnicas de análisis de datos del Procedimiento Superador

Como se mencionó en el apartado *Técnicas de análisis de datos*[125] los criterios de aplicación de dichas técnicas en el marco del Procedimiento Superador, excepto los correspondientes a la técnica Entrevista en profundidad (presentados en el referido apartado), se abordan a continuación en los apartados *Análisis de regresión lineal simple*[126] y *Análisis de regresión múltiple*[127]. Dado a la extensión de dichos apartados se ha optado por tratar los mencionados criterios después de establecer las conclusiones respecto a la viabilidad tanto del Procedimiento de Medición Face2.0 Existente en la Argentina como del Procedimiento Superador.

7. Análisis de regresión lineal

Como se mencionó en el apartado *Técnica de análisis de datos*[128] según Levine, D. M., Krehbiel, T. C., y Berenson, M. L. (2006) la regresión lineal es una técnica que indica si una o varias variables predicen a otra. "[En ese sentido] (…) a la variable que se va a predecir se le llama *Variable dependiente*[129]. A la variable o variables que se usan para predecir el valor de la variable dependiente se les llama *Variables independientes*[130]" (Anderson, Sweeney, y Williams, 2009, p. 545). Según Levine y otros (2006) a la variable dependiente también se le llama *Variable de respuesta*[131] y

[125] La cursiva es nuestra.
[126] Ibidem.
[127] Ibidem.
[128] Ibidem.
[129] Ibidem.
[130] Ibidem.
[131] Ibidem.

a las independientes *Variables explicatorias*[132].

7.1. Análisis de regresión lineal simple

Según Tacq, J. (1998) el análisis de regresión lineal simple también se conoce como *Análisis de regresión bivariado*[133]. De acuerdo a Anderson y otros (2009) por medio de esta técnica se estudia la relación lineal de una variable independiente llamada x_1 con una sola variable dependiente denominada *y*. Tacq (1998) señala que la relación entre dichas variables se puede representar en un sistema de coordenadas cartesiano con eje horizontal x_1 y eje vertical *y*. De acuerdo a Tacq (1998) cada punto o coordenada representa un caso, al cual le corresponde un valor x_1 y otro *y*. "Cuando dibujamos (…) [los pares (x_1,*y*)] en el sistema de coordenadas (…) [se obtiene] una nube de puntos llamada *Diagrama de dispersión*[134]" (Tacq, 1998, p. 101). Igualmente Tacq (1998) indica que, en el marco de una población, la relación lineal entre estas dos variables también se puede representar a través de un modelo matemático. Anderson y otros (2009) denominan al modelo en cuestión *Modelo de regresión lineal simple*[135] el cual se expresa como sigue:

$$y = \beta_0 + \beta_{y1}x_1 + \varepsilon$$

Según Tacq (1998) en esta fórmula *y* es la variable dependiente, β_0 es el intercepto, β_{y1} es el coeficiente de regresión y ε es el término del error. Tacq (1998) establece: "(…) [El intercepto β_0 constituye] el valor de *y* cuando x_1 es igual a cero" (p.100).

[132] La cursiva es nuestra.
[133] Ibidem.
[134] Ibidem.
[135] Ibidem.

Igualmente Tacq (1998) señala: "[El coeficiente de regresión β_{y1} representa] el cambio en *y* por unidad incrementada [o disminuida] en x_1" (p. 100). En tanto, Anderson y otros (2009) indican que el término de error ε da cuenta de la variabilidad de *y* que no puede ser explicada por la relación lineal entre x_1 y *y*. En ese contexto Anderson y otros (2009) señalan que x_1 puede interpretarse en función de dos modalidades. Así, Anderson y otros (2009) indican que según la primera x_1 puede reflejar el valor de un determinado caso de la población. Igualmente Anderson y otros (2009) establecen que conforme a la segunda x_1 puede representar el valor de una subpoblación de casos de la población total.

Un concepto asociado al referido modelo es la denominada *Ecuación de regresión lineal simple*[136]. De acuerdo a Anderson y otros (2009) se trata de la ecuación que describe la relación lineal de x_1 y *y* sin la consideración del término del error. Según Anderson y otros (2009) la misma se formula de la siguiente manera:

$$E(y) = \beta_0 + \beta_{y1} x_1$$

Anderson y otros (2009) afirman

> [En dicha ecuación] β_0 es la intersección de la recta de regresión con el eje *y*, β_{y1} es la pendiente y *E(y)* [puede representar el valor *y* para un x_1 dado o el promedio de las *y* para un valor x_1 común a un subconjunto de la población] (p. 546).

Asimismo, Anderson y otros (2009) señalan que el gráfico de la mencionada ecuación constituye una línea recta que se denomina *Recta de regresión*[137].

[136] La cursiva es nuestra.
[137] Ibidem.

Cabe destacar que según Anderson y otros (2009), tanto en el modelo de regresión como en la ecuación de regresión mencionados, β_0 y β_{y1} constituyen parámetros. Ya que una investigación poblacional considera todos los casos del universo de estudio los mencionados parámetros se pueden determinar a través de un censo. Según Anderson y otros (2009) si los mismos se conocieran se podría usar la ecuación [de regresión lineal simple] para calcular *E(y)*. En otras palabras, se pudiera obtener el valor *y* para un x_1 dado o el promedio de las *y* para un valor x_1 común a un subconjunto de la población. En cambio, según Anderson y otros (2009) al realizar un estudio muestral β_0 y β_{y1} son desconocidos y se estiman a través de los estadísticos muestrales b_0 y b_{y1}. Asimismo, de acuerdo a Anderson y otros (2009) ε representa una variable aleatoria. La misma también puede calcularse considerando datos poblacionales o muestrales. En este último caso, conforme a Tacq (1998) el término del error pasa a expresarse como *e* pues así se le denota cuando se hacen estimaciones. De esta manera Tacq (1998) señala que cuando se realiza un estudio muestral el modelo de regresión lineal simple se formula como sigue:

$$y = b_0 + b_{y1}x_1 + e$$

7.1.1. Ecuación de regresión estimada en regresión lineal simple

Anderson y otros (2009) establecen: "[Al omitir el término del error y sustituir] en la ecuación de regresión [β_0 y β_{y1}] por los valores de los estadísticos muestrales [b_0 y b_{y1}] se obtiene la *Ecuación de regresión estimada*[138]" (pp. 546-547). La misma de acuerdo a Anderson y otros (2009) es:

[138] La cursiva es nuestra.

$$\hat{y} = b_0 + b_{y1}x_1$$

Según Anderson y otros (2009) en esta fórmula la expresión $\hat{y}$ se denomina *Valor estimado de y*[139] o *Valor pronosticado de y*[140]. Tacq (1998) prefiere llamarla *y estimada*[141] o *y sombrero*[142], mientras que Levine y otros (2006) se refieren a la misma como *Valor predicho de y*[143]. A los fines de este estudio todas esas denominaciones de $\hat{y}$ son aceptadas y se consideran sinónimo.

En tanto, Anderson y otros (2009) establecen

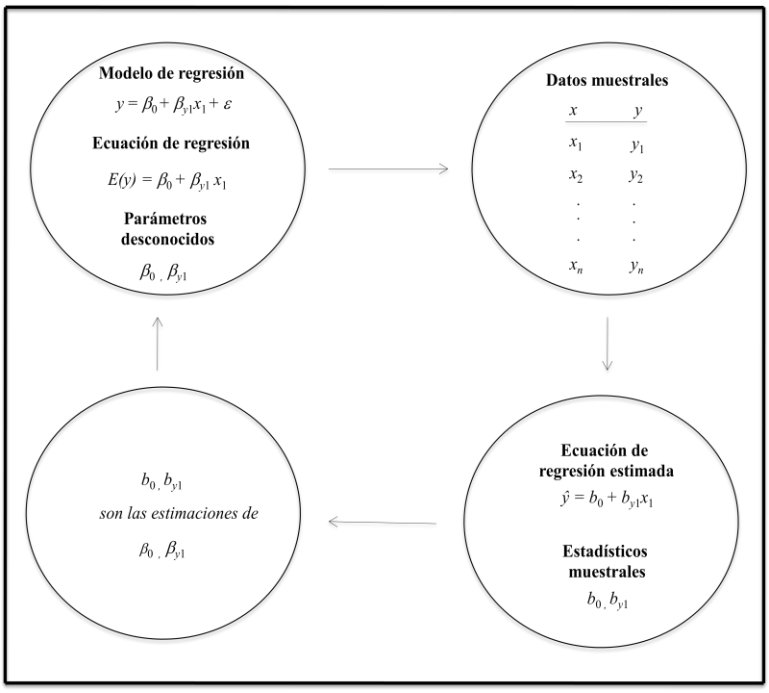
> A la gráfica de la ecuación de regresión simple estimada se le llama *Recta de regresión estimada*[144]. [En la misma] b_0 es la intersección con el eje y y b_{y1} es la pendiente (p.547).

Respecto a los demás componentes de dicha ecuación Anderson y otros (2009) agregan: "$\hat{y}$ es el estimador puntual de $E(y)$ [y x_1 es la variable independiente]" (p. 547). Así, conforme a Tacq (1998) a cada valor de x_1 le corresponde un valor de $\hat{y}$ que es situado en la recta.

Según Anderson y otros (2009) en la figura 1 se presenta de forma resumida el proceso de estimación en la regresión lineal simple.

[139] La cursiva es nuestra.
[140] Ibidem.
[141] Ibidem.
[142] Ibidem.
[143] Ibidem.
[144] Ibidem.

Figura 1. Resumen del proceso de estimación en la regresión lineal simple[145]



7.1.2. Método de los mínimos cuadrados en regresión lineal simple

Levine y otros (2006), en el marco de un estudio muestral, señalan

> El enfoque más común para encontrar b_0 y b_{y1} es el *Método de los mínimos cuadrados*[146]. Este método minimiza la suma del cuadrado de las diferencias entre los valores reales (y) y los valores predichos ($\hat{y}$) (…) [lo que se resume a través de la expresión] $\sum(y - \hat{y})^2$ (p. 413).

Conforme a Tacq (1998): "La diferencia entre y y $\hat{y}$ es el término del error (…)" (p. 101). Así, Tacq (1998) señala: "$e = y - \hat{y}$" (p. 101). En base a ese criterio se puede establecer que $\sum e^2 = \sum(y - \hat{y})^2$. Debido a que la ecuación de regresión estimada es $\hat{y} = b_0 + b_{y1}x_1$, Levine y otros (2006) indican que $\sum(y - \hat{y})^2 = \sum(y - b_0 - b_{y1}\, x_1)^2$. Levine y otros (2006) agregan

[145] Adaptado de *Estadística para administración y economía*, (p. 547), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[146] La cursiva es nuestra.

> Ya que (…) [la] fórmula [$\sum(y - b_0 - b_{y1} x_1)^2$] tiene dos incógnitas, b_0 y b_{y1}, la suma del cuadrado de la diferencia es una función de la intersección en y, [o sea, de] b_0, y de la pendiente de la muestra, b_{y1} (p. 414).

En ese sentido Levine y otros (2006) indican: "El método de mínimos cuadrados determina qué valores de b_0 y b_{y1} son los que minimizan la suma del cuadrado de las diferencias" (p. 414). Según Anderson y otros (2009) tales valores se pueden hallar por medio de las ecuaciones:

$$b_{y1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_{y1}\bar{x}_1$$

Conforme a Tacq (1998) después que se calcula los valores de b_{y1} y b_0 se puede hallar para cada valor x_1 tanto el valor de $\hat{y} = b_0 + b_{y1}x_1$ como los residuos $e = y - \hat{y}$.

7.1.3. Medidas de variación en regresión lineal simple

Levine y otros (2006), en el ámbito de una investigación muestral, establecen

> Al usar el método de mínimos cuadrados para encontrar (…) [el intercepto y el coeficiente de regresión de] un conjunto de datos existen tres medidas de variación que se necesita calcular. (…) [Por un lado,] la *Suma total de cuadrados*[147] (*SST* por sus siglas en inglés [o SCT por sus siglas en español]), es una medida de variación de los valores *y* alrededor de la media (…) [de *y*]. [Por otro lado,] (…) la variación total o la suma total de cuadrados se subdivide en *Variación explicada*[148]

[147] La cursiva es nuestra.
[148] Ibidem.

o *Suma de cuadrados de la regresión*[149] (*SSR* por sus siglas en inglés [o SCR por sus siglas en español]), la cual se debe a la relación entre x_1 y *y*, y la *Variación no explicada*[150] o *Error de la suma de cuadrados*[151] (*SSE* por sus siglas en inglés [o SCE por sus siglas en español]), la cual se debe a factores diferentes a la relación entre x_1 y *y*. (…) Las ecuaciones (…) [siguientes] definen las [referidas] medidas de variación.

(…) La suma total de cuadrados = suma de cuadrados de la regresión + el error de la suma de cuadrados [es decir:] *SST* = *SSR* + *SSE*

(…) [Igualmente] la suma total de cuadrados (*SST*) es igual a la suma del cuadrado de las diferencias entre cada valor *y* y [la media de *y*] (…) [o sea:]

$$\text{SST} = \text{suma total de cuadrados} = \sum(y_i - \bar{y})^2$$

(…) La suma de cuadrados de la regresión (*SSR*) es igual a la suma del cuadrado de las diferencias entre el valor predicho de *y* y (…) [la media de *y*] (…) [es decir:]

$$\text{SSR} = \text{variación explicada o suma de cuadrados de la regresión}$$
$$\text{SSR} = \sum(\hat{y} - \bar{y})^2$$

(…) [El] error de la suma de cuadrados (*SSE*) es igual a la suma del cuadrado de las diferencias entre el valor observado de *y* y el valor predicho de *y* (pp. 421-423):

$$\text{SSE} = \text{variación no explicada o error de la suma de cuadrados}$$
$$\text{SSE} = \sum(y_i - \hat{y})^2$$

[149] La cursiva es nuestra.
[150] Ibidem.
[151] Ibidem.

7.1.4. Coeficiente de determinación en regresión lineal simple

Levine y otros (2006) indican

> Por sí mismos *SSR*, *SSE* y *SST* aportan muy poca información. Sin embargo, (…) [la división] de la suma de cuadrados de la regresión (*SSR)* (…) [entre] la suma total de cuadrados *SST* mide la proporción de la variación en *y* que se explica por la variable independiente x_1 en el modelo de regresión. A esta razón se le llama *Coeficiente de determinación*[152] r^2 (…) (p. 424).

Según Levine y otros (2006) la fórmula utilizada para calcular el r^2 en una investigación muestral es:

$$r^2 = \frac{\text{suma de cuadrados de la regresión}}{\text{suma total de cuadrados}}$$

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

Conforme a Tacq (1998) el mencionado coeficiente adopta valores entre 0 y 1. En ese sentido Anderson y otros (2009) señalan que un r^2 igual a 1 indica que la variable independiente x_1 explica el total de la varianza de la variable dependiente *y*. Igualmente Anderson y otros (2009) dicen que un r^2 igual a 0 significa que la citada variable independiente no explica en ninguna proporción la varianza de la variable dependiente. Mientras, conforme a The University of Texas at Austin (s. f.) la interpretación de otros resultados de dicho coeficiente obedece a la siguiente regla general. De acuerdo a The University of Texas at Austin (s. f.) un r^2 mayor que 0 y

[152] La cursiva es nuestra.

menor que 0,04 indica que la variable independiente explica una proporción muy baja de la varianza de la variable dependiente. Según The University of Texas at Austin (s. f.) un r^2 mayor o igual que 0,04 y menor que 0,16 equivale a una proporción baja de la mencionada explicación. Para The University of Texas at Austin (s. f.) un r^2 mayor que 0,16 y menor que 0,36 implica una proporción moderada de la citada explicación. Conforme a The University of Texas at Austin (s. f.) un r^2 mayor o igual que 0,36 y menor que 0,64 representa una proporción alta de dicha explicación. Igualmente de acuerdo a The University of Texas at Austin (s. f.) un r^2 mayor que 0,64 y menor que 1 significa una proporción muy alta de la explicación en cuestión.

En el contexto del Procedimiento Superador la mencionada interpretación se realiza de la siguiente manera. Un r^2 igual a 1 equivale a un éxito total de la e-campaña. Un r^2 igual a 0 implica la ausencia total de éxito de la e-campaña. Un r^2 mayor que 0 y menor que 0,04 señala un éxito muy bajo de la referida e-campaña. Un r^2 mayor o igual que 0,04 y menor que 0,16 indica un éxito bajo de la mencionada e-campaña. Un r^2 mayor que 0,16 y menor que 0,36 establece un éxito moderado de la e-campaña en cuestión. Un r^2 mayor o igual que 0,36 y menor que 0,64 significa un éxito alto de la citada e-campaña. Mientras, un r^2 mayor que 0,64 y menor que 1 implica un éxito muy alto de la e-campaña.

7.1.5. Coeficiente de correlación en regresión lineal simple

"(…) El coeficiente de correlación (…) [es] una medida descriptiva de la intensidad [y del sentido] de la relación lineal entre dos variables x_1 y y" (Anderson y otros, 2009, p. 562). "Cuando se trata de dos variables solamente, [como en el caso de la regresión bivariada,] se habla de correlación simple (…)" (Spiegel, 1989, p. 241).

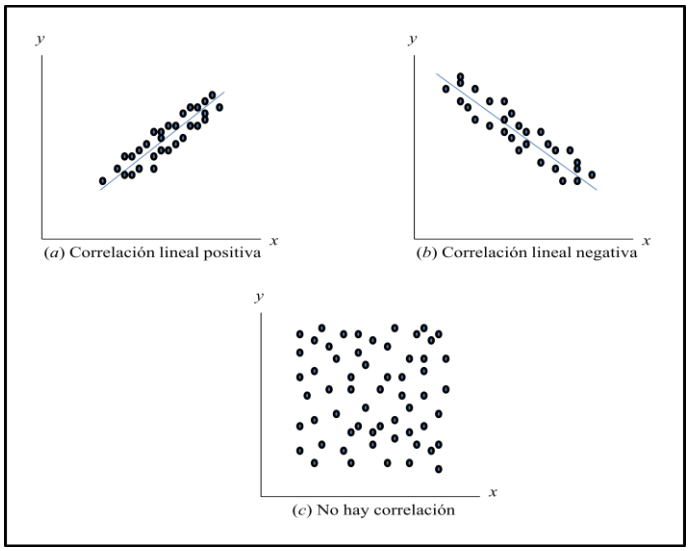
Tal como se mencionó en el apartado *Análisis de regresión lineal simple*[153], Spiegel (1989) señala: "Si x_1 y y denotan las dos variables que se consideran [en el análisis de regresión] un diagrama de dispersión muestra la localización de los puntos (x_1,y) en un sistema de coordenadas rectangulares" (p. 241). En ese sentido la figura 2 presenta tres gráficas con diferentes diagramas de dispersión en función de distintos posibles escenarios.

Conforme a Spiegel (1989)

> Si todos los puntos en (…) [el mencionado] diagrama (…) parecen encontrarse cerca de una recta, como en [los diagramas de dispersión] (*a*) y (*b*) de la (…) [figura 2], la correlación se dice lineal. (…) Si *y* tiende a incrementarse cuando se incrementa *x*, como en [el diagrama de dispersión] (*a*), la correlación se dice *Positiva*[154] o *Correlación directa*[155]. Si *y* tiende a disminuir cuando se incrementa x_1, como en [el diagrama de dispersión] (*b*), la correlación se dice *Negativa*[156] o *Correlación inversa*[157]. (…) [Pero] si no hay ninguna relación entre las variables, como en (…) [el diagrama de dispersión] (*c*), se dice que no hay correlación entre ellas, es decir, no están correlacionadas (p. 241).

[153] La cursiva es nuestra.
[154] Ibidem.
[155] Ibidem.
[156] Ibidem.
[157] Ibidem.

Figura 2. Gráficas con diferentes diagramas de dispersión según distintos posibles escenarios[158]



De acuerdo a Tacq (1998): "(…) [En un estudio muestral la fórmula del coeficiente de correlación simple] es igual a la covariación [de x_1 y y] dividida por la raíz cuadrada del producto de las variaciones [de x_1 y y] (p. 108). Según Tacq (1998) la traducción de dicha operación a fórmula es:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Anderson y otros (2009) afirman

> Los valores (…) [de este] coeficiente (…) son valores que van desde -1 hasta +1. El valor +1 indica que (…) [ambas] variables x_1 y y están perfectamente relacionadas en una relación lineal positiva. Es decir, los puntos de todos los datos se encuentran en una línea recta que tiene pendiente positiva. El valor -1 indica que x_1 y y están perfectamente relacionadas en una relación lineal negativa. (…) [Así,] todos los datos se encuentran en una línea recta que tiene pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que x_1 y y no están relacionadas linealmente (p. 562).

[158] Adaptado de *Estadística*, (p. 241), por M. R. Spiegel, 1989, España. McGraw-Hill. Derechos Reservados 1970 por Libros McGraw-Hill de México, S. A. de C. V.

Conforme a The University of Texas at Austin (s. f.) la interpretación de otros posibles resultados de dicho coeficiente se basa en la siguiente regla general. Independientemente del signo de *r*, el cual puede ser positivo o negativo, de acuerdo a The University of Texas at Austin (s. f.) un *r* mayor que 0 y menor que 0,02 equivale a una asociación muy baja de la variable independiente con la variable dependiente. De acuerdo a The University of Texas at Austin (s. f.) un *r* mayor o igual que 0,02 y menor que 0,04 representa una baja asociación entre ambas variables. Para The University of Texas at Austin (s. f.) un *r* mayor o igual que 0,04 y menor que 0,06 corresponde a una asociación moderada de las variables. Según The University of Texas at Austin (s. f.) un *r* mayor o igual que 0,06 y menor que 0,08 significa una asociación alta de dichas variables. Igualmente para The University of Texas at Austin (s. f.) un *r* mayor o igual que 0,08 y menor que 1 indica una asociación muy alta de las mencionadas variables.

En tanto, Anderson y otros (2009) señalan: "Cuando se ha realizado un análisis de regresión y se ha calculado el coeficiente de determinación r^2, el coeficiente de correlación (…) se puede calcular [a través de la ecuación siguiente:]" (p. 562.)

$$r = (\text{signo de } b_{y1})\sqrt{\text{Coeficiente de determinación}}$$

$$r = (\text{signo de } b_{y1})\sqrt{r^2}$$

"El signo del coeficiente de regresión (…) es positivo si la ecuación de regresión [estimada] tiene pendiente positiva ($b_{y1} > 0$) y es negativo si (…) [dicha] ecuación (…) tiene pendiente negativa ($b_{y1} < 0$)" (Anderson y otros, 2009, p. 563).

7.1.6. Análisis residual en regresión lineal simple

Como se indicó anteriormente en el apartado *Método de los mínimos cuadrados en regresión lineal simple*[159], Levine y otros (2006) establecen

> El residual o error (…) es la diferencia entre los valores observados (y) y los valores predichos ($\hat{y}$) (…). [Por ende] gráficamente aparece un residuo en el diagrama de dispersión como la distancia (…) entre un valor observado de y y la [recta de regresión estimada] (p. 428).

En ese sentido Anderson y otros (2009) establecen: "[El enfoque gráfico denominado *Análisis residual*[160]] (…) [se puede usar] para identificar observaciones que se pueden clasificar como observaciones atípicas o como observaciones especialmente influyentes sobre la ecuación de regresión estimada" (p. 597).

7.1.7. Detección de observaciones atípicas en regresión lineal simple

De acuerdo a Anderson y otros (2009): "(…) Una observación atípica, [es] un dato (una observación) que no sigue la tendencia del resto de los datos. (…) Son observaciones (…) sospechosas (…) que requieren un análisis cuidadoso" (p. 597).

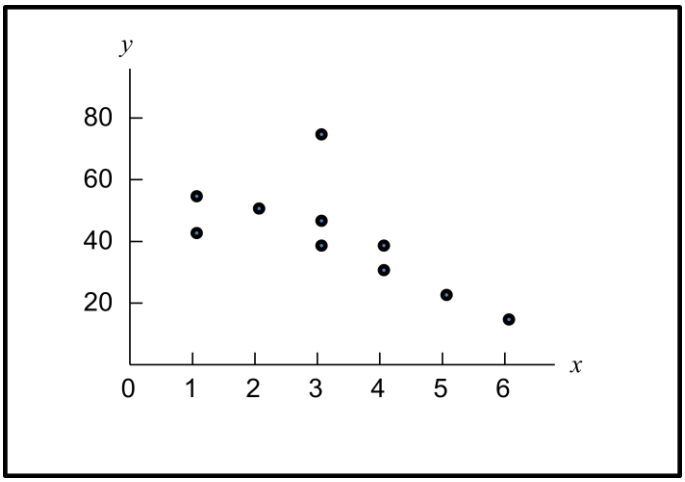
Anderson y otros (2009) agregan

> Puede tratarse de datos erróneos; si es así esos datos deben ser corregidos. Puede tratarse de una violación de las suposiciones del modelo; si es así habrá que [intentar satisfacerlas] (…). (…) [Del mismo modo] puede tratarse, simplemente, de valores inusuales que se presenten por casualidad. En ese caso (…) [dichos] valores deberán conservarse.

[159] La cursiva es nuestra.
[160] Ibidem.

Para mostrar cómo se detectan las observaciones atípicas considérense (…) el diagrama de dispersión [que se ilustra en la figura 3] (…). Con excepción de la observación (…) [$x = 3$, $y = 75$] estos datos parecen seguir un patrón que (…) [indica] una relación lineal negativa. En efecto, dado el patrón que parece seguir el resto de los datos se esperaría que [el valor de la *y* en cuestión] (…) fuera mucho más pequeño por lo que a esta observación se le considera como un dato atípico. En el caso de la regresión lineal simple las observaciones atípicas pueden detectarse mediante un simple examen del diagrama de dispersión (p. 598).

Figura 3. Diagrama de dispersión que muestra una observación atípica[161]



No obstante, a fin de identificar observaciones atípicas también se puede analizar los residuos como se señaló en el apartado *Análisis residual en regresión lineal simple* [162]. En ese sentido Anderson y otros (2009) afirman: "Para detectar observaciones atípicas (…) se puede usar los residuales estandarizados" (p. 508). De acuerdo a Anderson y otros (2009)

(…) La fórmula (…) para obtener el residual estandarizado de la observación *i* [es]

$$\frac{y - \hat{y}}{s_{y-\hat{y}}}$$

[161] De *Estadística para administración y economía*, (p. 598), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[162] La cursiva es nuestra.

donde

$$s_{y - \hat{y}} = \text{desviación estándar del residual } i$$

La fórmula general para obtener la desviación estándar del residual *i* está definida [así:]

$$s_{y - \hat{y}} = s\sqrt{1 - h}$$

donde

s = error estándar de estimación

h_i = influencia de la observación

(…) [En el apartado *Observaciones influyentes en regresión lineal simple* [163] se aborda tanto la definición como fórmula de la] influencia de una observación (…) (p. 658).

En ese contexto, "Si una observación se aleja mucho del patrón del resto de los datos (…) el valor absoluto del correspondiente residual estandarizado será grande" (Anderson y otros, 2009, p. 598). Conforme Anderson y otros (2009), si los errores están distribuidos normalmente, cualquier observación cuyo residual estandarizado sea menor a -2 o mayor a +2 será considerada como atípica. Así, de acuerdo a Anderson y otros (2009) solo 5% de los residuales estandarizados figurará fuera de dichos límites.

Anderson y otros (2009) afirman

[163] La cursiva es nuestra.

En general la presencia de una o más observaciones atípicas en un conjunto de datos tiende a incrementar *s*, el error estándar de estimación, y por lo tanto a incrementar $s_{y-\hat{y}}$, la desviación estándar del residual *i*. Dado que $s_{y-\hat{y}}$ aparece como denominador en la fórmula [*Residual estandarizado de la observación i*[164]] (…) el tamaño del residual estandarizado disminuirá a medida que *s* aumente. Esto da como resultado que aún cuando un residual estandarizado sea inusualmente grande el denominador de la fórmula [Residual estandarizado de la observación *i*], que será grande, hará que la regla del residual estandarizado falle para la identificación de una observación como observación atípica. Es posible sortear esta dificultad empleando una forma de los residuales estandarizados conocida como *Residuales estudentizados*[165] (p. 660).

7.1.7.1. Residuales estudentizados eliminados y observaciones atípicas en regresión lineal simple

Conforme a Anderson y otros (2009)

Supóngase que del conjunto de datos se elimina la observación *i* y que de las $n - 1$ observaciones restantes se obtiene una nueva ecuación de regresión estimada. Sea $s_{(i)}$ el error estándar de estimación obtenido del conjunto de datos en los que se ha eliminado la observación *i*. Si se calcula la desviación estándar del residual *i* usando $s_{(i)}$ en lugar de *s* y después se calcula el residual estandarizado de la observación *i* empleando el nuevo valor de $s_{y-\hat{y}}$, al residual estandarizado que se obtiene se le llama residual eliminado estudentizado. Si la observación *i* es una observación atípica, $s_{(i)}$ será menor a *s*. Por lo tanto el valor absoluto del residual eliminado estudentizado *i* será mayor que el valor absoluto del residual estudentizado. De esta manera, los residuos eliminados estudentizados pueden detectar observaciones atípicas que los residuales estandarizados no detectan.

(…) Para determinar si los residuales eliminados estudentizados indican la presencia de observaciones atípicas se emplea la distribución *t*. Recuérdese que *p*

[164] La cursiva es nuestra.
[165] Ibidem.

> denota el número de variables independientes y *n* el número de observaciones. Por lo tanto, si se elimina la observación *i*, el número de observaciones en el nuevo conjunto de datos es *n* – 1; en este caso la suma de cuadrados del error tiene (*n* – 1) – *p* – 1 grados de libertad (p. 660).

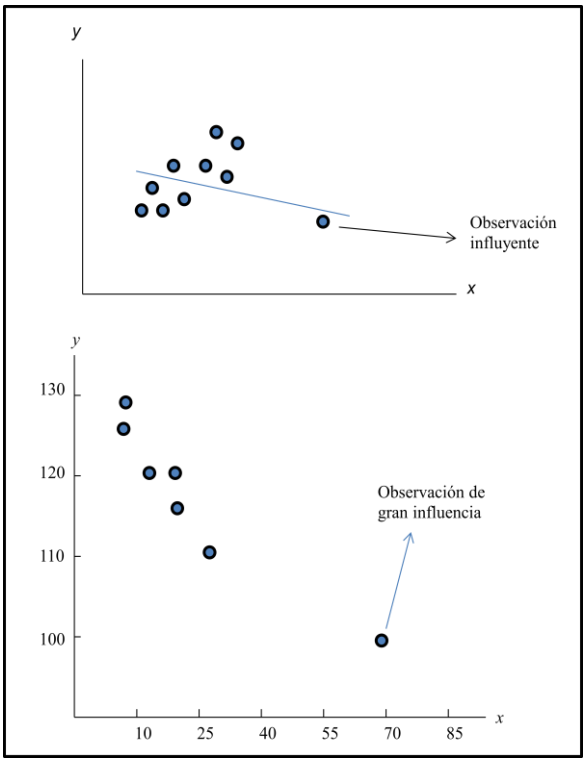
De ese modo según Anderson y otros (2009) se debe buscar en la tabla de distribución *t* el valor correspondiente a la combinación entre el nivel de significancia que se acepte y los grados de libertad que se calculen. Así, se concluye que la observación *i* es una observación atípica si su correspondiente residual eliminado estudentizado no figura dentro del intervalo ± el valor determinado por tabla.

7.1.8. Observaciones influyentes en regresión lineal simple

Según Anderson y otros (2009)

> Algunas veces una o más de las observaciones tienen una influencia fuerte sobre los resultados que se obtienen. En la figura [4] (…) se muestra un ejemplo de una observación influyente [y otro de una observación de gran influencia] en una regresión lineal simple. [En el caso de la primera] La recta de regresión estimada tiene pendiente negativa, pero si la observación influyente se elimina del conjunto de datos la pendiente de la recta de regresión estimada cambia de negativa a positiva y la intersección con el eje *y* es menor (p. 599).

Figura 4. Ejemplos de observación influyente y de gran influencia[166]



Anderson y otros (2009) agregan

> Es claro que esta sola observación tiene mucha más influencia sobre la recta estimada que [cualquier] (…) otra observación; el efecto que tiene la eliminación de cualquiera de las otras observaciones sobre la ecuación de regresión estimada es muy pequeño. Cuando solo se tiene una variable independiente las observaciones influyentes [también] pueden identificarse mediante un diagrama de dispersión.
>
> (…) Las observaciones influyentes deben examinarse cuidadosamente dado el gran efecto que tienen sobre la ecuación de regresión estimada. Lo primero que hay que hacer es verificar que no se haya cometido algún error al recolectar los datos. Si se cometió algún error se corrige y se obtiene una nueva ecuación de regresión estimada. Si la observación es correcta puede uno considerarse afortunado de tenerla. Tal dato, cuando es correcto, contribuye a una mejor comprensión del modelo adecuado y conduce a una mejor ecuación de regresión

[166] Adaptado de *Estadística para administración y economía*, (pp. 600-601), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

estimada.

(…) Las observaciones en las que la variable independiente [x_1] toma valores extremos se denominan *Datos* (…) *de gran influencia*[167] [siendo también conocidas como *Puntos u Observaciones de gran influencia*[168]]. La observación influyente de [la parte inferior de] la figura [4] (…) es un punto de gran influencia. La influencia de una observación depende de qué tan lejos está el valor de la variable independiente de su media. En el caso de una sola variable independiente la influencia (*Leverage*[169]) de la observación *i*, que se denota h_i, se calcula mediante la ecuación [siguiente:]

$$h = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

De acuerdo con esta fórmula es claro que entre más [baja] (…) se encuentre x_1 de su media mayor será la influencia (…) de la observación *i* (pp. 599-600).

Conforme a Anderson y otros (2009)

Un problema que puede presentarse al usar la influencia para identificar observaciones influyentes es que puede que se identifique una observación como una observación que tiene una gran influencia sin que necesariamente sea influyente en términos de la ecuación de regresión estimada que se obtiene (p. 661).

Sin embargo, como señala Anderson y otros (2009), antes de aceptar una conclusión final debe realizarse un nuevo análisis de regresión (ahora sin la observación que parece ser influyente) y verificar si se produce un cambio importante en los valores del intercepto y el coeficiente de regresión. Si hay un cambio de importancia se

[167] La cursiva es nuestra.
[168] Ibidem.
[169] Ibidem.

concluye que la observación es influyente, pero si no se registra un cambio sustancial se concluye que la observación no es influyente.

7.1.8.1. Uso de la medida de la distancia de Cook para identificar observaciones influyentes en regresión lineal simple

Según Anderson y otros (2009) una medición más confiable que la influencia por sí misma es la *Medida de la distancia de Cook*[170]. "[En ella] se utiliza tanto la influencia de la observación *i*, h_i, como el residual de la observación *i*, $(y - \hat{y})$, para determinar si una observación es influyente" (Anderson y otros, 2009, p. 662). De acuerdo a Anderson y otros (2009) la fórmula de la medida de la distancia de Cook es:

$$D_i = \frac{(y_i - \hat{y})^2}{(p-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

donde

D_i = medida de la distancia de Cook para la observación *i*

$(y_i - \hat{y}_i)$ = residual de la observación *i*

h_i = influencia de la observación *i*

[170] La cursiva es nuestra.

p = número de variables independientes

s = error estándar de estimación

Conforme a Anderson y otros (2009)

> Si el residual o la influencia es grande la medida de la distancia de Cook será grande e identificará una observación influyente. Como regla general se acepta que si $D_i > 1$ la observación i es influyente y debe ser analizada con más detenimiento (p. 663).

Anderson y otros (2009) establecen

> Los procedimientos para la detección de observaciones atípicas o de observaciones influyentes permiten estar alerta acerca de los efectos potenciales que algunas observaciones puedan tener en los resultados de la regresión. Cada observación atípica [o] (…) influyente justifica un examen cuidadoso. Si se encuentran errores en los datos [los mismos] se pueden corregir (…) y repetir el análisis de regresión. En general las observaciones atípicas y las observaciones influyentes no deben ser eliminadas del conjunto de datos a menos que haya una evidencia clara que indique que no provienen de elementos de la población de estudio y que no tenían que ser incluidos en el conjunto de datos original.
>
> [En tanto,] para determinar si el valor de una medida de la distancia de Cook D_i es lo suficientemente grande como para concluir que la observación i es influyente también puede compararse el valor de D_i con el percentil 50 de una distribución de F (denotado $F_{0,50}$) con $p + 1$ grados de libertad en el numerador y $n – p – 1$ grados de libertad en el denominador. Para esta prueba se necesita contar con tablas F a un nivel de significancia de (…) [0,05]. La regla práctica dada antes ($D_i > 1$) se basa en el hecho de que en muchos de los casos los valores en la tabla son cercanos a 1" (p. 663).

7.1.9. Suposiciones del modelo de regresión lineal simple

Según Anderson y otros (2009)

> Estas suposiciones son la base teórica para las pruebas *t* y *F* que se usan para determinar si la relación entre x_1 y y es significativa y para las estimaciones mediante intervalos de confianza y de predicción (…). Si las suposiciones acerca del término del error ε son dudosas puede ser que las pruebas de hipótesis acerca de la significancia de la relación de regresión y los resultados de la estimación por intervalo no sean correctos (p. 589).

Levine y otros (2006) establecen: "Las cuatro suposiciones de regresión (…) son las siguientes" (p. 428):

- Linealidad.
- Independencia de errores.
- Normalidad.
- Igual varianza (también llamada homocedasticidad).

Levine y otros (2006) afirman

> La primera suposición, *Linealidad*[171], establece que la relación entre variables es lineal. La segunda suposición, *Independencia de errores*[172], requiere que los errores ε sean independientes unos de otros. (…) La tercera suposición, *Normalidad*[173], requiere que los errores ε se distribuyan normalmente en cada valor de x_1. (…) Siempre y cuando la distribución de errores en cada nivel de x_1 no sea muy diferente a una distribución normal, las suposiciones acerca de β_0 y β_{y1} no se ven afectadas de forma importante. La cuarta suposición, *Igual varianza*[174] (…),

[171] La cursiva es nuestra.
[172] Ibidem.
[173] Ibidem.
[174] Ibidem.

requiere que la varianza [σ^2] de los errores ε sea constante para todos los valores de x_1. En otras palabras, la variabilidad de los valores de y será la misma tanto cuando x_1 es un valor bajo (…) [como] cuando (…) es un valor alto (p. 428).

Anderson y otros (2009) señalan: "Los [errores *e* o] residuales proporcionan la mejor información acerca de ε" (p.589). Así, conforme a Levine y otros (2006)

"(…) [El] (…) análisis [de los residuos, además de ayudar a identificar observaciones atípicas o influyentes como se mencionó en el apartado *Análisis residual en regresión lineal simple*[175], permite] evaluar las suposiciones y (…) determinar si el modelo de regresión seleccionado es apropiado (p. 428).

7.1.9.1. Evaluación de la linealidad en regresión lineal simple

Levine y otros (2006) indican

Para evaluar la linealidad (…) [se grafica] los residuos [*e*] en el eje vertical contra los valores (…) de la variable independiente [x_1] en el eje horizontal. Si el modelo lineal es apropiado para los datos no habría un patrón aparente en (…) [dicho] gráfico. [Un ejemplo de esto se presenta en la figura 5]. Sin embargo si el modelo lineal no es apropiado habrá una relación entre los valores x_1 y los residuos *e*. (…) [Dicho] patrón se observa en la figura (…) [6]. El panel A muestra una situación en la que, aunque hay una tendencia creciente en y conforme x_1 aumenta, la relación parece ser curvilínea porque la tendencia hacia arriba decrece conforme los valores de x_1 se incrementan. El efecto cuadrático se destaca en el panel B donde existe una clara relación entre x_1 y *e*. [Estas situaciones se hacen explícitas ya que] al graficar los residuos se remueve la tendencia lineal de x_1 con y exponiendo la falta de ajuste en el modelo lineal simple (p. 429).

Según Anderson y otros (2009) la figura 7 muestra otra variante de la gráfica de residuos contra *x* que también indica un efecto cuadrático en la relación.

[175] La cursiva es nuestra.

Figura 5. Gráfica de análisis residual evidenciando un modelo adecuado en la regresión bivariada[176]

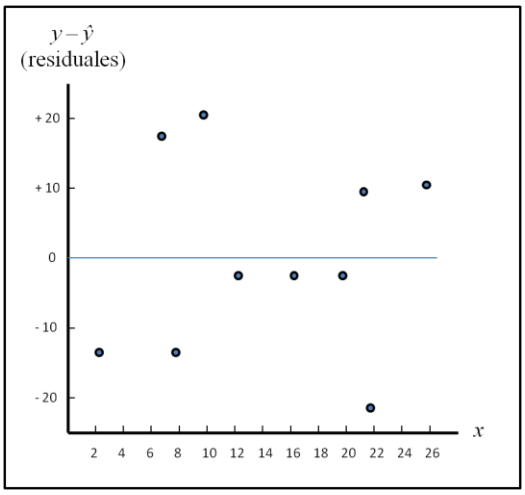


Figura 6. Gráficas de análisis residual evidenciando modelos inadecuados[177]

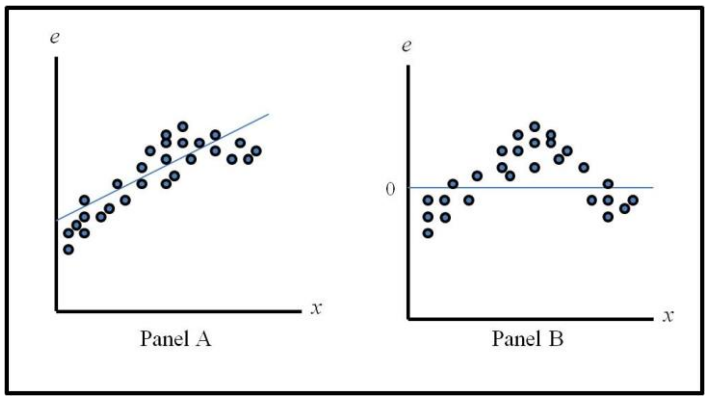
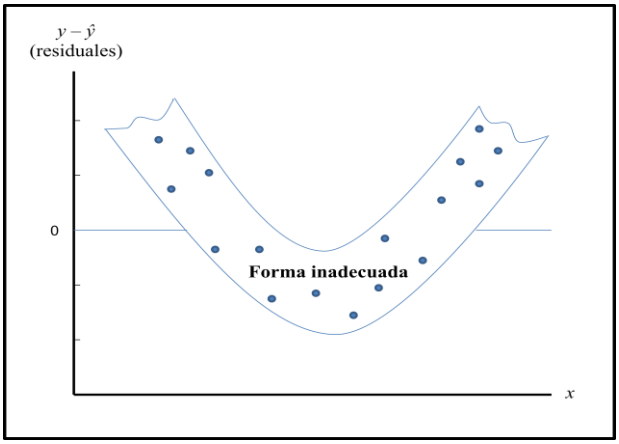


Figura 7. Forma inadecuada de la gráfica de residuales en la regresión lineal simple[178]



[176] Adaptado de *Estadística para administración y economía*, (p. 590), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[177] Adaptado de *Estadística para administración*, (p. 429), por D. M. Levine, T. C. Krehbiel y M. L. Berenson, 2006, México: Pearson Educación, Inc., publicada como PRENTICE HALL INC. Copyright 2006 por Pearson Educación de México, S.A. de C.V.

[178] Adaptado de *Estadística para administración y economía*, (p. 591), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

En tanto, según Tacq (1998) otro modo de evaluar este supuesto consiste en comparar la desviación estándar de los valores *y* con la de los errores *e*. Así, Tacq (1998) indica que si el desvío de los errores *e* es menor que el de los valores *y* se concluye que existe una relación lineal entre las variables x_1 y *y*.

De acuerdo a Anderson y otros (2009) igualmente se puede evaluar la linealidad a partir de los gráficos:

- Residuos tipificados o estandarizados contra valores de x_1.
- Residuos contra valores $\hat{y}$.
- Residuos tipificados o estandarizados contra valores $\hat{y}$.

A los fines de construir los gráficos que incluyen residuales tipificados o estandarizados Anderson y otros (2009) indican

> (…) Una variable aleatoria se estandariza sustrayéndole su media y dividiendo el resultado entre su desviación estándar. Cuando se emplea el método de mínimos cuadrados la media de los residuales es cero. Por lo tanto para obtener el residual estandarizado solo es necesario dividir cada residual entre su desviación estándar.
>
> Se puede demostrar que la desviación estándar del residual [*e*] depende del error estándar de estimación *s* y del valor correspondiente de la variable independiente x_1. [Así:]

$$s_{y-\hat{y}} = s\sqrt{1 - h}$$

donde

$s_{y-\hat{y}}$ = desviación estándar del residual [*e*]

s = error estándar de estimación

$$h = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

> [Por ende] (…) una vez calculada la desviación estándar de cada uno de los residuales, se puede calcular los residuales estandarizados dividiendo cada residual entre sus desviaciones estándar correspondientes $y - \hat{y}/s_{y-\hat{y}}$ (pp. 590-593).

En tanto, al generar cualquiera de los gráficos establecidos en este apartado se debe verificar la ausencia de patrón si existe linealidad. De lo contrario se apreciará el mencionado patrón señalando la falta de ajuste del modelo. En ese caso corresponde evaluar qué tipo de transformación mejora en mayor medida el ajuste en cuestión. Dicha temática se aborda en el apartado *Incumplimiento de los supuestos de regresión y uso de transformaciones*[179].

7.1.9.2. Evaluación de la independencia en regresión lineal simple

Anderson y otros (2009) señalan: "Si (…) [se viola la hipótesis de independencia] se pueden cometer errores serios cuando se realiza pruebas de significancia estadística basadas en el modelo de regresión (…)" (p. 732). Conforme a Levine y otros (2006) en casos en los que los datos se recolectan durante un mismo periodo de tiempo no es necesario evaluar dicho supuesto para los datos. Sin embargo Levine y otros (2006) indican que en los casos de datos recolectados en distintos periodos de tiempo a veces se verifica una relación entre las observaciones sucesivas. "(…) [Eso se debe a que] un residuo [*e*] (…) podría tender a ser semejante a los residuos en periodos adyacentes. A ese patrón (…) se le denomina autocorrelación" (Levine y otros, 2006,

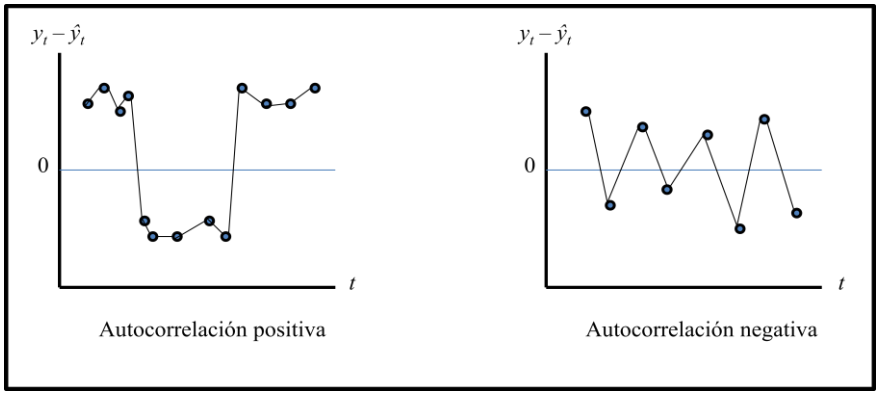
[179] La cursiva es nuestra.

p. 433). Anderson y otros (2009) también le llaman correlación serial.

En ese sentido Anderson y otros (2009) señalan

> Si el valor de *y* en el periodo *t* está relacionado con su valor en el periodo *t* – 1 existe correlación de primer orden. Si el valor de *y* en el periodo *t* está relacionado con su valor en el periodo *t* – 2 existe correlación de segundo orden y así sucesivamente.
>
> En el caso de la autocorrelación de primer orden el error en el periodo *t*, que se denota ε_t, estará relacionado con el error en el periodo *t* – 1, que se denota ε_{t-1}. En la figura [8] (...) se ilustran dos casos de autocorrelación de primer orden. [Por un lado] (…) una autocorrelación positiva [y por el otro] (…) una autocorrelación negativa. (pp. 731-732).

Figura 8. Dos conjuntos de datos con correlación de primer orden[180]



Como se observa en la figura 8, la cual se presenta arriba, Levine y otros (2006) señalan que la relación entre residuos consecutivos se revelará al graficar dichos residuos en la secuencia en que se recolectaron los datos. Por ende Levine y otros (2006) agregan que se puede evaluar la hipótesis de independencia de los errores por medio de la gráfica de residuos en contra del tiempo en que los datos se recolectaron.

[180] Adaptado de *Estadística para administración y economía*, (p. 732), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

En ese sentido Levine y otros (2006) establecen

> Si está presente un efecto positivo de autocorrelación habrá grupos de residuos con el mismo signo y podrá detectarse rápidamente un patrón aparente. Si existe una autocorrelación negativa los residuos tenderán a saltar hacia atrás y hacia delante, de positivo a negativo, luego a positivo y así sucesivamente. Este [último] tipo de patrón se observa rara vez en los análisis de regresión (p. 433).

"También se puede probar la autocorrelación (…) [a través del] estadístico de Durbin-Watson (…)" (Levine y otros, 2006, p. 430). "Este estadístico mide la correlación entre cada residuo y el residuo para el periodo de tiempo inmediatamente anterior al periodo de interés" (Levine y otros, 2006, p. 435). En ese sentido Anderson y otros (2009) agregan

> A continuación se mostrará cómo utilizar el estadístico de Durbin-Watson para detectar autocorrelación de primer orden.
>
> [Supóngase] (…) que los valores de ε no sean independientes sino que estén relacionados de la manera siguiente:
>
> $$\varepsilon = r\varepsilon_{t-1} + z_t$$
>
> donde r es un parámetro cuyo valor absoluto es menor que 1 y z_t es una variable aleatoria distribuida normal e independientemente, que tienen media cero y varianza σ^2. En la ecuación (…) [que se acaba de presentar] se ve que si $r = 0$ los términos del error no están relacionados y cada uno tiene media cero y varianza σ^2. En este caso no hay autocorrelación (…). Si $r > 0$ existe autocorrelación positiva; si $r < 0$ existe autocorrelación negativa. En cualquiera de estos casos se violan las suposiciones de la regresión acerca del término del error.
>
> En la prueba de Durbin-Watson para autocorrelación se usan los residuales para determinar si $r = 0$. Para simplificar la notación para el estadístico de Durbin-Watson el residual i se denota $e_i = y - \hat{y}$. El estadístico de prueba Durbin-Watson

se calcula como sigue:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Si valores sucesivos de los residuales se encuentran cercanos unos de otros (autocorrelación positiva) el valor del estadístico de prueba Durbin-Watson será pequeño. Si valores sucesivos de los residuales se encuentran alejados unos de otros (autocorrelación negativa) el valor del estadístico de prueba Durbin-Watson será grande.

El estadístico de prueba Durbin-Watson va de cero a cuatro, si su valor es dos (…) indica que no existe autocorrelación. Durbin y Watson elaboraron tablas para determinar cuándo su estadístico indica la existencia de autocorrelación. En la tabla [3] (…) se presentan límites inferiores y superiores (d_L y d_U) para las pruebas de hipótesis con $\alpha = 0{,}05$; *n* denota el número de observaciones [y los valores 1, 2, 3, 4 y 5 denotan la cantidad de variables independientes del modelo]. Siempre la hipótesis nula a probar es que no existe autocorrelación. [Así:]

$$H_0: r = 0$$

La hipótesis alternativa que se prueba en la autocorrelación positiva es:

$$H_a: r > 0$$

La hipótesis alternativa que se prueba en la autocorrelación negativa es:

$$H_a: r < 0$$

También se puede hacer una prueba de dos colas. En este caso la hipótesis alternativa es (pp. 732-734):

$$H_a: r \neq 0$$

Tabla 3. Valores críticos para la prueba de Durbin-Watson para autocorrelación[181]

α = 0,05										
	K = 1		*K* = 2		*K* = 3		*K* = 4		*K* = 5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

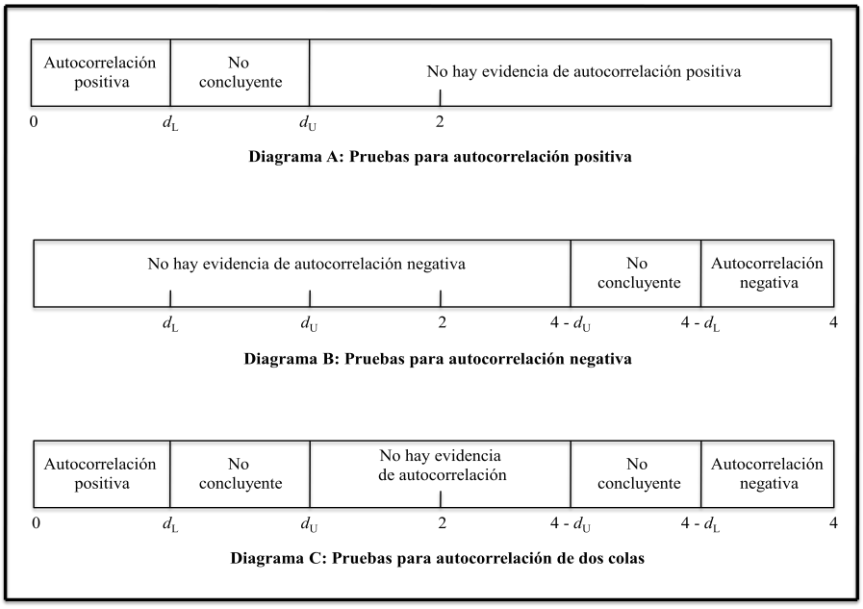
Anderson y otros (2009) indican: "(…) En las tablas de Durbin-Watson el menor valor para el tamaño de la muestra es 15. La razón es que para muestras menores la prueba suele ser no concluyente (…)" (p. 735). En cambio, cuando se analiza muestras de 15 ó más casos se puede esperar varios resultados. Así, conforme a Anderson y otros (2009)

> En la figura [9] (…) se muestra el uso de los valores d_L y d_U de la tabla (…) [3] para probar si existe autocorrelación. En el diagrama A se ilustra la prueba para autocorrelación positiva: Si $d < d_L$ se concluye que existe correlación positiva. Si $d_L \leq d \leq d_U$, (…) la prueba no es concluyente. Si $d > d_U$, se concluye que no hay evidencia de autocorrelación positiva.
>
> En el diagrama B se ilustra la prueba para autocorrelación negativa. Si $d > 4 - d_L$ se concluye que existe autocorrelación negativa. Si $4 - d_U \leq d \leq 4 - d_L$ se dice que la prueba no es concluyente. Si $d < 4 - d_U$, se concluye que no existe evidencia de autocorrelación negativa.

[181] Adaptado de *Estadística para administración y economía*, (p. 733), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

El diagrama C ilustra la prueba de dos colas. Si $d < d_L$ o $d > 4 - d_L$ se rechaza H_0 y se concluye que existe (…) [autocorrelación]. Si $d_L \leq d \leq d_U$ o si $4 - d_U \leq d \leq 4 - d_L$ se dice que la prueba no es concluyente. Si $d_U < d < 4 - d_U$ se concluye que no hay evidencia de autocorrelación" (p. 734).

Figura 9. Prueba de hipótesis para autocorrelación mediante la prueba de Durbin-Watson[182]



Cabe destacar que según Ramírez, D. C.:

La [mayoría de las] tablas son construidas para la autocorrelación positiva y en términos prácticos se prefiere utilizar el límite [inferior] de la tabla (d_L) como punto de significancia verdadero (se incluye la región no concluyente como parte de la región de rechazo) (s. f.).

En tanto, si se verifica que existe autocorrelación Anderson y otros (2009) señalan

(…) Se debe verificar si se omitieron una o varias variables independientes importantes que tengan un efecto de orden temporal sobre la variable dependiente. Si no se encuentran tales variables, incluir una variable independiente que mida el tiempo en el que se hace la observación (el valor de esta variable puede ser, por ejemplo, 1 para la primera observación, 2 para la segunda, etc.) algunas veces

[182] Adaptado de *Estadística para administración y economía*, (p. 734), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

> ayuda para eliminar o reducir la autocorrelación. Cuando no funcionan estos intentos para eliminar o reducir la autocorrelación, (…) [se puede recurrir al uso de] transformaciones (…) [lo cual se aborda en el apartado *Incumplimiento de los supuestos de regresión y uso de transformaciones*[183]] (…) (p. 735).

7.1.9.3. Evaluación de la normalidad en regresión lineal simple

Entre Levine y otros (2006) y Anderson y otros (2009) abordan los modos más utilizados para analizar el supuesto de normalidad. Así, Levine y otros (2006) establecen

> Se puede evaluar la (…) [hipótesis] de normalidad en los errores agrupando los residuos [*e*] dentro de la distribución de frecuencias y mostrando los resultados en un histograma (…). [Si la condición se satisface debe generarse una curva normal con forma de campana o una curva aproximadamente normal] (p. 430).

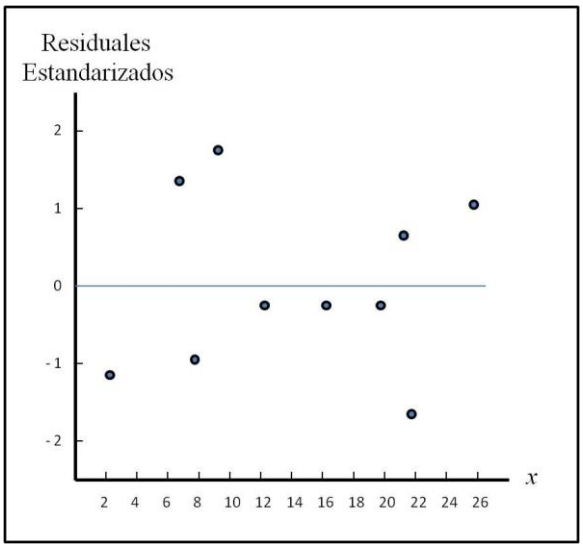
Igualmente es posible generar un histograma con curva normal a partir de los residuales tipificados o estandarizados. En ese caso se realiza el análisis del mismo modo que en el gráfico mencionado arriba.

Según Anderson y otros (2009) también se puede evaluar la hipótesis de normalidad a través de la gráfica de los residuos estandarizados contra los valores de x_1. Anderson y otros (2009) indican

> Si esta suposición se satisface debe parecer que la distribución de los residuales estandarizados proviene de una distribución de probabilidad normal estándar. Por lo tanto al observar la gráfica de los residuales estandarizados se espera encontrar que aproximadamente 95% de los residuales estandarizados (…) [estén] entre -2 y +2 [desviaciones estándar]. (…) De acuerdo a los residuales estandarizados (…) [la figura 10] no da razones para dudar de la suposición de que ε tiene una distribución normal" (p. 593).

[183] La cursiva es nuestra.

Figura 10. Gráfico residuales estandarizados contra variable independiente[184]



Asimismo, Anderson y otros (2009) afirman

> Otra manera de determinar la validez de la suposición de que el término del error tiene una distribución normal es la gráfica de probabilidad normal. Para mostrar cómo se elabora una gráfica de probabilidad normal se introduce el concepto de *Puntos normales*[185].
>
> Supóngase que de una distribución de probabilidad normal en la que la media es cero y la desviación estándar es uno se toman aleatoriamente 10 valores; supóngase que este proceso de muestreo se repite una y otra vez y que los 10 valores de cada muestra se ordenan de menor a mayor. Por ahora considérese únicamente el valor menor de cada muestra. A la variable aleatoria que representa el valor menor de estos varios muestreos se le conoce como el estadístico de primer orden.
>
> En la ciencia de la estadística se ha demostrado que en muestras de tamaño 10 tomadas de una distribución de probabilidad normal estándar el valor esperado del estadístico de primer orden es -1,55. A este valor esperado [no importa el orden del estadístico del que se trate] se le conoce como punto normal. En el caso de una muestra de tamaño $n = 10$, hay 10 estadísticos de orden y 10 puntos normales (…).

[184] Adaptado de *Estadística para administración y economía*, (p. 594), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[185] La cursiva es nuestra.

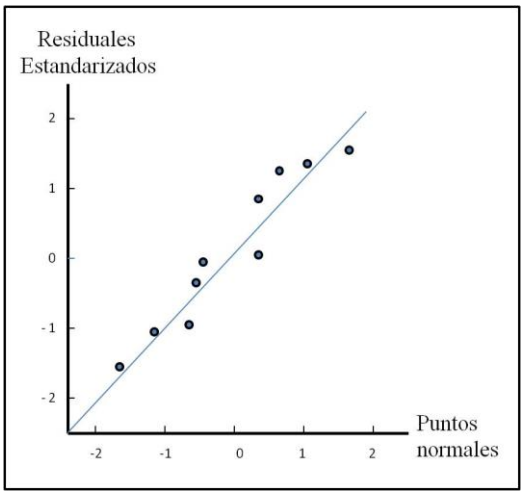
En general, un conjunto de datos que conste de *n* observaciones tendrá *n* estadísticos de orden y por lo tanto *n* puntos normales.

(…) Si se satisface la suposición de normalidad el menor residual estandarizado deberá tener un valor parecido al del menor punto normal, el siguiente residual estandarizado deberá tener un valor parecido al del siguiente punto normal y así sucesivamente. En el caso de que los residuos estandarizados se encuentren distribuidos de una manera aproximadamente normal, en una gráfica en la que los puntos normales corresponden al eje horizontal y los correspondientes residuales estandarizados al eje vertical, los puntos de la gráfica estarán situados cercanos a una línea recta a 45 grados que pase por el origen. A esta gráfica es a lo que se le conoce como *Gráfica de probabilidad normal*[186].

La figura [11] (…) es la gráfica de probabilidad normal de (…) [un ejemplo hipotético]. Para determinar si el patrón observado se desvía lo suficiente de la recta como para concluir que los residuales estandarizados no provienen de una distribución de probabilidad normal habrá que emplear el propio criterio. En la figura [11] (…) todos los puntos se encuentran cerca de esta recta. Se concluye por lo tanto que la suposición de que los términos del error tienen una distribución de probabilidad normal es razonable. En general, entre más cerca de la recta a 45 grados se encuentren los puntos más fuerte es la evidencia a favor de la suposición de normalidad. Cualquier curvatura sustancial en la gráfica de probabilidad normal es evidencia de que los residuales no provienen de una distribución de probabilidad normal (pp. 593-594).

[186] La cursiva es nuestra.

Figura 11. Gráfico de probabilidad normal[187]



En tanto, si se determina que no se cumple el supuesto de normalidad se debe recurrir a las transformaciones. Este tema es abarcado en el apartado *Incumplimiento de los supuestos de regresión y uso de transformaciones*[188].

7.1.9.4. Evaluación de la igualdad de varianza en regresión lineal simple

Levine y otros (2006) afirman: "(…) [Se] podrá evaluar la suposición de igual varianza a partir de una gráfica de los residuos [*e*] con x_1" (p. 431). Así, Levine y otros (2006) indican que se comprobará la igualdad de varianzas en caso de no encontrarse grandes diferencias en la variabilidad de los residuos para los diferentes valores de x_1.

Igualmente Levine y otros (2006) señalan

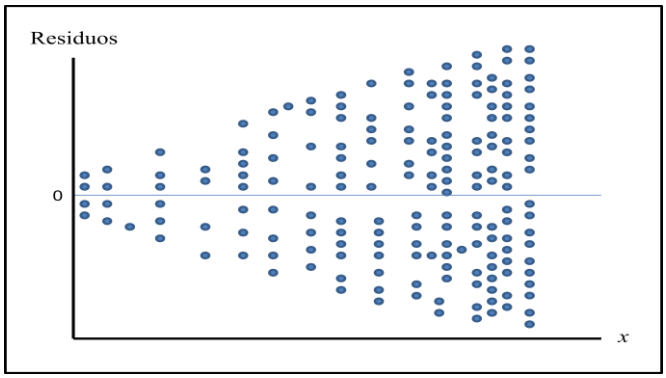
> Para examinar un caso en el que se infringe la suposición de igual varianza (…) [obsérvese] la figura [12] (…) que es una gráfica de los residuos con x_1 para un conjunto de datos hipotéticos. En esta gráfica la variabilidad de los residuos se

[187] Adaptado de *Estadística para administración y economía*, (p. 595), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[188] La cursiva es nuestra.

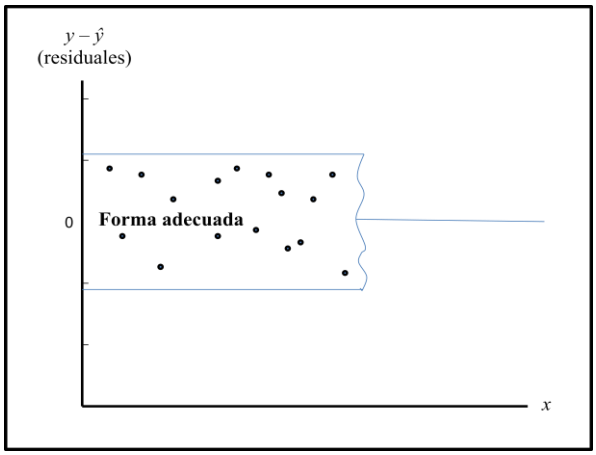
incrementa drásticamente conforme x_1 se incrementa, demostrándose así la falta de homogeneidad en las varianzas de y para cada nivel de x_1. Para estos datos la suposición de igual varianza se considera inválida (p. 431).

Figura 12. Caso hipotético de violación del supuesto igual varianza en la regresión lineal simple[189]



Conforme a Anderson y otros (2009) las figuras 13 y 14 presentan dos ejemplos de formas generales que pueden tener las gráficas de residuales cuando se evalúa la igualdad de varianza.

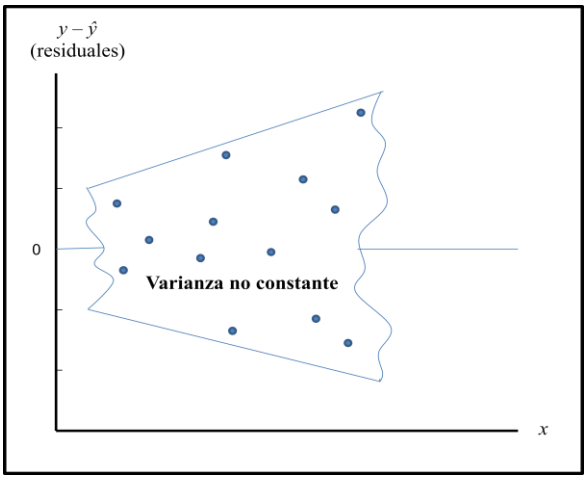
Figura 13. Forma de varianza adecuada de la gráfica de residuales[190]



[189] Adaptado de *Estadística para administración*, (p. 431), por D. M. Levine, T. C. Krehbiel y M. L. Berenson, 2006, México: Pearson Educación, Inc., publicada como Prentice Hall Inc. Copyright 2006 por Pearson Educación de México, S.A. de C.V.

[190] Adaptado de *Estadística para administración y economía*, (p. 591), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Figura 14. Forma de varianza inadecuada de la gráfica de residuales[191]



Anderson y otros (2009) señalan

> Si la suposición de que la varianza de ε es la misma para todos los valores de x_1 y si el modelo de regresión empleado representa adecuadamente la relación entre las variables, el aspecto general de la gráfica de residuales será la de una banda horizontal de puntos como en la figura (…) [13]. Pero si la varianza de ε no es la misma para todos los valores x_1 (…) el aspecto de la gráfica puede ser como (…) [las figuras 12 y 14]. En este caso se viola la suposición de que ε tiene una varianza constante (p. 589).

Para compensar esa situación se debe optar por las transformaciones. Dicha temática se aborda en el aparatado *Incumplimiento de los supuestos de regresión y uso de transformaciones*[192].

7.1.10. Prueba de significancia de la relación en regresión lineal simple

Anderson y otros (2009) establecen

[191] Adaptado de *Estadística para administración y economía*, (p. 591), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[192] La cursiva es nuestra.

[Tal como se observó en el apartado *Suposiciones del modelo de regresión lineal simple*[193]] (…) la ecuación de regresión estimada no debe ser usada hasta que se realice un análisis para determinar si el modelo empleado es adecuado. (…) [Otro] paso (…) para [determinar] (…) si el modelo (…) es [propicio] (…) es probar la significancia de la relación (p. 566).

Anderson y otros (2009) afirman

En una ecuación de regresión lineal simple (…) [el valor pronosticado] de y es una función lineal de x_1: $E(y) = \beta_0 + \beta_{y1}x_1$. Pero si el valor de β_{y1} es cero [entonces] $E(y) = \beta_0 + (0)x_1 = \beta_0$. En ese caso (…) [el valor pronosticado de y] no depende del valor de x_1 y por lo tanto se puede concluir que x_1 y y no están relacionadas linealmente. Pero si el valor de β_{y1} es distinto de cero se concluirá que las dos variables están relacionadas. Por lo tanto, para probar si existe una relación de regresión significante, se debe realizar una prueba de hipótesis para determinar si el valor de β_{y1} es distinto de cero. Hay dos pruebas que son las más usadas [la prueba t y la prueba F]. [Debido a que] en ambas se requiere una estimación de σ^2, la varianza de ε, (…) [a continuación se aborda la estimación de la varianza y luego las pruebas mencionadas] (p. 568).

7.1.10.1. Estimación de σ^2 y error de estimación en regresión lineal simple

Anderson y otros (2009) establecen

(…) La varianza de ε representa también la varianza de los valores de y respecto a la recta de regresión. Recuérdese que a las desviaciones de los valores de y de la recta de regresión estimada se les conoce como residuales. Por lo tanto SCE, la suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada. [Debido a la composición de la ecuación de regresión estimada] $\hat{y} = b_0 + b_{y1}x_1$, (…) se puede expresar (…) [que] $SCE = \Sigma(y - \hat{y})^2 = \Sigma(y - b_0 - b_{y1}x_1)^2$. (…) [En ese sentido] se ha demostrado que SCE tiene n - 2 grados de libertad, porque para (…) [calcularla] es

[193] La cursiva es nuestra.

necesario estimar dos parámetros (β_0 y β_{y1}) (p. 568).

Basados en las consideraciones recién señaladas Anderson y otros (2009) establecen que el *Error cuadrado medio*[194] (ECM) es un estimador insesgado de σ^2. Anderson y otros (2009) indican: "(…) [ECM] se calcula dividiendo SCE entre $n - 2$ [lo cual representa los grados de libertad de SCE]" (p. 568). Según Anderson y otros (2009) ya que ECM genera un estimado de σ^2 se puede apreciar que s^2 = ECM, lo que equivale a:

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n-2}$$

En tanto, Anderson y otros señalan

> Para estimar la σ se saca la raíz cuadrada de s^2. Al valor que se obtiene, s, se le conoce como (…) *Error estándar de la estimación*[195] [cuya fórmula es la siguiente:] (p. 568)

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n-2}}$$

7.1.10.2. Prueba *t* en regresión lineal simple

Anderson y otros (2009) establecen: "El objetivo de la prueba *t* es determinar si se puede concluir que $\beta_{y1} \neq 0$" (p. 569). Para ello se establecen las siguientes hipótesis:

$$H_0: \beta_{y1} = 0$$

[194] La cursiva es nuestra.
[195] Ibidem.

$$H_a: \beta_{y1} \neq 0$$

Anderson y otros (2009) afirman: "Si se rechaza H_0 se concluirá que $\beta_{y1} \neq 0$ y que entre las dos variables existe una relación estadísticamente significante" (p. 569).

En tanto, de acuerdo a Anderson y otros (2009): "(…) b_{y1} es un estimador insesgado de β_{y1}" (p. 569). Según Anderson y otros (2009) debido a que se desconoce el valor de σ se procede como se indica en la figura 15.

Figura 15. Determinación del error estándar de b_{y1}[196] para poder calcular el estadístico de prueba t[197]

Se obtiene una estimación de $\sigma_{b_{y1}}$ que se denota $s_{b_{y1}}$ o error estándar de b_{y1} mediante la fórmula

$$s_{b_{y1}} = \frac{s}{\sqrt{\sum(x_i - x)^2}}$$

Conforme a Anderson y otros (2009)

> (…) Para determinar si la relación es significativa se basa en el hecho de que el estadístico de prueba [que se ilustra en la figura 16] (…) sigue una distribución *t* con *n* - 2 grados de libertad. Si la hipótesis nula es verdadera entonces [se presenta el escenario establecido en la figura 17] (p. 570).

Otro modo de evaluar si la relación es significativa consiste en utilizar la prueba del valor-*p*. Conforme a Stats Direct (s. f.) el referido valor-*p* representa la probabilidad de rechazar la hipótesis nula (H_0) cuando la misma es cierta. Para calcularlo se

[196] Según Anderson y otros (2009) al error estándar de b_{y1} también se le conoce como desviación estándar estimada de b_{y1} y se calcula en base al error estándar de estimación (igualmente llamado desviación estándar estimada) y la raíz cuadrada del cuadrado de la suma de las diferencias.

[197] Adaptado de *Estadística para administración y economía*, (p. 569), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

recomienda utilizar el programa SPSS que lo emite automáticamente como parte del análisis, tanto en el caso de la prueba *t* como en el de la prueba *F*. Según Stats Direct (s. f.) la elección del valor-*p* por debajo del cual se rechaza H_0 es arbitraria pero generalmente se emplea los valores 0,05; 0,01 y 0,001.

De acuerdo a Anderson y otros (2009) la figura 16 muestra un resumen de las fórmulas competentes a la prueba *t* para la regresión lineal simple.

Figura 16. Fórmulas competentes a la prueba *t* para la regresión lineal simple[198]

HIPÓTESIS NULA Y ALTERNATIVA EN LA PRUEBA *t*

$H_0: \beta_{y1} = 0$
$H_a: \beta_{y1} \neq 0$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_{y1}}{s_{b_{y1}}}$$

REGLA DE RECHAZO

Método del valor-*p*: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ se toma de la distribución *t* con n – 2 grados de libertad.

Figura 17. Escenario cuando la H_0 es verdadera en la prueba *t* de la regresión lineal simple[199]

$$\beta_{y1} = 0 \quad \text{y} \quad t = \frac{b_{y1}}{s_{b_{y1}}}$$

[198] Adaptado de *Estadística para administración y economía*, (p. 570), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[199] Adaptado de *Estadística para administración y economía*, (p. 570), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

7.1.10.3. Determinación de la significancia a partir de la estimación del intervalo de confianza en regresión lineal simple

De acuerdo a Anderson y otros (2009)

La fórmula para un intervalo de confianza para β_{y1} es la siguiente:

$$b_{y1} \pm t_{\alpha/2}\, s_{b_{y1}}$$

[En dicha fórmula] el estimador puntual es b_{y1} y el margen de error (…) [está conformado por los componentes restantes como se muestra en la figura 18]. El coeficiente de confianza para este intervalo es 1 – α y $t_{\alpha/2}$ el valor *t* que proporciona un área α/2 en la cola superior de la distribución *t* con n – 2 grados de libertad (pp. 570 - 571).

Figura 18. Margen de error del intervalo de confianza en la prueba *t* de la regresión lineal simple[200]

$$t_{\alpha/2}\, s_{b_{y1}}$$

Anderson y otros (2009) señalan que si el valor hipotético de β_{y1} (que es 0) queda fuera del intervalo de confianza calculado se rechaza H_0 y se concluye que existe una relación significativa entre x_1 y *y*. Conforme a Anderson y otros (2009) de lo contrario no se rechaza H_0.

[200] Adaptado de *Estadística para administración y economía*, (p. 571), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

7.1.10.4. Prueba *F* en regresión lineal simple

Anderson y otros (2009) establecen

> Una prueba *F* basada en la distribución de probabilidad *F* puede emplearse también para probar la significancia en la regresión. Cuando solo se tiene una variable independiente la prueba *F* lleva a la misma conclusión que la prueba *t*, es decir, si la prueba *t* indica que $\beta_{y1} \neq 0$ y por la tanto existe una relación significante la prueba *F* también indicará que existe una relación significante. Pero cuando hay más de una variable independiente solo la prueba *F* puede usarse para probar que existe una relación significante general.
>
> La lógica detrás del uso de la prueba *F* para determinar si la relación de regresión es estadísticamente significativa se basa en la obtención de dos estimaciones independientes de σ^2. [En el apartado *Estimación de σ^2 y error de estimación en regresión lineal simple*[201]] se explicó cómo ECM proporciona una estimación de σ^2. Si la hipótesis nula H_0: $\beta_{y1} = 0$ es verdadera, la suma de cuadrados debida a la regresión, SCR, dividida entre sus grados de libertad proporciona otra estimación independiente de σ^2. A esta estimación se le llama el *Cuadrado medio debido a la regresión*[202] o simplemente el *Cuadrado medio de la regresión*[203], el cual se denota CMR. En general:
>
> $$\text{CMR} = \frac{\text{SCR}}{\text{Grados de libertad de la regresión}}$$
>
> En [el modelo de regresión lineal simple] (…) el número de grados de libertad de la regresión es siempre igual al número de variables independientes en el modelo (…) [o sea, 1. Por ende] CMR = SCR/1 = SCR.
>
> (…) Si la hipótesis nula es verdadera (H_0: $\beta_{y1} = 0$) CMR y ECM son dos

[201] La cursiva es nuestra.
[202] Ibidem.
[203] Ibidem.

estimaciones independientes de σ^2 y la distribución muestral de CMR/ECM sigue una distribución *F* en la que el número de grados de libertad en el numerador es igual a uno y el número de grados de libertad en el denominador es igual a *n* - 2. Por lo tanto, si $\beta_{y1} = 0$ el valor de CMR/ECM deberá ser un valor cercano a uno. Pero si la hipótesis nula es falsa, $\beta_{y1} \neq 0$, CMR sobreestimará σ^2 y el valor de CMR/ECM se inflará; de esta manera valores grandes de CMR/ECM conducirán al rechazo de H_0 y a la conclusión de que la relación entre x_1 y *y* es estadísticamente significante (pp. 571-572).

Conforme a Anderson y otros (2009) la figura 19 presenta un resumen de las fórmulas competentes a la prueba *F* de significancia para la regresión lineal simple.

Figura 19. Fórmulas competentes a la prueba *F* para la regresión lineal simple[204]

HIPÓTESIS NULA Y ALTERNATIVA EN LA PRUEBA *F*

$H_0: \beta_{y1} = 0$
$H_a: \beta_{y1} \neq 0$

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{ECM}}$$

REGLA DE RECHAZO

Método del valor-*p*: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $F \geq F_\alpha$

donde F_α es un valor de la distribución *F* con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

Según Anderson y otros (2009)

En la [figura 20] (…) se presenta la forma general de una tabla ANOVA [o Análisis de varianza] para la regresión lineal simple. (…) Regresión, error y total son (…) las tres fuentes de variación, y SCR, SCE y STC las sumas de cuadrados correspondientes que aparecen en la columna 2. En la columna 3 aparecen los grados de libertad 1 para SCR, *n* - 2 para SCE y *n* - 1 para STC. (…) [Las

[204] Adaptado de *Estadística para administración y economía*, (p. 572), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

fórmulas] de CMR y ECM aparecen en la Columna 4. [Del mismo modo] en la columna 5 aparece (…) [la fórmula] de F = CMR/ECM (…) (p. 572).

Figura 20. Imagen de la forma general de una tabla ANOVA para la regresión lineal simple[205]

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*
Regresión	SCR	1	$CMR = \frac{SCR}{1}$	$F = \frac{CMR}{CME}$
Error	SCE	$n - 2$	$CME = \frac{SCE}{n - 2}$	
Total	STC	$n - 1$		

7.1.10.5. Algunas advertencias acerca de la interpretación de las pruebas de significancia en regresión lineal simple

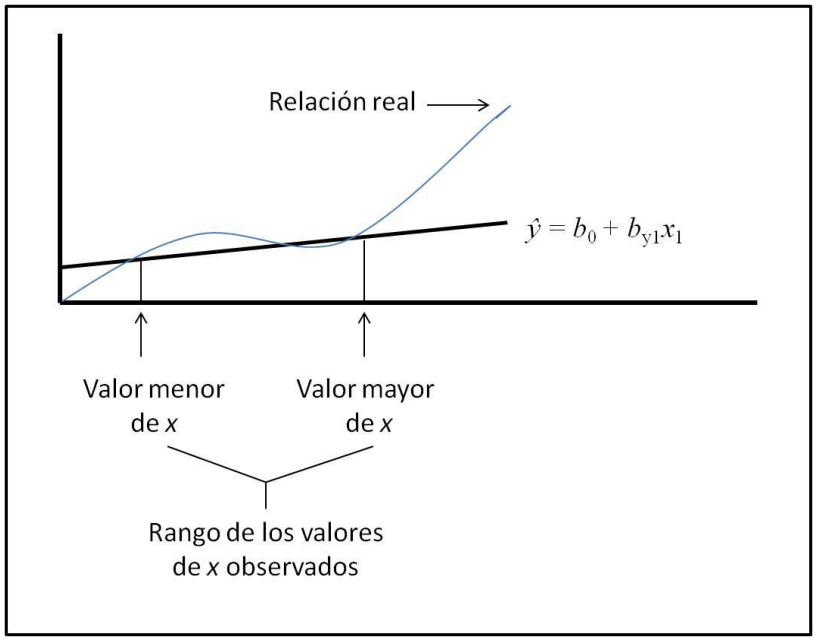
Anderson y otros (2009) indican

> Cuando se rechaza la hipótesis nula H_0: $\beta_{y1} = 0$, concluir que la relación que existe entre x_1 y y es significativa no permite que se concluya que existe una relación de causa y efecto entre x_1 y y. [Eso] (…) solo puede concluirse cuando el analista pueda dar justificaciones teóricas de que en efecto la relación es causal. (…) Además el hecho de que se pueda rechazar H_0: $\beta_{y1} = 0$ y demostrar que hay significancia estadística no permite concluir que la relación entre x_1 y y sea lineal. Lo único que se puede decir es que x_1 y y están relacionadas y que la relación lineal explica una porción significante de la variabilidad de y sobre el rango de los valores de x_1 observados en la muestra. En la figura [21] (…) se ilustra esta relación. La prueba de significancia lleva al rechazo de la hipótesis nula H_0: $\beta_{y1} = 0$ y a la hipótesis de que x_1 y y están significativamente relacionadas, pero en la figura se observa que la verdadera relación de x_1 y y no es lineal. Aunque la aproximación lineal proporcionada por $\hat{y} = b_0 + b_{y1}x_1$ es buena en el rango de los valores observados de x_1 en la muestra, [la misma] (…) se vuelve deficiente fuera

[205] Adaptado de *Estadística para administración y economía*, (p. 573), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

de ese rango. Dada una relación significante la ecuación de regresión estimada se puede usar con confianza para predicciones correspondientes a valores de x_1 dentro del rango de los valores de x_1 observados en la muestra. (…) Eso a menos que haya otras razones que indiquen que el modelo es válido más allá de este rango (pp. 573-574).

Figura 21. Ejemplo de una aproximación lineal para una relación que no es lineal[206]



7.1.11. Uso de la ecuación de regresión estimada para estimaciones y predicciones en regresión lineal simple

7.1.11.1. Estimación puntual a partir de la ecuación de regresión estimada en regresión lineal simple

Anderson y otros (2009) afirman

Con la ecuación de regresión estimada se puede obtener una estimación puntual del valor medio de y correspondiente a un determinado valor de x_1 [común a un

[206] Adaptado de *Estadística para administración y economía*, (p. 574), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

subconjunto de la población] o se puede predecir el valor de *y* que corresponde a un valor de x_1" (p. 577).

7.1.11.2. Estimación por intervalo a partir de la ecuación de regresión estimada en regresión lineal simple

"Las estimaciones puntuales no proporcionan información alguna acerca de la precisión de una estimación. Para eso es necesario obtener estimaciones por intervalo (…)" (Anderson y otros, 2009, p. 577).

De acuerdo a Anderson y otros (2009)

> La estimación puntual del valor medio de *y* es igual a la estimación puntual de un solo valor de *y*. Pero las estimaciones por intervalo que se obtienen para estos dos casos son diferentes [como se aprecia en los apartados que siguen a continuación denominados *Intervalo de confianza para el valor medio de y en regresión lineal simple*[207] e *Intervalo de predicción para un solo valor de y en regresión lineal simple*[208]] (p. 577).

7.1.11.2.1. Intervalo de confianza para el valor medio de *y* en regresión lineal simple

Anderson y otros (2009) establecen

> Con la ecuación de regresión estimada se [puede obtener] (…) una estimación puntual del valor medio de *y* que corresponde a un valor dado de x_1 [común a un subconjunto de la población. Así,] para obtener un intervalo de confianza se usa la notación siguiente (p. 578):
>
> x_p = valor dado de la variable independiente x_1 [común a un subgrupo de la población]

[207] La cursiva es nuestra.
[208] Ibidem.

y_p = valor de la variable dependiente *y* que corresponde al valor dado x_p [común a un subconjunto de la población]

$E(y_p)$ = valor medio (…) de la variable dependiente *y* que corresponde al valor dado de x_p [común a un subgrupo de la población]

$\hat{y}_p = b_0 + b_{y1}x_p$ = estimación puntual de $E(y_p)$ cuando $x_1 = x_p$

Según Anderson y otros (2009): "En general, no se puede esperar que $\hat{y}_p$ sea exactamente igual a $E(y_p)$. Para hacer una inferencia acerca de qué tan cerca está $\hat{y}_p$ de (…) $E(y_p)$ es necesario estimar la varianza de $\hat{y}_p$". (p. 578). Anderson y otros (2009) indican que la fórmula para estimarla es:

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Anderson y otros (2009) igualmente señalan: "Una estimación de la desviación estándar de $\hat{y}_p$ está dada por la raíz cuadrada de la ecuación (…) [arriba ilustrada]" (p. 578). Así:

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

De ese modo conforme a Anderson y otros (2009) el intervalo de confianza para $E(y_p)$, es:

$$\hat{y}_p \pm t_{\alpha/2}\, s_{\hat{y}_p}$$

"Obsérvese que la desviación estándar estimada de $\hat{y}_p$ (…) es menor cuando x_p = media de x_p y la cantidad x_p – media de x_p = 0. En este caso la desviación estándar estimada de $\hat{y}_p$ se convierte en" (Anderson y otros, 2009, p. 579):

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

Anderson y otros (2009) afirman

> Esto significa que cuando x_p = [media de x] (…) se obtiene la mejor estimación o la estimación más precisa del valor medio de y. Entre más alejada esté x_p de [la] media de x_p, mayor será [la diferencia] x_p – [media de x] (…). El resultado es que los intervalos de confianza para el valor medio de y son más amplios a medida que x_p se aleja de [la media de x] (p. 579).

7.1.11.2.2. Intervalo de predicción para un solo valor de *y* en regresión lineal simple

Conforme a Anderson y otros (2009) cuando, en lugar de estimar el valor medio de la variable dependiente en función de un valor dado de x_1 común a un subconjunto de la población, se quiere estimar un valor de y para un valor de x_1 determinado corresponde realizar una estimación a través de un intervalo de predicción.

Según Anderson y otros (2009)

> Para obtener un intervalo de predicción es necesario determinar primero la varianza correspondiente al uso de $\hat{y}_p$ como estimación de un valor individual de y

cuando (…) $x_1 = x_p$. Esta varianza está formada por la suma de los dos componentes siguientes:

[Primero] (…) la varianza de los valores individuales de y respecto a (…) $E(y_p)$, para la cual una estimación está dada por:

$$s^2$$

[Segundo] (…) la varianza correspondiente al uso de $\hat{y}_p$ para estimar $E(y_p)$, para la cual una estimación está dada por:

$$s^2_{\hat{y}_p}$$

[Abajo] la fórmula para [calcular la s^2] (…) de un valor individual de y_p que se denota s^2_{ind} :

$$s^2_{\text{ind}} = s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Por (…) [eso] una estimación de la desviación estándar de un solo valor de y_p es la dada por":

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

[Así], la fórmula general para un intervalo de predicción es como sigue:

$$\hat{y}_p \pm t_{\alpha/2}\, s_{\text{ind}}$$

(…) Tanto las estimaciones mediante un intervalo de confianza como las

estimaciones mediante un intervalo de predicción son más precisas cuando el valor de la variable independiente es x_p = (…) [media de x_p] (pp. 580-581).

7.2. Análisis de regresión múltiple

Como se indicó en el apartado *Análisis de regresión lineal*[209] y de acuerdo a Anderson y otros (2009): "El análisis de regresión múltiple estudia la relación de una variable dependiente con dos o más variables independientes" (p. 626).

7.2.1. Modelo de regresión y ecuación de regresión en regresión múltiple

"Los conceptos de modelo de regresión y ecuación de regresión vistos (…) [al abordar el análisis de regresión lineal simple] son aplicables en el caso de la regresión múltiple" (Anderson y otros, 2009, p. 626). Sin embargo como indica Tacq (1998) el modelo lineal ahora es más complejo pues hay dos o más variables independientes. "Para denotar el número de variables independientes se suele usar *p*" (Anderson y otros, 2009, p. 626).

Anderson y otros (2009) establecen

> A la ecuación que describe cómo está relacionada la variable dependiente *y* con las variables independientes x_1, x_2, …, x_p se le conoce como *Modelo de regresión múltiple*[210]. Se supone que el modelo de regresión múltiple toma la forma siguiente:

$$y = ß_0 + ß_{y1}x_1 + ß_{y2}x_2 + \ldots + ß_{yp}x_p + \varepsilon$$

[209] La cursiva es nuestra.
[210] Ibidem.

> En el modelo de regresión múltiple, $ß_0$, $ß_{y1}$, $ß_{y2}$, …, $ß_{yp}$, son parámetros y el término del error ε (…) es una variable aleatoria (p. 626).

En tanto, Tacq (1998) señala: "Los símbolos *y*, *x*, *i*, $ß_0$ y ε tienen el mismo significado que en el caso bivariado" (p. 115). No obstante, Tacq (1998) afirma que los coeficientes de regresión $ß_{y1}$, $ß_{y2}$, …, $ß_{yp}$ tienen un significado distinto al de $ß_{y1}$ en la regresión lineal simple pues ahora son coeficientes de regresión parcial y se denotan $ß_{y1.2\ldots i}$, $ß_{y2.1\ldots i}$, $ß_{yp.12\ldots i}$. De acuerdo a Tacq (1998) en ese sentido, contemplando por ejemplo 2 variables independientes, el modelo de regresión múltiple se expresa:

$$y = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}x_2 + \varepsilon$$

Según Tacq (1998) cada uno de los coeficientes $ß_{y1.2\ldots i}$, $ß_{y2.1\ldots i}$, $ß_{yp.12\ldots i}$ se interpreta como el cambio en *y* por cada unidad incrementada o decrecida en la correspondiente variable independiente cuando se controlan todas las demás, es decir, cuando el resto de las variables independientes permanece constante.

En ese sentido Anderson y otros (2009) indican

> Examinando con atención este modelo se ve que *y* es una función lineal de x_1, x_2, …, x_p (…) [(es decir, $ß_0 + ß_{y1.2\ldots i}x_1 + ß_{y2.1\ldots i}x_2 + \ldots + ß_{yp.12\ldots i}x_p$)] más el término del error ε. El término del error corresponde a la variabilidad en *y* que no puede atribuirse o explicarse al efecto lineal de las *p* variables independientes (p. 626).

En tanto, conforme a Anderson y otros (2009) a la ecuación que describe cómo está relacionada la variable independiente *y* con x_1, x_2, …, x_p sin la consideración del

término del error se le conoce como *Ecuación de regresión múltiple*[211]. De acuerdo a Tacq (1998) si se consideran 2 variables independientes la misma se expresa por medio de la fórmula:

$$E(y) = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}x_2$$

7.2.2. Ecuación de regresión múltiple estimada

Anderson y otros (2009) señalan

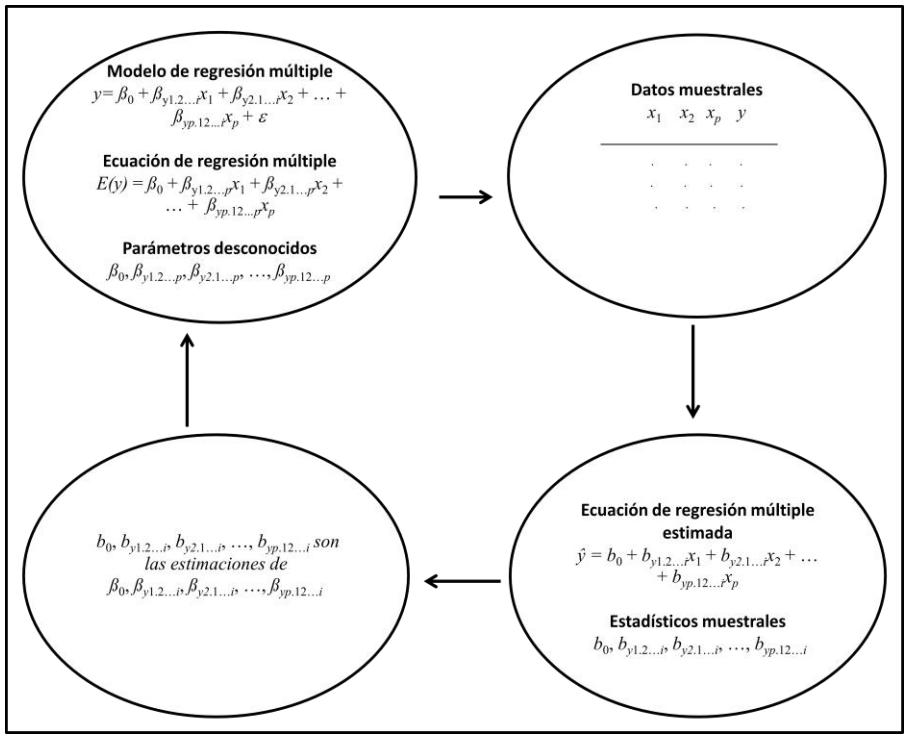
> Si se conocieran los valores de [$ß_0$, $ß_{y1.2\ldots i}$, $ß_{y2.1\ldots i}$, $ß_{yp.12\ldots i}$] (…) se podría usar la ecuación (…) [de regresión múltiple] para calcular la media de las *y* para valores dados de x_1, x_2, …, x_p [o el valor de *y* para valores dados de x_1, x_2, …, x_p]. Desafortunadamente los valores de estos parámetros no suelen conocerse, [y] es necesario estimarlos a partir de datos muestrales. (…) Con los estadísticos muestrales se obtiene la (…) *Ecuación de regresión múltiple estimada*[212]" (p. 626):

$$\hat{y} = b_0 + b_{y1.2\ldots i}x_1 + b_{y2.1\ldots i}x_2 + \ldots + b_{yp.12\ldots i}x_p$$

En dicha ecuación conforme a Anderson y otros (2009): [b_0, $b_{y1.2\ldots i}$, $b_{y2.1\ldots i}$, $b_{yp.12\ldots i}$] (…) son los estimadores de [$ß_0$, $ß_{y1.2\ldots i}$, $ß_{y2.1\ldots i}$, $ß_{yp.12\ldots i}$] (…) y $\hat{y}$ [es] el valor estimado de la variable dependiente. Este proceso de estimación (…) se muestra en la figura (…) [22] (p. 27).

[211] La cursiva es nuestra.
[212] Ibidem.

Figura 22. Proceso de estimación en la regresión múltiple[213]



7.2.3. Método de mínimos cuadrados en regresión múltiple

Anderson y otros (2009) establecen

> En [el contexto de la regresión lineal simple] (…) se usó el método de mínimos cuadrados para obtener la ecuación de regresión estimada que permitía aproximar mejor la relación lineal entre las variables dependiente e independiente. Este método también se usa para obtener la ecuación de regresión múltiple estimada. El criterio en el método de mínimos cuadrados, como [se indicó al abordar la regresión bivariada] (…), es el siguiente:
>
> $$\min \sum (y_i - \hat{y}_i)^2$$
>
> donde
>
> min = mínimo

[213] Adaptado de *Estadística para administración y economía*, (p. 627), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

y_i = valor observado en la variable dependiente en la observación *i*.

$\hat{y}_i$ = valor estimado para la variable dependiente en la observación *i* (pp. 627-628).

Según Anderson y otros (2009): "(…) [En el ámbito de la regresión lineal simple] fue posible usar (…) fórmulas para obtener b_0 y b_{y1} (…)" (p. 628). No obstante, en el caso de la regresión múltiple se puede seguir el procedimiento establecido por Tacq (1998). De acuerdo a Tacq (1998) tomando como ejemplo un caso en el que se evalúe un modelo que consta de dos variables independientes y una dependiente, se tiene la expresión:

$$\hat{y} = b_0 + b_{y1.2}x_1 + b_{y2.1}x_2 + e$$

Tacq (1998) indica que para calcular los coeficientes de regresión y el intercepto se debe tener en cuenta que los coeficientes $b_{y1.2}$ y $b_{y2.1}$ son ahora coeficientes de regresión parcial. Según Tacq (1998) eso significa que la relación entre dos variables es controlada por una tercera variable. De acuerdo a Tacq (1998) dicho control es necesario por lo siguiente. Tacq (1998) propone tomar por ejemplo b_{y1}. Para Tacq (1998) ese coeficiente no expresa la influencia de x_1 sobre *y*, pues x_2 está asociado tanto a x_1 como a *y*. Así, Tacq (1998) establece que para determinar la influencia de x_1 sobre *y* es necesario eliminar la influencia contaminante de x_2. Conforme a Tacq (1998) lo mismo ocurre en el caso de b_{y2} y la contaminación proveniente de x_1. En tanto, Tacq (1998) señala que para calcular $b_{y1.2}$ bastaría con remover la influencia de x_2 sobre x_1 y que para calcular $b_{y2.1}$ sería suficiente eliminar la influencia de x_1 sobre x_2. Así, Tacq (1998) afirma que para hallar $b_{y1.2}$ la eliminación de la influencia de x_2 sobre *y* es innecesaria al igual que para determinar $b_{y2.1}$ la eliminación de la influencia

de x_1 sobre *y* es prescindible. “Si solo la influencia de x_2 sobre x_1 es removida (y no la influencia de x_2 sobre *y*) en el cálculo de $b_{y1.2}$ el resultado parece ser el mismo” (Tacq, 1998, p. 119). Según Tacq (1998) igual sucede en el caso de $b_{y2.1}$, en el que se elimina la influencia de x_1 sobre x_2. De ese modo Tacq (1998) plantea que la secuencia de pasos para determinar los coeficientes $b_{y1.2}$ y $b_{y2.1}$ del ejemplo en cuestión es la siguiente. Conforme a Tacq (1998) en primer lugar se conduce un análisis de regresión bivariado entre x_2 y x_1, en el que x_2 constituye la variable independiente y x_1 la variable dependiente. Tacq (1998) establece que los valores estimados de x_1 (valores x_1 sombrero) se calculan y la diferencia entre x_1 y la x_1 sombrero ofrece los valores residuales. Tacq (1998) señala que la dispersión de dichos residuos indica la medida en que x_1 continúa variando tras ser eliminada la varianza común con x_2. De acuerdo a Tacq (1998) después de eso se realiza un análisis de regresión tomando como variable dependiente a *y* y como variable independiente la diferencia entre x_1 y x_1 sombrero. Tacq (1998) indica que el coeficiente de regresión resultante de ese análisis es el coeficiente de regresión parcial $b_{y1.2}$.

Según Tacq (1998) en segundo lugar se calcula el coeficiente de regresión parcial $b_{y2.1}$ de la misma manera que se calculó anteriormente el coeficiente de regresión parcial $b_{y1.2}$. Se conduce un análisis de regresión bivariado entre x_1 y x_2, en el que x_1 representa la variable independiente y x_2 la variable dependiente. Los valores estimados de x_2 (valores x_2 sombrero) se calculan y la diferencia entre x_2 y x_2 sombrero ofrece los valores residuales. La dispersión de dichos residuos indica la medida en que x_2 continúa variando tras ser eliminada la varianza común con x_1. Después de eso se realiza un análisis de regresión en el que *y* sea la variable dependiente y la diferencia entre x_2 y x_2 sombrero sea la variable independiente. El

coeficiente de regresión resultante de ese análisis es el coeficiente de regresión parcial $b_{y2.1}$. Así, de acuerdo a Tacq (1998) se está en condición de calcular el intercepto b_0 a través de la fórmula:

$$b_0 = \bar{y} - b_{y1.2}\bar{x}_1 - b_{y2.1}\bar{x}_2$$

Conforme a Tacq (1998) a pesar de haber establecido la ecuación de regresión múltiple estimada, en este caso $\hat{y} = b_0 + b_{y1.2}x_1 + b_{y2.1}x_2$, no se puede inferir automáticamente de la misma que el efecto de x_1 es mayor o menor que el de x_2. Según Tacq (1998) eso se debe a que los coeficientes de regresión parcial $b_{y1.2}$ y $b_{y2.1}$ no son apropiados para comparar los efectos. De acuerdo a Tacq (1998) dicha comparación solo se permite cuando las dispersiones de las variables X_1 y X_2 son iguales. Tacq (1998) establece que en caso contrario la comparación de los efectos se realiza a través de los *Coeficientes de regresión parcial estandarizados*[214], también llamados *Beta*[215], tema que se aborda en el apartado *Coeficientes de regresión parcial estandarizados en regresión múltiple*[216].

Cabe destacar que según Tacq (1998) la descontaminación de las variables independientes es innecesaria cuando dichas variables no están correlacionadas. "(…) En ese caso el coeficiente de regresión parcial es igual al coeficiente de regresión bivariado (…)" (Tacq, 1998, p. 121).

[214] La cursiva es nuestra.
[215] Ibidem.
[216] Ibidem.

7.2.4. Modelo lineal general

Según Anderson y otros (2009)

> Como marco general para el desarrollo de relaciones más complejas entre las variables independientes se introduce el concepto de *Modelo lineal general*[217]. (…) [Dicho modelo se compone de p variables independientes las cuales ya no se representan por x_i como en el modelo inicial sino por z_i. Por ende en el caso de un modelo que contemple 2 variables explicatorias el mismo se expresa como sigue:]
>
> $$y = ß_0 + ß_{y1.2}z_1 + ß_{y2.1}z_2 + \varepsilon$$
>
> El caso más sencillo es cuando solo se obtienen datos de una variable x_1 y se quiere estimar y por medio de una relación lineal. En ese caso $z_1 = x_1$ (…) (p. 695).

Según Anderson y otros (2009) teniendo una sola variable x_1 en el modelo el mismo se expresa en términos del modelo lineal general como sigue:

$$y = ß_0 + ß_{y1}z_1 + \varepsilon$$

De acuerdo a Anderson y otros (2009) dicha ecuación constituye el modelo de regresión lineal simple. Al respecto, Anderson y otros (2009) destacan

> (…) La palabra *Lineal*[218] en el término modelo lineal general se refiere únicamente al hecho de que (…) [los coeficientes de correlación de dicha ecuación] tienen, todos, exponente 1; [lo cual] no implica que la relación entre y y las x_i sea lineal (p. 697).

[217] La cursiva es nuestra.
[218] Ibidem.

7.2.5. Interacción

Conforme a Levine y otros (2006)

> Una interacción se presenta si el efecto de una variable independiente sobre la variable respuesta depende del valor de una segunda variable independiente. Por ejemplo, es posible que la publicidad tenga un gran efecto sobre las ventas de un producto cuando el precio es bajo. Sin embargo, si el precio (…) es muy elevado, aumentar la publicidad no cambiará drásticamente las ventas. En este caso se dice que interactúan precio y publicidad (p. 484).

Según Levine y otros (2006) para modelar el efecto de interacción se utiliza un *Término de interacción*[219], también conocido como *Término de producto cruzado*[220] o *Variable de interacción*[221], el cual está constituido por el producto entre las dos variables independientes que interactúan.

En tanto, de acuerdo a Levine y otros (2006) para demostrar que el efecto de interacción contribuye significativamente al modelo de regresión se realiza una prueba de hipótesis de pendientes iguales. En ese sentido se supone que el efecto de una variable analizada (x_1, por ejemplo) sobre la variable dependiente (y) es independiente de la otra variable en cuestión (x_2, digamos).

Levine y otros (2006) indican

> En otras palabras, (…) [se supone] que la pendiente del valor estimado de acuerdo con (…) [la variable x_1] es igual para (…) [la variable x_2]. Si estas dos pendientes son distintas entonces existe una interacción entre (…) [ambas variables] (p. 484).

[219] La cursiva es nuestra.
[220] Ibidem.
[221] Ibidem.

Según Anderson y otros (2009) dicha interacción se representa por la siguiente expresión:

$$z_3 = x_1 * x_2$$

En este caso de acuerdo a Levine y otros (2006) para determinar si la variable de interacción contribuye significativamente al modelo de regresión, se utiliza la hipótesis nula H_0: $ß_3 = 0$ contra la hipótesis alternativa H_1: $ß_3 \neq 0$. Según Levine y otros (2006) si el valor-*p* es mayor que 0,05 no se rechaza la hipótesis nula y se concluye que la interacción no hace un aporte significativo al modelo, pero si dicho valor es menor que 0,05 se rechaza la hipótesis nula y se concluye que la interacción contribuye significativamente al modelo.

Conforme a Anderson y otros (2009): "Cuando hay interacción entre dos variables (…) solo es posible obtener conclusiones claras si se considera el efecto conjunto que tienen las dos variables sobre la [variable] respuesta" (p. 701).

Según Anderson y otros (2009): "Para tomar en cuenta el efecto de interacción se usará el siguiente modelo de regresión": (p. 701).

$$y = ß_0 + ß_{y1.2(12)}x_1 + ß_{y2.1(12)}x_2 + ß_{y(12).12}x_1x_2 + \varepsilon$$

Por ende a partir del Modelo Lineal General la expresión de arriba pudiera presentarse también como sigue:

$$y = ß_0 + ß_{y1.23}z_1 + ß_{y2.13}z_2 + ß_{y3.12}z_3 + \varepsilon$$

A modo de explicar la interpretación de los coeficientes University of Delaware (s. f.) plantea un ejemplo que resulta útil.

Tómese el modelo $\hat{y} = b_0 + b_{y1.23}z_1 + b_{y2.13}z_2 + b_{y3.12}z_3 + \varepsilon$

donde

y = Incremento salarial (en dólares extra por mes).

$z_1 = x_1$ = Sexo (0 = Hombre, 1 = Mujer).

$z_2 = x_2$ = Evaluación de desempeño (0 puntos = mínimo, 100 puntos = máximo).

$z_3 = x_1x_2$ = Interacción Sexo*Evaluación de desempeño.

Asúmase que el análisis de regresión generó los siguientes resultados:

$$\hat{y} = 59{,}94 - 29{,}71\ x_1 + 4{,}84\ x_2 - 4{,}05\ x_1x_2 + \varepsilon$$

Los hombres, como se aprecia, cuando la evaluación de desempeño es 0 y no se contempla el efecto de interacción ($x_1x_2 = 0$) pueden esperar, en promedio, un aumento salarial de USD$59,94. Por cada unidad (punto) incrementada en la evaluación lograrían, en promedio, USD$4,84 extra. Esto así debido a que:

$$\hat{y} = 59{,}94 - 29{,}71(0) + 4{,}84\ x_2 - 4{,}05\ (0)x_2$$

Por tanto

$$\hat{y} = 59{,}94 + 4{,}84\, x_2$$

En cambio las mujeres pueden esperar, en promedio, un aumento salarial de USD$30,23 cuando la evaluación de desempeño es 0. Igualmente alcanzarían, en promedio, USD$0,79 extra por cada unidad (punto) incrementada en la evaluación de desempeño. Esto así ya que:

$$\hat{y} = 59{,}94 - 29{,}71(1) + 4{,}84\, x_2 - 4{,}05\,(1)x_2$$

Entonces

$$\hat{y} = (59{,}94 - 29{,}71) + (4{,}84x_2 - 4{,}05x_2)$$

Por lo que

$$\hat{y} = 30{,}23 + 0{,}79x_2$$

En tanto, cuando está presente el efecto de interacción se recomienda centrar las variables. Como se indicó en el apartado *Multicolinealidad*[222], Preacher (2003) indica que el referido centrado se obtiene restando a cada variable su media.

7.2.6. Coeficiente de determinación múltiple

Anderson y otros (2009) afirman

[222] La cursiva es nuestra.

En el (…) [apartado *Coeficiente de determinación en regresión lineal simple*[223]] se empleó el coeficiente de determinación r^2 = SCR/STC para medir la bondad de ajuste de la ecuación de regresión estimada. El mismo concepto es válido en la regresión múltiple. El término *Coeficiente de determinación múltiple*[224] indica que mide la bondad de ajuste de la ecuación de regresión múltiple estimada. El coeficiente de determinación múltiple, que se denota R^2, se calcula como sigue:

$$R^2 = \frac{\text{SCR}}{\text{STC}}$$

El coeficiente de determinación múltiple puede interpretarse como la proporción de la variabilidad en la variable [dependiente] (…) que es explicada por la ecuación de regresión estimada. Por lo tanto el producto de este coeficiente por 100, se interpreta como el porcentaje de la variabilidad en *y* que es explicada por la ecuación de regresión estimada (p. 636).

De todos modos, según Anderson y otros (2009)

Muchos analistas prefieren ajustar R^2 al número de variables independientes para evitar sobreestimar el efecto que [produce] (…) agregar una variable independiente sobre la cantidad de variabilidad explicada por la ecuación de regresión estimada. Siendo *n* el número de observaciones y *p* el número de variables independientes el *Coeficiente de determinación ajustado*[225] se calcula como sigue (p. 637):

$$R^2 = \frac{1 - (1 - R^2)n - 1}{n - p - 1}$$

Anderson y otros (2009) indican que un R^2 igual a 1 señala que las variables independientes explican el total de la varianza de la variable dependiente. Igualmente

[223] La cursiva es nuestra.
[224] Ibidem.
[225] Ibidem.

Anderson y otros (2009) establecen que un R^2 igual a 0 significa que las citadas variables independientes no explican en ninguna proporción la varianza de la variable dependiente. Mientras, según The University of Texas at Austin (s. f.) la interpretación de otros resultados de dicho coeficiente responde a la siguiente regla general. Para The University of Texas at Austin (s. f.) un R^2 mayor que 0 y menor que 0,04 refleja que las variables independientes explican una proporción muy baja de la varianza de la variable dependiente. Conforme a The University of Texas at Austin (s. f.) un R^2 mayor o igual que 0,04 y menor que 0,16 equivale a una proporción baja de la mencionada explicación. De acuerdo a The University of Texas at Austin (s. f.) un R^2 mayor o igual que 0,16 y menor que 0,36 implica una proporción moderada de la citada explicación. Según The University of Texas at Austin (s. f.) un R^2 mayor o igual que 0,36 y menor que 0,64 representa una proporción alta de dicha explicación. Igualmente de acuerdo a The University of Texas at Austin (s. f.) un R^2 mayor o igual que 0,64 y menor que 1 significa una proporción muy alta de la explicación en cuestión.

En el marco del Procedimiento Superador la referida interpretación se realiza del siguiente modo. Un R^2 igual a 1 representa un éxito total de la e-campaña. Un R^2 igual a 0 indica la ausencia total de éxito de la e-campaña. Un R^2 mayor que 0 y menor que 0,04 señala un éxito muy bajo de la referida e-campaña. Un R^2 mayor o igual que 0,04 y menor que 0,16 implica un éxito bajo de la mencionada e-campaña. Un R^2 mayor que 0,16 y menor que 0,36 significa un éxito moderado de la e-campaña en cuestión. Un R^2 mayor o igual que 0,36 y menor que 0,64 establece un éxito alto de la citada e-campaña. Mientras, un R^2 mayor que 0,64 y menor que 1 refleja un éxito muy alto de la e-campaña.

7.2.7. Coeficiente de correlación múltiple

De acuerdo a Morales Vallejo (s. f.) el coeficiente de correlación múltiple, denominado *R*, equivale al coeficiente de correlación entre la variable dependiente y el conjunto de variables independientes. Dicho coeficiente puede calcularse obteniendo la raíz cuadrada de R^2.

Como en la regresión lineal simple los valores de este coeficiente oscilan entre -1 y +1. El valor +1 indica que las variables independientes y la dependiente están perfectamente relacionadas en una relación lineal positiva. El valor -1 significa que las variables están perfectamente relacionadas en una relación lineal negativa. Los valores del coeficiente de correlación próximos a cero implican que las citadas variables no están relacionadas linealmente. Para The University of Texas at Austin (s. f.) la interpretación de otros posibles resultados de dicho coeficiente se fundamenta en la siguiente regla general. Sin importar el signo de *r*, el cual puede ser positivo o negativo, conforme a The University of Texas at Austin (s. f.) un $R > 0 <$ 0,02 equivale a una asociación muy baja de las variables independientes con la variable dependiente. Según The University of Texas at Austin (s. f.) un *R* mayor o igual que 0,02 y menor que 0,04 representa una baja asociación entre dichas variables. Para The University of Texas at Austin (s. f.) un *R* mayor o igual que 0,04 y menor que 0,06 corresponde a una asociación moderada de las variables. Según The University of Texas at Austin (s. f.) un *R* mayor que 0,06 y menor que 0,08 significa una asociación alta de las variables en cuestión. Igualmente para The University of Texas at Austin (s. f.) un *R* mayor o igual que 0,08 y menor que 1 indica una asociación muy alta de las referidas variables.

7.2.8. Detección de observaciones atípicas en regresión múltiple

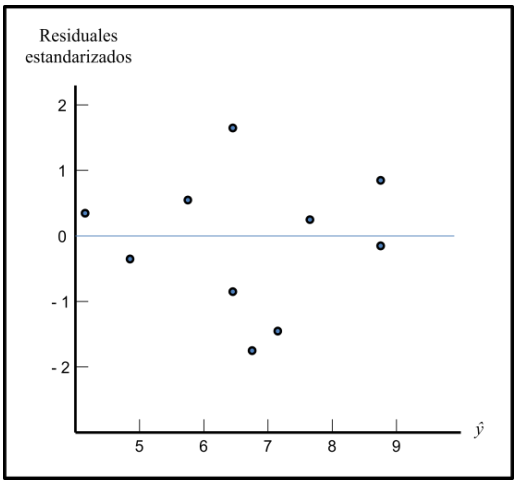
Anderson y otros (2009) señalan

> [Como se indicó en el apartado *Detección de observaciones atípicas en regresión lineal simple*[226]] una observación atípica es una observación que es inusual en relación con el resto de los datos; en otras palabras, una observación atípica no sigue el patrón del resto de los datos (p. 659).

Contrario a la regresión lineal simple, en la regresión múltiple el diagrama de dispersión no puede determinar este tipo de observaciones y es preciso recurrir al análisis de los residuos detallado en los apartados *Detección de observaciones influyentes en regresión lineal simple*[227] y *Residuales estudentizados eliminados y observaciones atípicas en regresión lineal simple*[228]. En ese sentido la figura 23 muestra un ejemplo de gráfica de residuales estandarizados. Como se aprecia la totalidad de los residuales estandarizados oscila entre -2 y +2 lo que indica la ausencia de observaciones atípicas o de gran influencia. Como se especificó en el apartado *Detección de observaciones influyentes en regresión lineal simple*[229], para un análisis más profundo corresponde realizar un análisis de residuales estudentizados.

[226] La cursiva es nuestra.
[227] Ibidem.
[228] Ibidem.
[229] Ibidem.

Figura 23. Ejemplo de gráfica de residuales estandarizados[230]



7.2.9. Observaciones influyentes en regresión múltiple

En el apartado *Detección de observaciones influyentes en regresión lineal simple*[231] se indicó que de acuerdo a Anderson y otros (2009) cuando el modelo cuenta con una sola variable la influencia de la observación *i* se obtiene a través de la fórmula:

$$h = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - x)^2}$$

No obstante, si el modelo implica dos o más variables independientes Anderson y otros (2009) establecen: "(…) Para detectar observaciones influyentes (…) [se puede emplear] la regla $h_i > 3(p + 1)/n$" (, p. 661).

[230] Adaptado de *Estadística para administración y economía*, (p. 659), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[231] La cursiva es nuestra.

7.2.9.1. Distancia de Cook e identificación de observaciones influyentes

El modo en que se emplea la distancia de Cook para detectar observaciones influyentes se abordó en el apartado *Uso de la medida de la distancia de Cook para identificar observaciones influyentes en regresión lineal simple*[232]. El mismo procedimiento que aplica para la regresión lineal simple, funciona en la regresión múltiple.

7.2.10. Supuestos del modelo de regresión múltiple

Anderson y otros (2009) afirman

> Las suposiciones acerca del término del error ε en el modelo de regresión múltiple son análogas a las suposiciones en el modelo de regresión lineal simple. [Así:]

1. El término del error ε es una variable aleatoria cuya media o valor esperado es cero, es decir, $E(\varepsilon) = 0$.

   Consecuencia:

   Para valores dados de $x_1, x_2, \ldots, x_p$, el valor esperado o valor promedio de y está dado por $E(y) = ß_0 + ß_{y1.2\ldots i}x_1 + ß_{y2.1\ldots i}x_2 + \ldots + ß_{yp.12\ldots i}x_p$.

2. La varianza de ε se denota σ^2 y es la misma para todos los valores de las variables independientes $x_1, x_2, \ldots, x_p$.
   Consecuencia:

   La varianza de y respecto a la línea de regresión es σ^2 y es la misma para todos los valores de $x_1, x_2, \ldots, x_p$.

[232] La cursiva es nuestra.

3. Los valores de ε son independientes.

   Consecuencia:

   El valor de ε para un determinado conjunto de valores de las variables independientes no está relacionado con el valor de ε de ningún otro conjunto de valores.

4. El término del error ε es una variable aleatoria distribuida normalmente y que refleja la desviación entre el valor de *y* y el valor [estimado] (…) de *y* dado por $ß_0 + ß_{y1.2\ldots i}x_1 + ß_{y2.1\ldots i}x_2 + \ldots + ß_{yp.12\ldots i}x_p$.

   Consecuencia:

   (…) La variable dependiente *y* es (…) una variable aleatoria distribuida normalmente (p.639).

Levine y otros (2006) indican: "[En el apartado *Suposiciones del modelo de regresión lineal simple*[233] se] utilizó el análisis residual para evaluar la conveniencia de utilizar el modelo de regresión lineal simple para un conjunto de datos" (p. 476). Para representar la aplicación de dicho análisis en la regresión múltiple tómese como ejemplo un modelo de regresión múltiple de dos variables independientes. Según Levine y otros (2006) en ese caso se requiere analizar las siguientes gráficas:

- Residuos contra $\hat{y}$.

- Residuos contra x_1.

[233] La cursiva es nuestra.

- Residuos contra x_2.

- Residuos contra el tiempo.

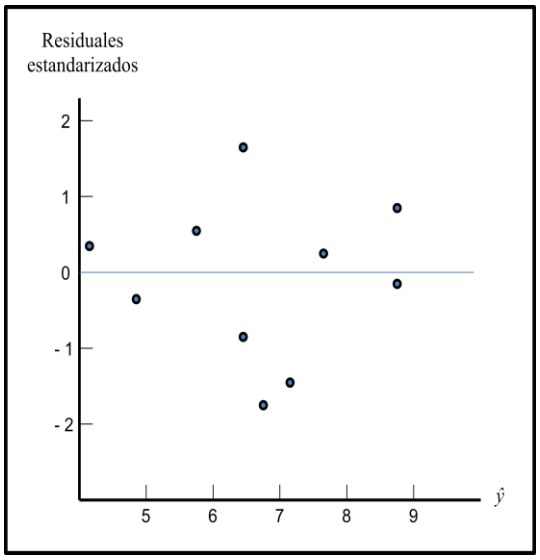
Levine y otros (2006) agregan

> La primera gráfica residual examina el patrón de los residuos contra los valores pronosticados de *y*. Si los residuos muestran un patrón para distintos valores pronosticados de *y* hay evidencia de un posible efecto cuadrático en al menos una variable independiente, [o] una posible violación de la suposición de una varianza igual (…). La segunda y tercera (…) [gráfica residual] incluyen a las variables independientes. Los patrones de una gráfica de residuos contra una variable independiente pueden indicar la existencia de un efecto cuadrático (…). [Igualmente estos gráficos pueden señalar una situación de heterocedasticidad]. La cuarta gráfica se utiliza para investigar patrones de los residuos, con objeto de validar la suposición de independencia al recopilar datos en orden cronológico. En relación con esta gráfica residual (…) es conveniente calcular el estadístico de Durbin-Watson para determinar la existencia de una autocorrelación positiva entre los residuos" (p. 476).

La figura 24 muestra el ejemplo de la gráfica de residuales estandarizados presentado anteriormente en la figura 23. Sin embargo en esta ocasión el análisis de dicho gráfico se limita a la evaluación de los supuestos de regresión. De acuerdo a Anderson y otros (2009)

> (…) [En la misma] no se observa ninguna anormalidad. [Parece cumplirse los supuestos de linealidad y homocedasticidad]. Además todos los residuales estandarizados se encuentran entre -2 y +2; por lo tanto no hay ninguna razón para cuestionar la suposición de que el término del error esté distribuido normalmente (p. 658).

Figura 24. Ejemplo de gráfica de residuales estandarizados[234]



De todos modos, Anderson y otros (2009) indican

> Para determinar si la distribución de ε parece ser normal puede emplearse también una gráfica de probabilidad normal. En (…) [el apartado *Evaluación de la normalidad en regresión lineal simple*[235]] se discutió el procedimiento y la interpretación de una gráfica de probabilidad normal. Ese mismo procedimiento es adecuado para la regresión múltiple (p. 659).

7.2.10.1. Incumplimiento de los supuestos de regresión y uso de transformaciones

Universidad Rafael Urdaneta (s. f.) afirma

> En algunas situaciones es evidente que un modelo lineal en todas las variables independientes es inadecuado. (…) [Por ejemplo, al] ajustar un modelo de primer orden y después representar gráficamente los residuos (…) [del mismo] con cada variable independiente [se puede determinar que] un modelo más apropiado contiene un término en segundo grado x_1^2 (...).

En ese sentido Anderson y otros (2009) establecen

[234] Adaptado de *Estadística para administración y economía*, (p. 659), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[235] La cursiva es nuestra.

Con la ecuación (…) [del modelo lineal general] se pueden modelar relaciones más complejas [que las de primer orden tales como la mostrada en el ejemplo que presenta Universidad Rafael Urdaneta (s. f.) arriba]. Para ilustrar esto se (…) [analiza] un problema que se le presentó a la empresa Reynolds, Inc., fabricante de balanzas industriales y de equipo para laboratorio. Los gerentes de Reynolds desean investigar la relación que existe entre la antigüedad de sus vendedores y el número de balanzas electrónicas para laboratorio que venden. (…) En la figura (…) [25] se presenta el diagrama de dispersión de estos datos. En (…) [dicho] diagrama (…) se observa que es posible que exista una relación curvilínea entre *Antigüedad de un empleado*[236] y *Número de balanzas que vende*[237]. Antes de considerar cómo obtener una relación curvilínea para este problema (…), (…) [analícese] los resultados (…) que corresponden a un modelo simple de primer orden [(ver tablas 4, 5 y 6) cuya] (…) ecuación estimada de regresión es:

Sales (Ventas) = 111 + 2,38 Months (Meses)

donde

Sales (Ventas) = Número de balanzas electrónicas para laboratorio vendidas.

Months (Meses) = Antigüedad del vendedor, en meses.

La figura (…) [26] es la gráfica de residuales estandarizados correspondiente. Aunque los resultados (…) indican que la relación sí es significativa (valor-p = 0,000) y que se explica un porcentaje grande de la variabilidad en las ventas (R^2 = 78,1%), la gráfica de residuales estandarizados sugiere que se necesita una relación curvilínea.

Para obtener una relación curvilínea en (…) [el modelo lineal general] se hace $z_1 = x_1$ y $z_2 = x_1^2$, así resulta el modelo [$y = ß_0 + ß_{y1.2}z_1 + ß_{y2.1}z_2 + \varepsilon$.]

(…) Para proporcionar la ecuación estimada de regresión correspondiente a este

[236] La cursiva es nuestra.
[237] Ibidem.

modelo de segundo orden (…), [al modelo original se le agrega los valores de la variable dependiente (*Meses de antigüedad*[238]) elevados al cuadrado] (…). En (…) [las tablas 7, 8 y 9] se presenta los resultados (…) correspondientes al modelo de segundo orden, cuya ecuación estimada de regresión es:

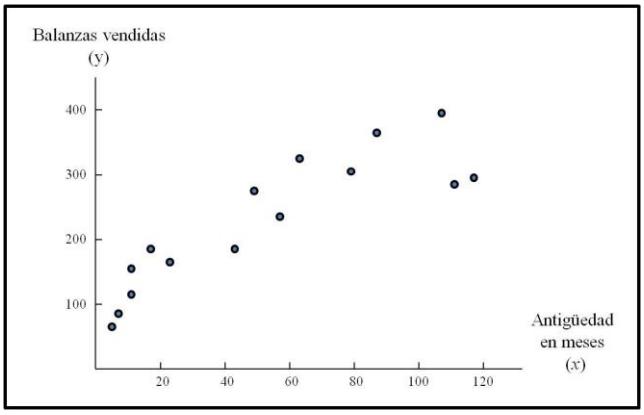
Sales (Ventas) = 45,3 + 6,34 Months (Meses) – 0,0345 Months^2 (Meses^2)

donde

Months^2 = Cuadrado del número de meses que ha trabajado el vendedor (…).

La figura (…) [27] es la gráfica de residuales estandarizados correspondiente al modelo de segundo orden. En esta (…) [figura] se observa que el patrón curvilíneo anterior ha desaparecido. Al emplear como nivel de significancia 0,05 los resultados indican que el modelo general es significativo (el valor-*p* para la prueba *F* es 0,000); observe también que el valor-*p* correspondiente al cociente *t* de MonthsSq (valor-*p* = 0,002) es menor que 0,05, por lo que se puede concluir que la adición de (…) [Months^2] al modelo es significativa. Como el valor de (…) [R^2 ajustado] es 88,6% se puede estar satisfecho con el ajuste que proporciona esta ecuación estimada de regresión (p. 695-697).

Figura 25. Diagrama de dispersión del caso Reynolds[239]



[238] La cursiva es nuestra.

[239] Adaptado de *Estadística para administración y economía*, (p. 696), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Tabla 4. Tabla Coeficientes del caso Reynolds[240]

Predictor	Coeficiente	Error estándar del coeficiente	T	P
Constante	111,23	21,63	5,14	0,000
Meses	2,3768	0,3489	6,81	0,000

Tabla 5. Tabla ANOVA del caso Reynolds[241]

Fuente	Grados libertad	Varianza	Cuadrado Medio	F	P
Regresión	1	113783	113783	46,41	0,000
Error residual	13	31874	2452		
Total	14	145657			

Tabla 6. Medidas de variación del caso Reynolds[242]

Desvío estándar	Coeficiente determinación	Coeficiente determinación ajustado
49,52	0,781	0,764

[240] Adaptado de *Estadística para administración y economía*, (p. 696), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[241] Adaptado de *Estadística para administración y economía*, (p. 696), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[242] Adaptado de *Estadística para administración y economía*, (p. 696), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Figura 26. Gráfica residuales estandarizados caso Reynolds modelo de primer orden[243]

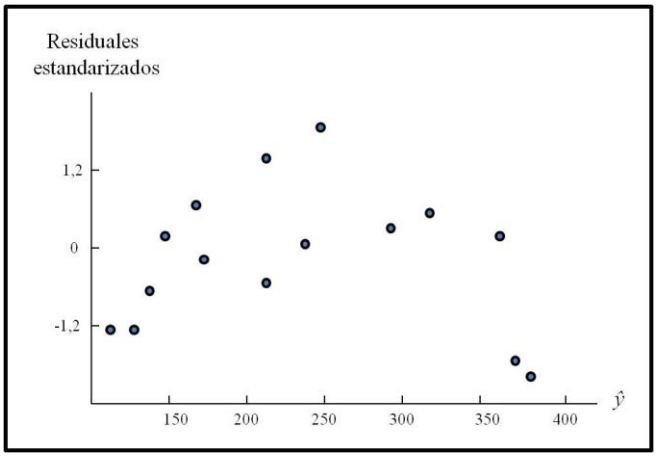


Tabla 7. Tabla Coeficientes en el modelo de segundo orden[244]

Predictor	Coeficiente	Error estándar del coeficiente	T	P
Constante	45,35	22,77	1,99	0,070
Meses	6,345	1,058	6,00	0,000
Meses al cuadrado	-0,034486	0,008948	-3,85	0,002

Tabla 8. Tabla ANOVA en el modelo de segundo orden[245]

Fuente	Grados libertad	Varianza	Cuadrado Medio	F	P
Regresión	2	131413	65707	55,36	0,000
Error residual	12	14244	1187		
Total	14	145657			

[243] Adaptado de *Estadística para administración y economía*, (p. 697), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

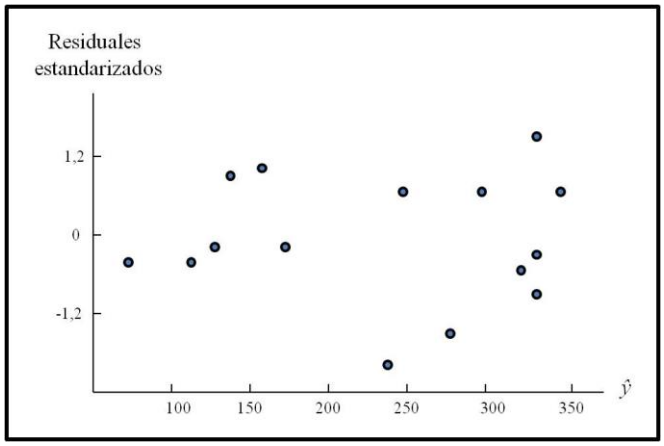
[244] Adaptado de *Estadística para administración y economía*, (p. 696), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[245] Adaptado de *Estadística para administración y economía*, (p. 698), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Tabla 9. Medidas de variación en el modelo de segundo orden[246]

Desvío estándar	Coeficiente determinación	Coeficiente determinación ajustado
34,45	0,902	0,886

Figura 27. Gráfica de residuales estandarizados correspondiente al modelo de segundo orden[247]



Universidad Rafael Urdaneta (s. f.) afirma

> (...) [Como en el caso ejemplificado arriba por Anderson y otros (2009), cuando se trabaja con modelos de segundo orden] (...) la gráfica de los residuos ($e = y - \hat{y}$) contra x_i muestra un patrón no lineal. (…) [Al respecto, aunque] existen una gama muy amplia de modelos [no lineales] (…) [los mismos] se pueden clasificar en (…) los intrínsecamente lineales, es decir, [aquellos que] se pueden transformar a la forma lineal [para luego ser analizados a partir del método de los mínimos cuadrados], y los no lineales propiamente dichos que no se pueden transformar a la forma lineal.

Así, Universidad de Alicante (s. f.) establece: "(...) Las transformaciones generalmente utilizadas [para linealizar los modelos intrínsecamente lineales son] las

[246] Adaptado de *Estadística para administración y economía*, (p. 698), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[247] Adaptado de *Estadística para administración y economía*, (p. 698), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

No lineales monotónicas[248] (...) [las cuales] no solo pueden hacer que el modelo recobre la linealidad sino que pueden solucionar problemas de normalidad y variabilidad (...)". Según Universidad Rafael Urdaneta (s. f.) los tipos más comunes de dichas transformaciones son Logarítmica en *x*, Logarítmica en *y*, Cuadrática en *x*, Raíz cuadrada de *x*, Inversa en *x*, y Recíproca en *y*. Igualmente Universidad Rafael Urdaneta (s. f.) indica que la fórmula correspondiente a cada una de las transformaciones mencionadas es[249]:

- Logarítmica en x_1: $y = b_0 + (b_{y1})(\log x_1)$
- Logarítmica en y: $\log y = b_0 + b_{y1}x_1$
- Cuadrática en x_1: $y = b_0 + b_{y1}z_1 + b_{y2}z_2$, donde $z_1 = x_1$ y $z_{2\,=}\ x_1^{\,2}$
- Raíz cuadrada de x_1: $y = b_0 + b_{y1}\sqrt{x_1}$
- Inversa en x: $y = b_0 + (b_{y1})(1/x_1)$
- Recíproca en y: $1/y = b_0 + b_{y1}x_1$

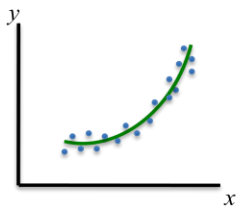
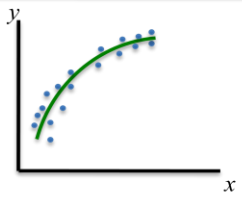
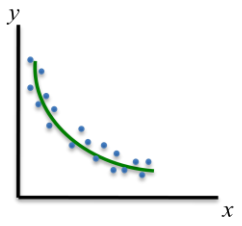
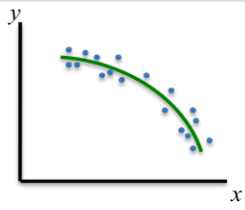
Así como existe una versión de la transformación logarítmica tanto para *y* como para *x*, VCE Further Maths (2011) señala una modalidad de la transformación cuadrática para *y*. Según VCE Further Maths (2011) la misma es denominada Cuadrática en *y* y está expresada por $y^2 = b_0 + b_{y1}x_1$. Igualmente Marín Diazaraque, J. M. (s. f.) indica un tipo de la transformación raíz cuadrada llamada Raíz cuadrada de *y* la cual se expresa como $\sqrt{y} = b_0 + b_{y1}x_1$.

[248] Según Universidad de Alicante (s. f.) las transformaciones no lineales monotónicas cambian el valor absoluto y la distancia entre los datos a la vez que conservan el orden de los mismos. "(…) [Por ende] las (…) [mismas] suponen expresar los datos en una escala diferente manteniendo su originalidad a la vez que se favorece el cumplimiento de determinados supuestos." (Universidad de Alicante, s. f.).

[249] La fórmula de cada transformación se presenta en función de un modelo simple (con una única variable independiente) para simplificar la expresión representada. Para plantear estas fórmulas en base a un modelo múltiple (con dos o más variables independientes) se agrega a las mismas todas las variables que compongan al modelo quedando afectadas por la transformación que se realice las distintas variables que corresponda.

De acuerdo a Universidad de Alicante (s. f.): "(...) [se pueden emplear las transformaciones que sean] oportunas sobre la [/s] variable/s que provoca/n la desviación del supuesto [de linealidad, normalidad o variabilidad] (…)". De esa manera, en caso de falta de linealidad según VCE Further Maths (2011) McRae (2005) plantea un conjunto de transformaciones que se pueden aplicar de acuerdo a la forma que adopte el diagrama de dispersión entre *y* y *x*. La figura 28 resume dicha información.

Figura 28. Transformaciones conforme al diagrama de dispersión[250]

Curvas	Transformaciones
	x^2, $\log y$, o $1/y$
	y^2, $\log x$, o $1/x$
	$\log x$, $\log y$, $1/x$ o $1/y$
	x^2, o y^2

De acuerdo a VCE Further Maths (2011) al elevar una variable al cuadrado los

[250] Adaptado de *Maths tutorial: Question on data transformation (statistics)*, recuperado de http://www.youtube.com/watch?v=EJ6EhfenqNs, por VCE Further Maths, 2011.

valores más pequeños de la misma sufren ligeras modificaciones mientras los valores más grandes se expanden en mayor medida. Igualmente VCE Further Maths (2011) indica que tanto cuando se calcula el logaritmo base 10 de una variable o cuando se divide $1/y$ o $1/x$ se comprimen en mayor forma los valores mayores en tanto los valores menores sufren leves alteraciones. Asimismo, conforme a Marín Diazaraque (s. f.) cuando se busca comprimir los valores de una variable también se puede aplicar a la misma la transformación de raíz cuadrada. Igualmente con el interés de solucionar la falta de linealidad Marín Diazaraque (s. f.) propone basarse en la asimetría evidenciada a través de un histograma. Marín Diazaraque (s. f.) indica: "[Ante] (...) distribuciones de frecuencias con asimetría negativa (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar la transformación $y = x^2$. (...) Para distribuciones asimétricas positivas se usan las transformaciones $\sqrt{x}$, $\ln(x)$ y $1/x$ (...)". Asimismo, Anderson y otros (2009) señalan que ante la presencia de autocorrelación se puede hacer transformaciones de las variables independientes. Igualmente, según The University of Texas at Austin (s. f.) si se identifica una violación de la normalidad se puede optar por una transformación ya sea logarítmica, raíz cuadrada, o inversa. Del mismo modo, conforme a Anderson y otros (2009) para compensar una violación de variabilidad se puede transformar la variable dependiente y de manera logarítmica o recíproca. De igual modo según Universidad de Alicante (s. f.), aplicar transformaciones en base a raíces también puede solucionar problemas de variabilidad. Para mostrar un ejemplo del empleo de las transformaciones mencionadas Anderson y otros (2009) presentan un caso con dificultades de variabilidad. De acuerdo a Anderson y otros (2009)

> (…) [Este ejemplo consiste en un caso que se ha denominado Millas-Peso, pues analiza la relación entre] millas por galón y pesos de 12 automóviles. El diagrama de dispersión de la figura (…) [29] indica que entre estas dos variables existe una

relación lineal negativa. Por tanto, se usa un modelo simple de primer orden para relacionar estas dos variables. En (…) [las tablas 10, 11, 12 y 13] se (…) [presenta] los resultados (…) [respecto al análisis de regresión correspondiente]. La ecuación estimada de regresión es:

$$MPG = 56{,}1 - 0{,}0116 \text{ Weight (Peso)}$$

donde

MPG = Rendimiento en millas por galón.
Weight (Peso) = Peso del automóvil dado en libras.

Figura 29. Diagrama de dispersión del caso Millas-Peso[251]

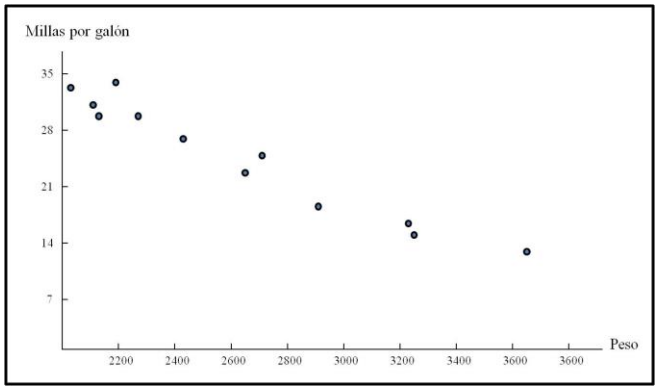


Tabla 10. Tabla Coeficientes en el modelo del caso Millas-Peso[252]

Predictor	Coeficiente	Error estándar del coeficiente	T	P
Constante	56,096	2,582	21,72	0,000
Peso	-0,01	0,00	-12,03	0,000

[251] Adaptado de *Estadística para administración y economía*, (p. 703), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[252] Adaptado de *Estadística para administración y economía*, (p. 703), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Tabla 11. Tabla ANOVA en el modelo del caso Millas-Peso[253]

Fuente	Grados libertad	Varianza	Cuadrado Medio	F	P
Regresión	1	403,98	403,98	144,76	0,000
Error residual	10	27,91	2,79		
Total	11	431,88			

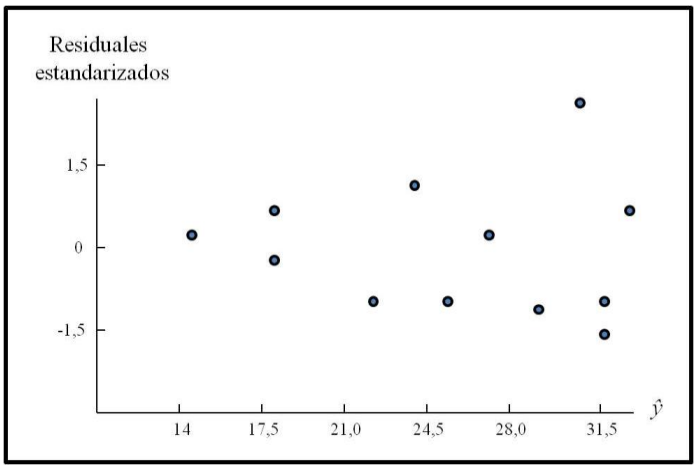
Tabla 12. Medidas de variación en el modelo del caso Millas-Peso[254]

Desvío estándar	Coeficiente determinación	Coeficiente determinación ajustado
1,671	0,935	0,929

Tabla 13. Observaciones inusuales en el modelo del caso Millas-Peso[255]

Observaciones inusuales	Residual	Residual estandarizado
3	3,487	2,26

Figura 30. Gráfica de residuales estandarizados del modelo de primer orden del caso Millas-Peso[256]



El modelo es significativo (el valor-*p* en la prueba *F* es 0,000) y el ajuste es muy

[253] Adaptado de *Estadística para administración y economía*, (p. 703), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[254] Adaptado de *Estadística para administración y economía*, (p. 703), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[255] Adaptado de *Estadística para administración y economía*, (p. 703), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[256] Adaptado de *Estadística para administración y economía*, (p. 704), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

bueno ([R^2] (…) = 93,5%). Sin embargo, en la (…) [tabla 13] se ve que la observación 3 ha sido identificada como una observación cuyo residual estandarizado es grande.

La figura (…) [30] es la gráfica de los residuales estandarizados correspondientes al modelo de primer orden. Su forma no parece ser la de la banda horizontal que se esperaría observar si las suposiciones acerca del término del error fueran válidas. La variabilidad de los residuales parece aumentar a medida que aumenta el valor de la $\hat{y}$. En otras palabras, se observa la forma de cuña (…) [indicando] una varianza que no es constante. Si las suposiciones para el modelo de esta prueba no parecen satisfacerse entonces no se justifica sacar conclusiones acerca de la significancia estadística de la ecuación estimada de regresión que se obtiene.

El problema de una varianza no constante suele corregirse al transformar la variable dependiente a otra escala. Por ejemplo, si se trabaja con el logaritmo de la variable dependiente en lugar de la variable dependiente original los valores de la variable dependiente se comprimirán (estarán más cercanos unos a otros) y con esto disminuirán los efectos de la varianza no constante. La mayor parte de los paquetes de software para estadística proporcionan la posibilidad de aplicar *Transformaciones logarítmicas*[257] mediante logaritmos base 10 (logaritmos comunes) o logaritmos base e = 2,71828… (logaritmos naturales). [A los fines de transformar la variable dependiente (MPG) del ejemplo anterior] (…) se empleará (…) la ecuación estimada de regresión que relaciona el peso con el logaritmo natural de las Millas por galón [donde la variable dependiente se rotula como] (…) LogeMPG. En (…) [las tablas 14, 15 y 16] se presenta (…) [los resultados del análisis correspondiente].

[257] La cursiva es nuestra.

Tabla 14. Tabla Coeficientes en el modelo de segundo orden del caso Millas-Peso[258]

Predictor	Coeficiente	Error estándar del coeficiente	T	P
Constante	4,52423	0,09932	45,55	0,000
Peso	-0,0005010	0,00003722	-13,46	0,000

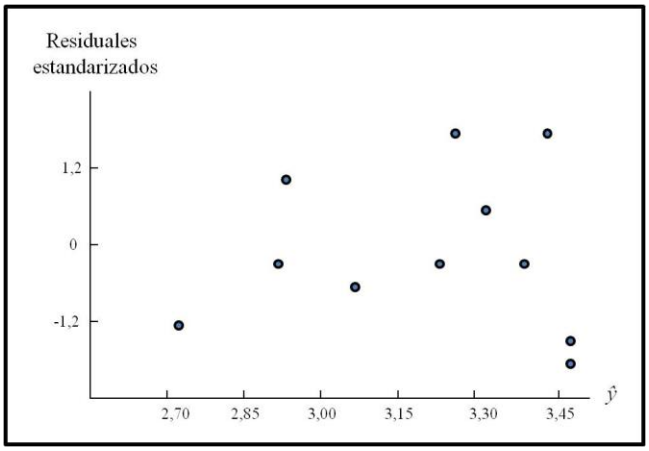
Tabla 15. Tabla ANOVA en el modelo de segundo orden del caso Millas-Peso[259]

Fuente	Grados libertad	Varianza	Cuadrado Medio	F	P
Regresión	1	0,74822	0,74822	181,22	0,000
Error residual	10	0,04129	0,00413		
Total	11	0,78950			

Tabla 16. Medidas de variación en el modelo de segundo orden del caso Millas-Peso[260]

Desvío estándar	Coeficiente determinación	Coeficiente determinación ajustado
0,06425	0,948	0,942

Figura 31. Forma de cuña ausente en el modelo de segundo orden del caso Millas-Peso[261]



[258] Adaptado de *Estadística para administración y economía*, (p. 704), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[259] Adaptado de *Estadística para administración y economía*, (p. 704), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[260] Adaptado de *Estadística para administración y economía*, (p. 704), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[261] Adaptado de *Estadística para administración y economía*, (p. 705), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2008, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

> Al observar la gráfica de residuales de la figura (…) [31] se ve que la forma de cuña ha desaparecido. Además, ninguna de las observaciones ha sido identificada como una observación cuyo residual estandarizado sea grande. El modelo en el que se emplea como variable dependiente el logaritmo de las Millas por galón es estadísticamente significativo y proporciona un ajuste excelente a los datos observados. Por tanto se recomendará usar la ecuación estimada de regresión:
>
> LogeMPG = 4,52 – 0,000501 Weight (Peso)
>
> Para estimar el rendimiento en Millas por galón de un automóvil que pese 2.500 libras se obtiene primero una estimación del logaritmo del rendimiento de millas por galón.
>
> LogeMPG = 4,52 – 0,000501 (2.500)
>
> LogeMPG = 3,2675
>
> La estimación de las millas por galón se obtiene al hallar el número cuyo logaritmo natural es 3,2675. Al emplear una calculadora con función exponencial o elevar *e* a la potencia 3,2675 se obtiene 26,2 millas por galón.
>
> (...) [Igualmente para afrontar este caso se puede] usar como variable dependiente 1/*y* en lugar de *y*. (…) [Cabe destacar que] no hay manera de determinar qué transformación funcionará mejor, si una transformación logarítmica o una transformación recíproca si no es probándolas (pp. 701-705).

Cabe destacar que en caso de un análisis de regresión lineal simple entre una variable dependiente *y* y otra independiente *x* se dispone de una serie distinta de transformaciones a considerar. De acuerdo a Supo, J. (2011) se puede determinar cuál es más apropiada evaluando los gráficos correspondientes a diversos modelos. Según Supo (2011) una selección de dichos modelos está disponible en SPSS. En la tabla 17 Rodríguez, M. (2012) plantea la lista de los mismos y sus consecuentes

transformaciones. De acuerdo a Supo (2011) para evaluar en SPSS la pertinencia de uno o más de esos modelos se elige en el menú *Analizar*[262] la opción *Regresión*[263] en cuya lista desplegable se selecciona la alternativa *Estimación curvilínea…*[264]. Conforme a Supo (2011) se especifica las variables dependiente e independiente en sus respectivos cuadros de texto. Después Supo (2011) indica que en la sección *Modelos*[265] del cuadro de diálogo *Estimación curvilínea*[266] se marca/n el/los modelo/s Lineal, Logarítmico en *x*, Inverso en *x*, Cuadrático en *x*, Cúbico en *x*, Potencia, Compuesto, G, Crecimiento, y Exponencial según se desee. Luego se cliquea el botón *Aceptar*[267] respetando la elección predeterminada que hace SPSS de las opciones *Incluir la constante en la ecuación*[268] y *Representar los modelos*[269].

[262] La cursiva es nuestra.
[263] Ibidem.
[264] Ibidem.
[265] Ibidem.
[266] Ibidem.
[267] Ibidem.
[268] Ibidem.
[269] Ibidem.

Tabla 17. Modelos y transformaciones según Rodríguez (2012)[270]

Modelo	**Ecuación estimada**	**Ecuación estimada transformada**
Lineal	$\hat{y} = b_0 + b_{y1}x_1$	Se aplica tal cual
Logarítmico en x_1	$\hat{Y} = b_0 + (by1)(\ln x_1)$	Se aplica tal cual
Inverso en x_1	$\hat{Y} = b_0 + (b_{y1})(1/x_1)$	Se aplica tal cual
Cuadrático en x_1	$\hat{Y} = b_0 + b_{y1}z_1 + b_{y2}z_2$ donde $z_1 = x_1$ y $z_2 = x_1^2$	Se aplica tal cual
Cúbico en x_1	$\hat{y} = b_0 + b_{y1}z_1 + b_{y2}z_2 + b_{y3}z_3$ donde $z_1 = x_1$; $z_2 = x_1^2$; $z_3 = x_1^3$	Se aplica tal cual
Potencia	$\hat{y} = (b_0)(x_1^b)$	$\text{Ln}\hat{y} = \ln b_0 + (b_{y1})(\ln x_1)$
Compuesto	$\hat{y} = (b_0)(b_{y1}^x)$	$\text{Ln}\hat{y} = \ln b_0 + (\ln b_{y1})(x_1)$
G	$\hat{y} = \exp(b_0 + (by_1)(1/x_1))$	$\ln\hat{y} = b_0 + (by_1)(1/x_1)$
Crecimiento	$\hat{y} = \exp(b_0 + b_{y1}x_1)$	$\ln\hat{y} = b_0 + b_{y1}x_1$
Exponencial	$\hat{y} = (b_0)(\exp(b_{y1}x_1))$	$\ln\hat{y} = \ln b_0 + (b_{y1})(x1)$

Nótese que las transformaciones propuestas por Rodríguez (2012) no incluyen las transformaciones Recíproca en y ($1/y$), Logarítmica en y ($\ln y$), Raíz cuadrada en y ($\sqrt{y}$), Cuadrática en y (y^2), ni Raíz cuadrada en x ($\sqrt{x}$). De esa manera, no se puede determinar si alguna de las mismas resulta mejor opción. En ese sentido, Universidad Rafael Urdaneta (s. f.) plantea un modo más integrador de emplear las transformaciones mencionadas en este apartado el cual consiste en aplicar todas sin excepción. Así, resulta útil transformar las variables del modelo de primer orden que se desea linealizar en función de dichas transformaciones. De acuerdo a Universidad Rafael Urdaneta (s. f.) después de obtener los modelos transformados se elige el que mejor ajuste presente, es decir, el de mayor coeficiente de determinación (R^2).

[270] Adaptado de *Regresión lineal*, recuperado de http://cristian415.wikispaces.com/file/view/5+-+Regresión+Lineal+-+Curvilinea+-+Multiple+(sin+PH).pdf, por M. Rodríguez, 2012.

En SPSS estas transformaciones se generan seleccionando en el menú Transformar la opción Calcular. Posteriormente se procede a crear la/s variable/s transformada/s con la/s cual/es se realiza un nuevo análisis de regresión. Según Universidad Rafael Urdaneta (s. f.) la función a aplicar para crear dicha/s variable/s según las transformaciones Logarítmica en *x*, Logarítmica en *y*, Cuadrática en *x*, Cuadrática en *y*, Raíz cuadrada de *x*, Raíz cuadrada de *y*, Inversa en *x*, y Recíproca en *y* es según cada caso como sigue[271]:

- Logarítmica en x_1: $x_1 = \ln(x_1)$
- Logarítmica en *y*: $y = \ln(y)$
- Cuadrática en x_1: $z_2 = x_1^2$; ejecutando el comando x_1**2
- Cuadrática en *y*: $y = y^2$; ejecutando el comando *y***2
- Raíz cuadrada de x_1: $x_1 = \sqrt{x_1}$; ejecutando el comando *SQRT*(x_1)
- Raíz cuadrada de *y*: $y = \sqrt{y}$; ejecutando el comando *SQRT*(*y*)
- Inversa en x_1: $x_1 = 1/x_1$
- Recíproca en *y*: $y = 1/y$

Si el modelo transformado es apto entonces se puede utilizar para predecir a *y*. En la regresión lineal simple al sustituir *x* por un valor determinado se resuelve la ecuación estimada de regresión y se obtiene el valor *y* resultante. En la regresión múltiple tras sustituir un valor determinado por las *x* del modelo se soluciona la ecuación estimada de regresión y se halla el valor *y* correspondiente. En tanto, la interpretación del/de los coeficiente/s de regresión demanda ciertas consideraciones. Sánchez Mangas, R.

[271] La/s variable/s a crear según cada transformación se presenta en función de un modelo simple (con una única variable independiente) para simplificar la explicación. Si en un modelo múltiple en vez de x_1 se desea transformar por ejemplo x_3 se aplica la transformación en cuestión a x_3 lo cual se expresa sustituyendo en la fórmula de la variable creada a x_1 por x_3.

(s. f.) indica que para las siguientes transformaciones dicha interpretación es:

- Logarítmica en x_1 [$y = b_0 + (b_{y1})(\ln x_1)$]:

  Cuando x_1 varía 1%, y varía en promedio en $(b_{y1})(0{,}01)$ unidades

- Logarítmica en y [$\ln y = b_0 + b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(b_{y1})(100)$ porciento

Igualmente Montero Lorenzo, J. M. (2007) indica que para la transformación del modelo compuesto la interpretación del coeficiente de regresión es:

- Transformación del modelo compuesto [$\ln y = \ln b_0 + \ln b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(\ln b_{y1})(100)$ porciento

Alonso, C. (s. f.) señala que conforme a las siguientes transformaciones la interpretación del coeficiente de regresión es:

- Inversa en x [$y = b_0 + (b_{y1})(1/x_1)$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(-b_{y1})(1/x^2)$ unidades

- Cuadrática en x_1 [$y = b_0 + b_{y1}z_1 + b_{y2}z_2$, donde $z_1 = x_1$ y $z_2 = x_1^2$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(b_{y1} + 2b_{y2}x_1)$ unidades

  Ni b_{y1} ni b_{y2} tienen interpretación por separado

En tanto, a partir de las interpretaciones del coeficiente de regresión establecidas por

Sánchez Mangas (s. f.), Montero Lorenzo (2007) y Alonso (s. f.) se aprecia que según las siguientes transformaciones la interpretación de dicho coeficiente es:

- Cúbica en x_1 [$y = b_0 + b_{y1}z_1 + b_{y2}z_2 + b_{y3}z_3$, donde $z_1 = x_1$, $z_2 = x_1^2$, y $z_3 = x_1^3$]:

  Al x_1 variar 1 unidad, y variará en promedio en ($b_{y1} + 2b_{y2}x_1 + 3b_{y3}x_1^2$) unidades si se cumplen las siguientes condiciones:

  - b_{y3} y b_{y1} son de un mismo signo (ambos positivos o negativos)
  - $b_{y2} \geq \sqrt{3b_{y3}b_{y1}}$

  Si se verifican los dos requisitos la fórmula general de la ecuación cuadrática $-b \pm \sqrt{(b^2 - 4ac)}/2a$ de la derivada $\Delta y/\Delta x = b_{y1} + 2b_{y2}x_1 + 3b_{y3}x_1^2 = 0$ genera dos soluciones reales. En ese caso se interpreta el coeficiente de regresión basado en la expresión $b_{y1} + 2b_{y2}x_1 + 3b_{y3}x_1^2$. Pero si no se satisface las dos condiciones el discriminante $b^2 - 4ac$ produce un valor negativo. De ese modo, según Sapiensman (s. f.) la fórmula general de la ecuación cuadrática arroja dos soluciones imaginarias. Dada esa circunstancia se invalida el uso de la expresión $b_{y1} + 2b_{y2}x_1 + 3b_{y3}x_1^2$ para interpretar el coeficiente de regresión. Cabe destacar que ni b_{y1} ni b_{y2} ni b_{y3} tienen interpretación por separado

- Cuadrática en y [$y^2 = b_0 + b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $\sqrt{b_{y1}}$ unidades

- Recíproca en y [$1/y = b_0 + b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $1/b_{y1}$ unidades

- Raíz cuadrada de y [$\sqrt{y} = b_0 + b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en b_{y1}^2 unidades

- Raíz cuadrada de x_1 [$y = b_0 + b_{y1}\sqrt{x_1}$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en b_{y1} unidades

- Potencia [$\ln\hat{y} = \ln b_0 + (b_{y1})(\ln x_1)$]:

  Cuando x_1 varía 1%, y varía en promedio en $(b_{y1})(100)(0{,}01)$ porciento. Eso equivale a indicar que cuando x_1 varía 1%, y varía en promedio en (b_{y1}) porciento

- G [$\ln\hat{y} = b_0 + (by_1)(1/x_1)$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(b_{y1})(100)$ porciento

- Crecimiento [$\ln\hat{y} = b_0 + b_{y1}x_1$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(b_{y1})(100)$ porciento

- Exponencial [$\ln\hat{y} = \ln b_0 + (b_{y1})(x1)$]:

  Cuando x_1 varía 1 unidad, y varía en promedio en $(b_{y1})(100)$ porciento

7.2.11. Multicolinealidad

Anderson y otros (2009) indican: "(…) El término *Variables independientes*[272] no significa que (…) [dichas] variables (…) sean estadísticamente independientes entre

[272] La cursiva es nuestra.

ellas (p. 644). De hecho, Anderson y otros (2009) señalan: "(…) La mayoría de las variables independientes están, en cierto grado, correlacionadas (…) [entre sí]. Conforme a Anderson y otros (2009) dicha correlación entre variables independientes se denomina *Multicolinealidad*[273].

Según Anderson y otros (2009)

> Por lo general la multicolinealidad no afecta la manera en que se realiza el análisis de regresión o en que se interpretan los resultados de un estudio. Pero si la multicolinealidad es severa se pueden tener dificultades al interpretar los resultados de las pruebas *t* acerca de cada uno de los parámetros. (…) [Igualmente] se ha demostrado que los casos severos de multicolinealidad [también] dan como resultado estimaciones por mínimos cuadrados con signo erróneo. Esto es, en estudios simulados en los que los investigadores crearon el modelo de regresión subyacente y después emplearon el método de mínimos cuadrados para obtener estimaciones de [$ß_{y1.2\ldots i}$, $ß_{y2.1\ldots i}$, $ß_{yp.12\ldots i}$] (…) se ha demostrado que en condiciones de fuerte multicolinealidad las estimaciones obtenidas por mínimos cuadrados pueden tener signo opuesto al del parámetro que se estima. Por ejemplo, [$ß_{y2.1\ldots i}$] (…) puede ser en realidad +10 y su estimación [$b_{y2.1\ldots i}$] (…) puede resultar ser -2. Por lo tanto si existe una fuerte multicolinealidad podrá tenerse poca confianza en los coeficientes (p. 645).

En ese sentido Anderson y otros (2009) agregan

> (…) En las pruebas *t* para la significancia de cada uno de los parámetros la dificultad ocasionada por la multicolinealidad (…) [puede hacer] posible concluir que ninguno de los parámetros es significativamente distinto de cero cuando la prueba *F* sobre la ecuación de regresión múltiple general indica que hay una relación significante. [Estos problemas no se verifican] (…) cuando existe poca correlación entre las variables independientes.
>
> Para determinar si la multicolinealidad es lo suficientemente alta para ocasionar

[273] La cursiva es nuestra.

problemas se han desarrollado diversas pruebas. De acuerdo con la prueba de la regla práctica la multicolinealidad es un problema potencial si el valor absoluto del coeficiente de correlación (…) es mayor a 0,7 para cualquier par de variables independientes.

(…) Siempre que sea posible, debe evitarse incluir variables independientes que estén fuertemente correlacionadas (p. 644).

Conforme a The University of Texas at Austin (s. f.) otras pruebas pertinentes son las denominadas *Tolerancia*[274] y *VIF*[275] (inflación de la varianza, por sus siglas en inglés). Por un lado, Tacq (1998) señala que la tolerancia se calcula por medio de la fórmula $1 - R^2$, siendo R^2 el coeficiente de determinación múltiple de una determinada variable independiente respecto al resto de las variables explicatorias del modelo. Por otro lado, The University of Texas at Austin (s. f.) indica que la inflación de la varianza (VIF) equivale al inverso de la tolerancia, o sea, a la fórmula $1 / 1 - R^2$. Así, según The University of Texas at Austin (s. f.) las variables con tolerancia por debajo de 0,10 o con inflación de la varianza (VIF) por encima de 10, indicarán la presencia de multicolinealidad grave con otra variable del modelo.

Igualmente se puede analizar el índice de condición y la proporción de varianza. En ese sentido Universidad Complutense Madrid (s. f.) establece

"Los índices de condición son la raíz cuadrada del cociente entre el autovalor más grande y cada uno del resto de los autovalores. (…) Indices mayores que 30 delatan un serio problema de (…) [multicolinealidad]. Las proporciones de varianza recogen la proporción de varianza de cada coeficiente (…) que está explicada por cada dimensión o factor. En condiciones de no-colinealidad cada dimensión suele explicar gran cantidad de varianza de un solo coeficiente (excepto

[274] La cursiva es nuestra.
[275] Ibidem.

> en lo que se refiere al coeficiente $ß_0$ o constante, que siempre aparece asociado a uno de los otros coeficientes (…). [SPSS brinda ambos datos, los cuales conviene analizar en conjunto]. La colinealidad es un problema cuando una dimensión o factor con índice de condición alto contribuye a explicar gran cantidad de la varianza de los coeficientes de dos o más variables".

En caso de verificarse problemas de multicolinealidad se pueden tomar varias medidas. Conforme a Tacq (1998) se puede eliminar del modelo una o varias variables independientes, por ejemplo, aquellas que están fuertemente relacionadas con otras.

Según Tacq (1998) otro método consiste en asignar la varianza común a una de las variables independientes y removerla de las otras. En el caso de un modelo con dos variables independientes (x_1 y x_2) sería mantener a x_1 como está y eliminar la parte que x_2 comparte con x_1. Eso implica que el residuo $x_2 - x_2$ sombrero se conserva en el modelo, lo que conduce a un análisis de regresión que contempla a y como función de x_1 y $x_2 - x_2$ sombrero.

Asimismo según Preacher, K. J. (2003) centrar las variables es particularmente útil para reducir la multicolinealidad cuando hay variables cuadráticas o cúbicas en el modelo. Según Preacher (2003) dicho centrado se logra restando a cada variable su media.

En tanto, de acuerdo a Tacq (1998) también se puede llevar a cabo un análisis de componentes principales (ACP, por sus siglas en español). Según Tacq (1998) dicha técnica consiste en agrupar a las variables originales del modelo en nuevas variables

llamadas *Componentes*[276] o *Factores*[277]. De acuerdo a Tacq (1998) tales factores pudieran estar mutuamente correlacionados, pero se puede elegir un procedimiento de análisis factorial en que los factores sean ortogonales. Tacq (1998) indica que esos componentes son menores en número en comparación a las referidas variables independientes originales. Conforme a Tacq (1998) cada factor actúa como un representante de cada grupo de variables que están fuertemente asociadas, así, si se conduce un análisis de regresión múltiple con *y* como variable dependiente y los factores como variables independientes el problema de multicolinealidad se resolverá ya que los componentes no están correlacionados entre sí. Según Tacq (1998) de todos modos se genera un obstáculo a partir de esta solución, pues se debe hallar un nombre que represente a las variables de cada factor generado y además inferir el significado de dicho componente.

La manera en que se realiza un ACP se ilustra en el apartado *Caso de regresión múltiple en SPSS con análisis de componentes principales como solución de la multicolinealidad*[278], donde figura el análisis de un ejemplo realizado a través de SPSS. No obstante, en las líneas siguientes se aborda los criterios que se deben tener en cuenta cuando se efectúa un ACP. En esta tesis solo se expone el modo de llevar a cabo un ACP a través de SPSS, porque es un mecanismo más usado que el cálculo manual. La verificación del cálculo manual se puede realizar en diversos textos estadísticos destacando los autores Tacq (1998) y López Pérez (2005) cuyas referencias se detalla en la bibliografía de esta tesis.

En tanto, según López Pérez (2005) existen varios supuestos para realizar un ACP.

[276] La cursiva es nuestra.
[277] Ibidem.
[278] Ibidem.

Conforme a López Pérez (2005): “Las variables deben ser cuantitativas, (…) los datos deben tener una distribución normal bivariada para cada pareja de variables y las observaciones deben ser independientes” (p. 505). Asimismo The University of Texas at Austin (s. f.) señala que el ACP se puede realizar siempre que el tamaño de la muestra sea mayor a 50 casos. Igualmente The University of Texas at Austin (s. f.) advierte que el análisis cuyo tamaño de muestra sea mayor a 50 y menor a 100 casos debe interpretarse con cautela y, a la vez, indicar dicha precaución en el referido análisis.

De la misma forma una vez que el análisis se realiza en SPSS se debe constatar el cumplimiento de otros supuestos. The University of Texas at Austin (s. f.) establece que al menos algunas variables contempladas en el ACP deben tener una correlación mayor de 0,30. “También se exige que el determinante de la matriz de los datos iniciales sea muy pequeño [cercano a 0] para que realmente exista la opción de poder [aplicar el análisis]” (López Pérez, 2005, p. 505). “[Si el determinante es cercano a 0] indica que el grado de intercorrelación entre las variables es muy alto” (López Pérez, 2005, p. 508). De igual modo López Pérez (2005) señala que el estadístico KMO debe ser grande (cercano a 1), lo que indica una buena adecuación de la muestra al análisis. En ese sentido The University of Texas at Austin (s. f.) indica que el valor KMO debe ser mayor que 0,50 para cada variable individual y para el conjunto de variables. En tanto, The University of Texas at Austin (s. f.) establece que la prueba de esfericidad de Bartlett debe ser significativa con una probabilidad asociada menor que el nivel de significancia considerado. “[Dicha prueba] (…) permite contrastar formalmente la existencia de correlación entre las variables” (López Pérez, 2005, p. 508). Igualmente López Pérez (2005) señala: “Los elementos de la diagonal de (…)

[la] matriz [anti-imagen] son similares al estadístico KMO para cada par de variables e interesa que estén cercanos a (1)" (p. 509).

Si se satisface los supuestos, se puede aceptar el ACP. En ese caso los resultados presentados por SPSS brindan la información pertinente para determinar si las variables independientes se pueden agrupar en un conjunto de variables menor. De acuerdo a López Pérez (2005) la tabla *Varianza total explicada*[279] indica el porcentaje de la varianza que cada componente explica y la varianza explicada por el conjunto de componentes resultantes. Según The University of Texas at Austin (s. f.) la varianza total explicada por dicho grupo de componentes debe ser igual o mayor que 60%. Igualmente López Pérez (2005) indica que la tabla *Varianza total explicada*[280] también permite determinar la cantidad de componentes que se tomará aceptando solo los componentes cuyo autovalor sea mayor que 1. Asimismo, López Pérez (2005) señala que la tabla *Comunalidades*[281] expone para cada variable la parte de su variabilidad que es explicada por los factores. Conforme a The University of Texas at Austin (s. f.) cada variable con comunalidad menor que 0,50 se debe eliminar del ACP. Según The University of Texas at Austin (s. f.) cuando se logra que la comunalidad de todas las variables sea mayor que 0,50, se verifica que la *Matriz de componentes rotados*[282] solo exhiba variables con la denominada *Estructura simple*[283]. De acuerdo a The University of Texas at Austin (s. f.) el concepto *Estructura simple*[284] se refiere a que una variable debe tener una correlación mayor o igual que 0,40 en un único componente. The University of Texas at Austin (s. f.) señala que toda variable que se identifique con estructura compleja, o sea, que tenga

[279] La cursiva es nuestra.
[280] Ibidem.
[281] Ibidem.
[282] Ibidem.
[283] Ibidem.
[284] Ibidem.

una correlación mayor que 0,40 en más de un componente, debe eliminarse del análisis. Finalmente, The University of Texas at Austin (s. f.) establece que luego se revisa las comunalidades nuevamente para asegurar que se está explicando una porción suficiente de la varianza de todas las variables originales, es decir para garantizar que todas las variables de los componentes puntúen mayor que 0,50. Posteriormente el resultado del ACP se observa en la *Matriz de componentes rotados*[285], la cual muestra cómo se agrupan las variables en función de los diferentes componentes.

7.2.12. Prueba de Significancia en regresión múltiple

Anderson y otros (2009) afirman

> Las pruebas de significancia que se usaron en la regresión lineal simple fueron la prueba *t* y la prueba *F*. En (…) [dicho tipo de] regresión (…) estas dos pruebas llevan a la misma conclusión, es decir, si se rechaza la hipótesis nula se concluye que $b_{y1} \neq 0$. [Sin embargo] en la regresión múltiple la prueba *t* y la prueba *F* tienen propósitos diferentes:
>
> 1) La prueba *F* se usa para determinar si existe una relación de significancia entre la variable dependiente y el conjunto de todas las variables independientes. [Por ende] a esta prueba se le llama *Prueba de significancia global*[286].
>
> *2)* Si la prueba *F* indica que hay significancia global se (…) [puede usar] la prueba *t* para ver si cada una de las variables individuales es significativa. [En ese sentido] para cada una de las variables independientes (…) se realiza una prueba *t*. A cada una de estas pruebas *t* se le conoce como *Prueba de significancia individual*[287]" (p. 640).

[285] La cursiva es nuestra.
[286] Ibidem.
[287] Ibidem.

7.2.12.1. Prueba *F* en regresión múltiple

Conforme a Anderson y otros (2009) ya que en regresión múltiple $y = ß_0 + ß_{y1.2\ldots i}x_1 + ß_{y2.1\ldots i}x_2 + \ldots + ß_{yp.12\ldots i}x_p + \varepsilon$ es el modelo de regresión, la hipótesis de la prueba *F* comprende los parámetros del mismo. Así:

$$H_0: ß_{y1.2\ldots i} = ß_{y2.1\ldots i} = \ldots = ß_{yp.12\ldots i} = 0$$

$$H_a: \text{Uno o más de los parámetros es distinto de cero}$$

Según Anderson y otros (2009)

> Cuando se rechaza H_0 la prueba proporciona evidencia estadística suficiente para concluir que uno o más de los parámetros no es igual a cero y que la relación global entre *y* y el conjunto de variables independientes $x_1, x_2, \ldots, x_p$ es significativa. En cambio, si no se puede rechazar H_0, no se tiene evidencia suficiente para concluir que exista una relación significativa.
>
> Antes de describir los pasos de la prueba *F*, es necesario revisar el concepto de *Cuadrado medio*[288]. Un cuadrado medio es una suma de cuadrados dividida entre sus correspondientes grados de libertad. En el caso de la regresión múltiple la suma de cuadrados del total tiene $n - 1$ grados de libertad, la suma de cuadrados debida a la regresión (SCR) tiene p grados de libertad y la suma de cuadrados debida al error [(SCE)] tiene $n - p - 1$ grados de libertad. Por tanto el cuadrado medio debido a la regresión (CMR) es SCR/p y el cuadrado medio debido al error (CME) es SCE/$(n - p - 1)$. [Así:]

$$\text{CMR} = \frac{\text{SCR}}{p}$$

[288] La cursiva es nuestra.

y

$$\text{CME} = \frac{\text{SCE}}{n - p - 1}$$

Como se vio [en el apartado *Estimación de σ^2 y error de estimación en regresión lineal simple*[289]] (…) CME proporciona una estimación insesgada de σ^2, la varianza del término del error. Si $H_0 = ß_{y1.2\ldots i} = ß_{y2.1\ldots i} = \ldots = ß_{yp.12\ldots i} = 0$ es verdadera, CMR también proporciona un estimador insesgado de σ^2 y el valor de CMR/CME será cercano a 1. Pero si H_0 es falsa el CMR sobreestima [la] σ^2 y el valor de CMR/CME será mayor. Para determinar qué tan grande debe ser CMR/CME para que se rechace H_0 se hace uso del hecho de que, si H_0 es verdadera y las suposiciones acerca del modelo de regresión múltiple son válidas, la distribución muestral de CMR/CME es una distribución F con p grados de libertad en el numerador y $n - p - 1$ en el denominador. (…) [La figura 32] presenta un resumen de la prueba F de significancia para la regresión múltiple (p. 641).

Figura 32. Resumen de la prueba F de significancia para la regresión múltiple[290]

HIPÓTESIS NULA Y ALTERNATIVA EN LA PRUEBA F

$H_0 = ß_{y1.2\ldots i} = ß_{y2.1\ldots i} = \ldots = ß_{yp.12\ldots i} = 0$

H_a: *uno o más de los parámetros no son iguales a cero.*

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{CME}}$$

REGLA DE RECHAZO

Valor aproximado a p: Rechazar H_0 si valor $p \leq \alpha$

Valor crítico aproximado: Rechazar H_0 si $F \geq F_\alpha$

donde F_α pertenece a la distribución F con p grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador.

[289] La cursiva es nuestra.

[290] Adaptado de *Estadística para administración y economía*, (p. 641), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Igualmente, a modo de esquema, se muestra información pertinente a la prueba *F* en la figura 33.

Figura 33. Imagen tabla ANOVA para modelo de regresión múltiple con *p* variables independientes[291]

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*
Regresión	SCR	p	$CMR = \frac{SCR}{p}$	$F = \frac{CMR}{CME}$
Error	SCE	$n - p - 1$	$CME = \frac{SCE}{n - p - 1}$	
Total	STC	$n - 1$		

7.2.12.2. Prueba *t* en regresión múltiple

Como se mencionó en el apartado *Prueba F en regresión lineal simple*[292], Anderson y otros (2009) afirman

> Si la prueba *F* indica que la relación de regresión múltiple es significativa se puede realizar una prueba *t* para determinar la significancia de cada uno de los parámetros. (…) [La figura 34] presenta la prueba *t* de significancia para cada uno de los parámetros (p. 643).

[291] De *Estadística para administración y economía*, (p. 643), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[292] La cursiva es nuestra.

Figura 34. Prueba *t* de significancia para cada uno de los parámetros[293]

HIPÓTESIS NULA Y ALTERNATIVA EN LA PRUEBA *t*

$H_0: \beta_{yp.i} = 0$
$H_a: \beta_{yp.i} \neq 0$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_{yp.i}}{s_{b_{yp.i}}}$$

REGLA DE RECHAZO

Método del valor-*p*: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ se toma de la distribución *t* con n - *p* - 1 grados de libertad.

Anderson y otros (2009) establecen: "En el estadístico de prueba (…) [el componente que se ilustra en la figura (35)] (…) es la estimación de la desviación estándar de $b_{yp.i}$" (p. 643).

Figura 35. Componente del estadístico de prueba *t* que estima la desviación estándar de *byp.i*[294]

$$s_{b_{yp.i}}$$

De esa manera según Levine y otros (2006) la prueba *t* evalúa cada una de las variables del modelo de regresión múltiple. Así, por ejemplo, para determinar si la variable independiente x_1 tiene efecto significativo en la variable dependiente *y* las hipótesis nula y alternativa son:

[293] Adaptado de *Estadística para administración y economía*, (p. 643), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[294] Adaptado de *Estadística para administración y economía*, (p. 643), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

$$H_0: ß_{y1.2\ldots i} = 0$$

$$H_1: ß_{y1.2\ldots i} \neq 0$$

Así, Anderson y otros (2009) indican que a partir de la ecuación que se ilustra en la figura 36, un nivel de significancia dado y $n - p - 1$ grados de libertad se determina si existe una relación significativa entre x_1 y y. Conforme a Levine y otros (2006) si se establece una relación estadísticamente significante se puede concluir que existe una relación lineal entre x_1 y y.

Figura 36. Estadístico de prueba t en función de la variable independiente x_1[295]

$$t = \frac{b_{y1.2\ldots i}}{s_{b_{y1.2\ldots i}}}$$

En tanto, para determinar si la variable independiente x_2 tiene efecto significativo en la variable dependiente y las hipótesis nula y alternativa son:

$$H_0: ß_{y2.1\ldots i} = 0$$

$$H_1: ß_{y2.1\ldots i} \neq 0$$

Como se mencionó arriba al analizar la relación entre x_1 y y, Anderson y otros (2009) indican que a partir de la ecuación que se ilustra en la figura 37, un nivel de significancia dado y $n - p - 1$ grados de libertad se determina si existe una relación

[295] Adaptado de *Estadística para administración y economía*, (p. 643), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

significativa entre x_2 y y. Igualmente según Levine y otros (2006) si se establece una relación estadísticamente significante se puede concluir que existe una relación lineal entre x_2 y y.

Figura 37. Estadístico de prueba t en función de la variable independiente x_2[296]

$$t = \frac{b_{y2.1\ldots i}}{s_{b_{y2.1\ldots i}}}$$

De acuerdo a Levine y otros (2006)

> (…) La prueba de significancia para un coeficiente de regresión en particular es, en realidad, una prueba de la significancia de añadir una variable específica a un modelo de regresión dado en el que ya se encuentra incluida (…) otra variable. Por lo tanto la prueba *t* del coeficiente de regresión equivale a probar la aportación de cada variable independiente (p. 479).

7.2.13. Determinación de cuándo agregar o eliminar variables en regresión múltiple

Anderson y otros (2009) indican que la prueba *F* sirve para determinar si conviene agregar una o más variables independientes al modelo de regresión múltiple. Según Anderson y otros (2009): "Esta prueba se basa en determinar la disminución del valor de la suma de cuadrados debidos al error al agregar una o más variables independientes al modelo" (p. 710).

Conforme a Anderson y otros (2009)

[296] Adaptado de *Estadística para administración y economía*, (p. 643), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

> Se [utilizará] (…) SCE(x_1) para denotar la suma de cuadrados debidos al error cuando x_1 es la única variable independiente del modelo, SCE(x_1, x_2) para denotar la suma de cuadrados debidos al error cuando tanto x_1 como x_2 son las variables del modelo y así sucesivamente. Por tanto la disminución del valor de la SCE que se obtuvo al adicionar x_2 al modelo que solo tenía como variable independiente a x_1 es:
>
> $$\text{SCE}(x_1) - \text{SCE}(x_1, x_2)$$
>
> (…) Para determinar si esta reducción es significativa se realiza una prueba *F*. El numerador del estadístico *F* es la disminución en el valor de SCE dividida entre la cantidad de variables independientes agregadas al modelo original (pp. 710-711).

Anderson y otros (2009) indican que si, por ejemplo, se agrega una sola variable, x_2, el numerador del estadístico *F* es:

$$\frac{\text{SCE}_{(x_1)} - \text{SCE}_{(x_1, x_2)}}{1}$$

"Lo que se obtiene es una medida de la disminución de SCE por variable independiente añadida al modelo" (Anderson y otros, 2009, p. 711).

En tanto, "El denominador del estadístico *F* es el cuadrado medio debido al error en el modelo que contiene (…) [el mayor número de] variables independientes. [Por ende:]" (Anderson y otros, 2009, p. 711)

$$\text{CME} = \frac{\text{SCE}_{(x_1, x_2)}}{n - p - 1}$$

Anderson y otros (2009) señalan que, como en el caso en cuestión, si el modelo de regresión múltiple que contiene la mayor cantidad de variables independientes tiene solo dos, x_1 y x_2, entonces $p = 2$.

Según Anderson y otros (2009) la siguiente fórmula integra tanto al numerador como al denominador recién descritos:

$$F = \frac{\dfrac{\text{SCE}_{(x_1)} - \text{SCE}_{(x_1, x_2)}}{1}}{\dfrac{\text{SCE}_{(x_1, x_2)}}{n - p - 1}}$$

Como se mencionó en el apartado *Prueba F en regresión lineal simple*[297] Anderson y otros (2009) destacan: "El número de grados de libertad en el numerador de este estadístico *F* es igual al número de variables agregadas al modelo y el número de grados en el denominador es igual a $n - p - 1$" (p. 711).

Anderson y otros (2009) establecen que se rechaza la hipótesis de que x_2 no sea estadísticamente significativa cuando el valor calculado a través del estadístico *F* resulta mayor que el valor encontrado a través de la tabla de la distribución de *F* para un nivel de significancia determinado.

Según Anderson y otros (2009)

> Cuando se desea probar la significancia de agregar solo una variable independiente al modelo, el resultado que se obtiene con la prueba *F* que se acaba de describir

[297] La cursiva es nuestra.

también se obtiene con la prueba *t* para la significancia de uno solo de los parámetros (…). El estadístico *F* que se acaba de calcular es el cuadrado del estadístico *t* que se usa para probar la significancia de un solo parámetro.

Puesto que, cuando se agrega una sola variable independiente al modelo, la prueba *t* es equivalente a la prueba *F*, esto permite aclarar el uso adecuado de la prueba *t* para probar la significancia de uno de los parámetros. Si uno de los parámetros no es significativo la variable correspondiente puede ser eliminada del modelo (p. 711).

Pero Anderson y otros (2009) agregan

(…) [A pesar de que] la prueba *t* [indique] (…) que hay dos o más parámetros que no son significativos nunca se debe eliminar del modelo más de una variable independiente con base en la prueba *t*; cuando se elimina una variable puede resultar que una segunda variable, que inicialmente no era significativa, se vuelva significativa.

(…) Cabe considerar si la adición de más de una variable independiente – como conjunto – da como resultado que haya una reducción significativa de la suma de los cuadrados debidos al error (pp. 711-712).

Anderson y otros (2009) proponen considerar el siguiente modelo de regresión múltiple:

$$y = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}x_2 + \varepsilon$$

Según Anderson y otros (2009)

Si a este modelo se le agregan las variables (…) [x_3] y (…) [x_4] se obtiene (…) [el modelo que sigue]:

$$y = ß_0 + ß_{y1.234}x_1 + ß_{y2.134}x_2 + ß_{y3.124}x_3 + ß_{y4.123}x_4 + \varepsilon$$

Para probar si la adición de (…) [x_3] y (…) [x_4] es estadísticamente significativa las hipótesis nula y alternativa pueden plantearse (…) [así]:

H_0: $ß_{y3.124} = ß_{y4.123} = 0$

H_a: Uno o más de los parámetros no es igual a cero

El siguiente estadístico *F* aporta la base para probar si la adición de estas variables independientes es estadísticamente significativa.

$$F = \frac{\dfrac{\text{SCE}_{(x_1, x_2)} - \text{SCE}_{(x_1, x_2, x_3, x_4)}}{\text{Variables agregadas al modelo}}}{\dfrac{\text{SCE}_{(x_1, x_2, x_3, x_4)}}{n - p - 1}}$$

Este valor *F* calculado se compara con F_α, el valor en la tabla [*F*] (…). Si $F > F_\alpha$ se rechaza H_0 y se concluye que el conjunto de variables independientes agregadas es estadísticamente significativo.

(…) Para dar una descripción un poco más sencilla de este cociente *F*, al modelo que tiene la menor cantidad de variables independientes se le denomina *Modelo reducido*[298] y al modelo que tiene la mayor cantidad de variables independientes se le denomina *Modelo completo*[299]. Si SCE(reducido) denota la suma de los cuadrados debido al error del modelo reducido y SCE(completo) denota la suma de los cuadrados debido al error del modelo completo, el numerador (…) [del estadístico de prueba *F*] se expresa como:

[298] La cursiva es nuestra.
[299] Ibidem.

$$\frac{\text{SCE}_{(\text{reducido})} - \text{SCE}_{(\text{completo})}}{\text{Número de términos extra}}$$

> (…) [Obsérvese] que "número términos extra" denota la diferencia entre el número de variables independientes en el modelo completo y el número de variables independientes en el modelo reducido. El denominador (…) [que acompaña al numerador arriba presentado] es la suma de los cuadrados debidos al error en el modelo completo dividida entre los (…) [$n - p - 1$] (p. 712).

De esa manera, conforme a Anderson y otros (2009), el denominador recién descrito representa el cuadrado medio debido al error en el modelo completo. Por ende al denotarse el cuadrado medio debido al error del modelo completo como CME(completo) se puede expresar:

$$F = \frac{\dfrac{\text{SCE}_{(\text{reducido})} - \text{SCE}_{(\text{completo})}}{\text{Número de términos extra}}}{\text{CME}_{(\text{completo})}}$$

Anderson y otros (2009) ilustran con un ejemplo el uso del estadístico *F*

> (…) [Supóngase] que se tiene un problema de regresión que tiene 30 observaciones. En un modelo en el que intervienen las variables independientes x_1, x_2 y x_3 la suma de los cuadrados debida al error es 150, y en un segundo modelo en el que las variables independientes son x_1, x_2, x_3, x_4, x_5 la suma de los cuadrados debida al error es 100. ¿La adición de las variables x_4 y x_5 produjo una reducción significativa de la suma de los cuadrados debida al error?
>
> (…) [Obsérvese] primero que (…) [$n - p - 1 = 30 - 5 - 1 = 24$], entonces CME(completo) = 100/24 = 4,17. Así, el estadístico *F* es:

$$F = \frac{\frac{150 - 100}{2}}{4{,}17} = 6$$

Este valor *F* que se ha calculado se compara con el valor *F* que se encuentra en la tabla para 2 grados de libertad en el numerador y 24 grados de libertad en el denominador. Para el nivel de significancia 0,05 en la tabla (…) [*F*] se encuentra $F_{0,05}$ = 3,40. Como *F* = 6,00 es mayor que 3,40 se concluye que la adición de las variables x_4 y x_5 es estadísticamente significativa (p. 713).

7.2.13.1. Uso del valor-*p* en regresión múltiple

De acuerdo a Anderson y otros (2009)

También puede usarse el criterio del valor-*p* para determinar si resulta ventajoso agregar una o más variables independientes a un modelo de regresión múltiple. En el ejemplo anterior se mostró cómo realizar la prueba *F* para determinar si la adición de dos variables independientes, x_4 y x_5, a un modelo con tres variables independientes, x_1, x_2 y x_3, era estadísticamente significativo. En ese ejemplo el valor que se obtuvo para el estadístico *F* fue 6,00 y se concluyó (por comparación de *F* = 6,00 con el valor crítico $F_{0,o5}$ = 3,40) que la adición de las variables x_4 y x_5 era significativa. El valor-*p* que corresponde a *F* = 6,00 (2 grados de libertad en el numerador y 24 grados de libertad en el denominador) es 0,008. Como el valor-*p* = 0,008 < α = 0,05, también se concluye que la adición de las dos variables independientes es significativa (p. 713).

7.2.13.2. Análisis de un problema mayor en regresión múltiple

Para mostrar en detalle los temas vinculados a la selección de variables en un modelo Anderson y otros (2009) plantean analizar un conjunto de datos que consta de 25 observaciones con 8 variables independientes. Así, Anderson y otros (2009) indican

> El doctor David W. Cravens del Departamento de Marketing de la Texas Christian University otorgó el permiso para emplear estos datos. Por esta razón a este conjunto de datos se le llamará datos de Cravens.
>
> Los datos de Cravens son de una empresa que tiene varios territorios de ventas, cada uno de los cuales le está asignado a un solo representante de ventas. Para determinar si diversas variables (independientes) predictoras podían explicar las ventas en cada uno de los territorios se realizó un análisis de regresión. A partir de una muestra de 25 territorios se obtuvieron los datos que se muestran en la tabla [18] (…); en la tabla [19] (…) se presenta la definición de las variables.
>
> Como paso preliminar se considerarán los coeficientes de correlación entre cada par de variables. En la figura [38] (…) se presenta la matriz de correlación [entre las variables] (…). Observe que el coeficiente de correlación (…) entre Sales y Time es 0,623, entre Sales y Poten es 0,598 y así sucesivamente.
>
> Si observa los coeficientes de correlación entre las variables independientes se dará cuenta de que la correlación entre Time y Accounts es 0,758; por tanto, si Accounts se usa como una de las variables independientes, Time no agregaría mucho poder explicatorio al modelo. Recuerde [que] (…) la multicolinealidad puede causar problemas si el valor absoluto del coeficiente de correlación (…), entre cualesquiera dos de las variables independientes, es mayor que 0,7 (p. 717).

Según Anderson y otros (2009) siempre que se pueda se evitará incluir, al mismo tiempo, a las variables Time y Accounts en el modelo. "También el coeficiente de correlación (…) entre Change y Rating, que es 0,549, es elevado y merece ser considerado más cuidadosamente" (Anderson y otros, 2009, p. 718).

Tabla 18. Datos de Cravens[300]

Ventas	Antigüedad	Potencial	GastPubl	Participación	Cambio	Cuentas	Trabajo	Evaluación
3.669,88	43,10	74.065,1	4.582,9	2,51	0,34	74,86	15,05	4,9
3.473,95	108,13	58.117,3	5.539,8	5,51	0,15	107,32	19,97	5,1
2.295,10	13,82	21.118,5	2.950,4	10,91	-0,72	96,75	17,34	2,9
4.675,56	186,18	68.521,3	2.243,1	8,27	0,17	195,12	13,40	3,4
6.125,96	161,79	57.805,1	7.747,1	9,15	0,50	180,44	17,64	4,6
2.134,94	8,94	37.806,9	402,4	5,51	0,15	104,88	16,22	4,5
5.031,66	365,04	50.9353	3.140,6	8,54	0,55	256,10	18,80	4,6
3.367,45	220,32	35.602,1	2.086,2	7,07	-0,49	126,83	19,86	2,3
6.519,45	127,64	46.176,8	8.846,2	12,54	1,24	203,25	17,42	4,9
4.876,37	105,69	42.053,2	5.673,1	8,85	0,31	119,51	21,41	2,8
2.468,27	57,72	36.829,7	2.761,8	5,38	0,37	116,26	16,32	3,1
2.533,31	23,58	33.612,7	1.991,8	5,43	-0,65	142,28	14,51	4,2
2.408,11	13,82	21.412,8	1.971,5	8,48	0,64	89,43	19,35	4,3
2.337,38	13,82	20.416,9	1.737,4	7,80	1,01	84,55	20,02	4,2
4.586,95	86,99	36.272,0	10.694,2	10,34	0,11	119,51	15,26	5,5
2.729,24	165,85	23.093,3	8.618,6	5,15	0,04	80,49	15,87	3,6
3,289.40	116,26	26.878,6	7.747,9	6,64	0,68	136,58	7,81	3,4
2.800,78	42,28	39.572,0	4.565,8	5,45	0,66	78,86	16,00	4,2
3.264,20	52,84	51.866,1	6.022,7	6,31	-0,10	136,58	17,44	3,6
3.453,62	165,04	58.749,8	3.721,1	6,35	-0,03	138,21	17,98	3,1
1.741,45	10,57	23.990,8	861,0	7,37	-1,63	75,61	20,99	1,6
2.035,75	13,82	25.694,9	3.571,5	8,39	-0,43	102,44	21,66	3,4
1.578,00	8,13	23.736,3	2.845,5	5,15	0,04	76,42	21,46	2,7
4.167,44	58,44	34.314,3	5.060,1	12,88	0,22	136,58	24,78	2,8
2.799,97	21,14	22.809,5	3.552,0	9,14	-0,74	88,62	24,96	3,9

[300] Adaptado de *Estadística para administración y economía*, (p. 718), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Tabla 19. Definición de las variables en los datos de Cravens[301]

Variable	Definición
Ventas	Total de ventas acreditadas al representante de ventas
Antigüedad (Time)	Antigüedad del empleado en meses
Potencial (Poten)	Potencial de mercado: ventas industriales totales en unidades en el territorio de ventas*
GastPubl (AdvExp)	Gastos del territorio en publicidad
Participación (Share)	Participación en el mercado: promedio ponderado de los últimos cuatro años
Cambio (Change)	Cambio, en los últimos cuatro años en participación en el mercado
Cuentas (Accounts)	Número de cuentas asignadas a los representantes de ventas*
Trabajo (Work)	Carga de trabajo: índice ponderado basado en compras anuales y concentración de cuentas
Evaluación (Rating)	Evaluación general del representante de ventas sobre ocho dimensiones de desempeño: una evaluación agregada en una escala del 1 -7

*Estos datos fueron codificados para proteger la confidencialidad.

Figura 38. Coeficientes de correlación de los datos de Cravens[302]

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0,623							
Poten	0,598	0,454						
AdvExp	0,596	0,249	0,174					
Share	0,484	0,106	-0,211	0,264				
Change	0,489	0251	0,268	0,377	0,085			
Accounts	0,754	0,758	0,479	0,200	0,403	0,327		
Work	0,117	0,179	-0,259	-0,272	0,349	-0,288	-0,199	
Rating	0,402	0,101	0,359	0,411	-0,024	0,549	0,229	-0,277

Anderson y otros (2009) indican

> Al observar los coeficientes de correlación (…) entre Sales y cada una de las variables independientes se puede tener una rápida idea de cuáles de las variables independientes son, en sí mismas, buenos predictores. Se encuentra que el mejor predictor de Sales es Accounts debido a que su coeficiente de correlación (…) es el más alto (0,754) (pp. 718-719).

Igualmente Anderson y otros (2009) establecen

[301] Adaptado de *Estadística para administración y economía*, (p. 718), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

[302] Adaptado de *Estadística para administración y economía*, (p. 719), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

> (…) Accounts explica (…) [el] 56,85% de la variabilidad en Sales. Las variables independientes que siguen en importancia son Time, Poten y AdvExp, cada una con un coeficiente de correlación (…) de 0,6, aproximadamente.
>
> [En tanto,] (…) el coeficiente de determinación ajustado para este modelo de regresión múltiple con ocho variables es 88,3%. Observe, sin embargo, que los valores-*p* de las pruebas *t* para cada uno de los parámetros indican que solo Poten, AdvExp y Share son significativos a un nivel de significancia $\alpha = 0{,}05$ (…). Por tanto se deseará investigar los resultados que se obtienen si se usan solamente estas tres variables. (pp. 719-720).

"(…) El coeficiente de determinación ajustado para esta ecuación estimada de regresión es 82,7% el cual, aunque no es tan bueno como el de la ecuación estimada de regresión con ocho variables, es alto" (Anderson y otros, 2009, p. 720).

En ese sentido, Anderson y otros (2009) plantean

> ¿Cómo se puede encontrar una ecuación estimada de regresión que dé mejores resultados, dada la información (…) que se dispone? Una posibilidad es calcular todas las regresiones posibles. Es decir, obtener ocho ecuaciones estimadas de regresión con una sola variable (cada una de las cuales corresponde a una de las variables independientes), 28 ecuaciones estimadas de regresión con dos variables independientes (que es el número de combinaciones de ocho variables tomadas de dos en dos) y así sucesivamente. Para los datos de Cravens se necesitan, en total, 255 ecuaciones estimadas de regresión conteniendo una o más de las variables independientes (p. 720).

Según Anderson y otros (2009) para elegir el subconjunto de variables independientes que proporcione la mejor ecuación estimada de regresión conviene aplicar alguno de los métodos más conocidos entre los que simplifican la selección. A saber:

- La regresión por pasos.
- La selección hacia adelante.
- La selección hacia atrás.
- La regresión del mejor subconjunto.

Anderson y otros (2009) afirman

> Dado un conjunto de datos en el que hay variables independientes estos procedimientos permiten determinar con qué variables independientes se obtiene el mejor modelo. Los tres primeros procedimientos son iterativos; en cada paso del procedimiento se agrega o se elimina una variable independiente y se evalúa el nuevo modelo. El procedimiento continúa hasta que un criterio de detención indica que el procedimiento ya no puede hallar un modelo mejor. El último procedimiento (mejores subconjuntos) no es un procedimiento que evalúe las variables de una en una, sino que evalúa modelos de regresión en los que intervienen distintos subconjuntos de variables independientes.
>
> En los procedimientos regresión por pasos, selección hacia adelante y eliminación hacia atrás, en cada paso el criterio para elegir una variable independiente, para agregarla o eliminarla del modelo, se basa en el estadístico F (…). Suponga, por ejemplo, que se desea considerar si agregar x_2 a un modelo en el que interviene x_1 o eliminar x_2 de un modelo en el que intervienen x_1 y x_2. Para probar si la adición o la eliminación de x_2 es estadísticamente significativa las hipótesis nula y alternativa pueden establecerse como sigue:
>
> $$H_0: ß_{y2.1} = 0$$
>
> $$H_a: ß_{y2.1} \neq 0$$
>
> [Anteriormente] (…) se mostró que:

$$F = \frac{\dfrac{\text{SCE}_{(x_1)} - \text{SCE}_{(x_1, x_2)}}{1}}{\dfrac{\text{SCE}_{(x_1, x_2)}}{n - p - 1}}$$

> [Igualmente se indicó que dicha fórmula es usada] (…) como criterio para determinar si la presencia de x_2 en el modelo causa una reducción significativa de la suma de los cuadrados debidos al error. El valor-*p* correspondiente a este estadístico *F* es el criterio que se emplea para determinar si se debe agregar o eliminar una variable independiente del modelo de regresión. Para el rechazo se emplea la regla usual: rechazar H_0 si valor-$p \leq \alpha$ (pp. 720-721).

7.2.13.2.1. Regresión por pasos

Conforme a Anderson y otros (2009)

> El procedimiento de regresión por pasos empieza por determinar en cada paso si alguna de las variables que ya se encuentran en el modelo debe ser eliminada. Para esto primero se calcula el estadístico *F* y el correspondiente valor-*p* para cada una de las variables independientes que intervienen en el modelo (p. 721)

Si el valor-*p* de alguna de las variables independientes resulta ser mayor que el valor del nivel de significancia establecido para determinar si una variable se debe eliminar del modelo, Anderson y otros (2009) indican: “La variable independiente que tenga el mayor valor-*p* se elimina del modelo y el proceso de regresión por pasos empieza un nuevo paso (p. 721).

Anderson y otros (2009) agregan

> Si ninguna de las variables independientes puede ser eliminada del modelo el procedimiento trata de ingresar otra variable independiente al modelo. Para hacer esto primero se calcula el estadístico *F* y el valor-*p* de cada variable independiente

> que no está en el modelo. (…) La variable independiente que tiene el menor valor-*p* es ingresada al modelo siempre que su valor-*p* sea menor que (…) [el valor del nivel de significancia establecido para ingresar alguna variable independiente a dicho modelo]. Este procedimiento continúa de la misma forma hasta que no haya ninguna variable independiente que pueda ser eliminada o agregada al modelo (p. 721).

Anderson y otros (2009) señalan que el procedimiento regresión por pasos aplicado a los datos de Cravens, con 0,05 como valor del nivel de significancia tanto para eliminar como para ingresar variables independientes al modelo, culminó en cuatro pasos y una ecuación estimada de regresión cuyo R^2 ajustado = 88,05, la cual se compone de la siguiente manera:

$$\hat{y} = -1.441{,}93 + 9{,}2\ \text{Accounts} + 0{,}175\ \text{AdvExp} + 0{,}0382\ \text{Poten} + 190\ \text{Share}$$

En resumen, de acuerdo a Anderson y otros (2009) en cada paso del procedimiento de regresión por pasos primero se determina si se puede eliminar alguna variable independiente del modelo que se tiene. Si eso no es posible el procedimiento verifica si se puede ingresar alguna de las variables independientes de las que no intervienen en el modelo. Conforme a Anderson y otros (2009)

> Debido a la naturaleza del procedimiento de regresión por pasos puede ser que una variable independiente sea ingresada al modelo en un paso, en un paso subsiguiente eliminada y después ingresada al modelo en un paso posterior. El procedimiento se detiene cuando no hay ya ninguna variable independiente que pueda ser eliminada del modelo ni agregada al modelo (p. 722).

7.2.13.2.2. Selección hacia adelante

Anderson y otros (2009) establecen

> En el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente y se van agregando variables de una en una con el mismo procedimiento que se usa en la regresión por pasos para determinar si una variable independiente debe ser ingresada al modelo. Pero en el procedimiento de selección hacia adelante no se permite que se elimine del modelo una variable que ha sido ingresada. El procedimiento se detiene cuando el valor-*p* de cada una de las variables independientes que no están en el modelo es mayor que (…) [el valor del nivel de significancia establecido para ingresar alguna variable independiente al modelo].
>
> La ecuación estimada de regresión obtenida mediante el procedimiento de selección hacia adelante (…) es:
>
> $$\hat{y} = -1.441{,}93 + 9{,}2\ \text{Accounts} + 0{,}175\ \text{AdvExp} + 0{,}0382\ \text{Poten} + 190\ \text{Share}$$
>
> Por tanto en el caso de los datos de Cravens, con el procedimiento de selección hacia adelante (con 0,05 como (…) [el valor del nivel de significancia establecido para ingresar alguna variable independiente al modelo]) se llega a la misma ecuación estimada de regresión que con el procedimiento por pasos (p. 722).

7.2.13.2.3. Eliminación hacia atrás

Según Anderson y otros (2009)

> En el procedimiento de eliminación hacia atrás se empieza con un modelo en el que se incluyen todas las variables independientes. Después, de una en una, se van eliminando variables independientes mediante el mismo procedimiento que en la regresión por pasos. Sin embargo, en el procedimiento de eliminación hacia atrás no se permite que una variable que ya ha sido eliminada vuelva a ser ingresada al

modelo. El procedimiento se detiene cuando ninguna de las variables independientes del modelo tenga un valor-*p* mayor que (…) [el valor del nivel de significancia establecido para eliminar alguna variable independiente al modelo].

La ecuación estimada de regresión obtenida con el procedimiento de eliminación hacia atrás (…) aplicado a los datos de Cravens (con 0,05 como (…) [el valor del nivel de significancia establecido para eliminar alguna variable independiente al modelo]) es:

$$\hat{y} = -1.312 + 3{,}8\ \text{Accounts} + 0{,}0444\ \text{AdvExp} + 0{,}152\ \text{Poten} + 259\ \text{Share}$$

Al comparar la ecuación estimada de regresión obtenida mediante el procedimiento de eliminación hacia atrás con la ecuación estimada de regresión obtenida con el procedimiento de selección hacia adelante, se ve que hay tres variables independientes comunes a los dos procedimientos: AdvExp, Poten y Share. Pero en el procedimiento de eliminación hacia atrás se incluyó Time en lugar de Accounts.

La selección hacia adelnate y la eliminación hacia atrás son dos extremos en la construcción de modelos; en el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente en el modelo y, una por una, se van agregando variables independientes, mientras que en el procedimiento de eliminación hacia atrás se empieza teniendo todas las variables independientes en el modelo y, de una en una, se eliminan variables. Con los dos procedimientos se puede llegar a ecuaciones estimadas de regresión diferentes, como ocurre en el caso de los datos de Cravens. ¿Por cuál de las ecuaciones estimadas de regresión decidirse? Esto es algo que queda a discusión. Al final el analista tiene que aplicar su propio criterio (p. 723).

7.2.13.2.4. Regresión de los mejores subconjuntos

De acuerdo a Anderson y otros (2009)

En (…) [los] resultados [de este método] aparecen las dos mejores ecuaciones de

> regresión estimada con una sola variable, las dos mejores ecuaciones con dos variables, las dos mejores ecuaciones con tres variables, etc. El criterio que se emplea para determinar cuáles son las mejores ecuaciones estimadas de regresión con un determinado número de predictores es el valor del coeficiente de determinación (…) [R^2]. Por ejemplo, [en la tabla 20] Accounts proporciona la mejor ecuación estimada de regresión con una sola variable independiente, (…) [R^2] = 56,8%; al usar AdvExp y Accounts se obtiene la mejor ecuación estimada de regresión con dos variables independientes, (…) [R^2] = 77,5%; y con Poten, AdvExp y Share se obtiene la mejor ecuación estimada de regresión con tres variables independientes, (…) [R^2] = 84,9%. Para los datos de Cravens el mayor coeficiente de determinación ajustado ([R^2] (…) = 89,4%) es el del modelo con seis variables independientes, Time, Poten, AdvExp, Share, Change y Accounts. Sin embargo el coeficiente de determinación ajustado del mejor modelo con cuatro variables independientes (Poten, AdvExp, Share y Accounts) es casi igual de alto (88,1%). Por lo general se prefiere el modelo más sencillo con el menor número de variables (p. 723).

Cabe destacar que SPSS no cuenta con este método de selección de variables. En ese sentido el ejemplo de los datos de Cravens está basado en resultados generados por el programa estadístico denominado *Minitab*[303].

[303] La cursiva es nuestra.

Tabla 20. Parte de los resultados obtenidos con la regresión de los mejores subconjuntos de Minitab[304]

Vars	R-sq	Adj. R-sq	S	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
1	56,8	55,0	881,09						X		
1	38,8	36,1	1049,3	X							
2	77,5	75,5	650,39			X			X		
2	74,6	72,3	691,11		X			X			
3	84,9	82,7	545,52		X	X	X				
3	82,8	80,3	582,64		X	X			X		
4	90,0	88,1	453,84		X	X	X		X		
4	89,6	87,5	463,93	X	X	X	X				
5	91,5	89,3	430,21	X	X	X	X	X			
5	91,2	88,3	436,75		X	X	X	X	X		
6	92,0	89,4	427,99	X	X	X	X	X	X		
6	91,6	88,9	438,20		X	X	X	X	X	X	
7	92,2	89,0	435,66	X	X	X	X	X	X	X	
7	92,0	88,8	440,29	X	X	X	X	X	X		X
8	92,2	88,3	449,29	X	X	X	X	X	X	X	X

7.2.13.2.5. Elección final

Anderson y otros (2009) establecen

> El análisis de los datos de Cravens hecho hasta ahora es una buena preparación para (…) [decidirse] por un modelo, pero antes habrá (…) también otros análisis. (…) [Así] es necesario hacer un (…) análisis de los residuales. Se desea que la gráfica de los residuales del modelo elegido parezca una banda horizontal. Suponga que en los residuales no se encuentre ningún problema y que se desee emplear los resultados del procedimiento de los mejores subconjuntos para decidirse por un modelo.
>
> El procedimiento de los mejores subconjuntos indica que el mejor modelo (…) es el que contiene las variables independientes Poten, AdvExp, Share y Accounts. Este modelo resulta ser el (…) [mismo] modelo (…) encontrado mediante el procedimiento de regresión por pasos (p. 724).

[304] Adaptado de *Estadística para administración y economía*, (p. 724), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Según Anderson y otros (2009)

> En la tabla [20] (…) se ve que el modelo que solo tiene AdvExp y Accounts es bueno. Su coeficiente de determinación ajustado es 75,5%, y con el modelo con las cuatro variables solo se logra un aumento de 12,6 puntos porcentuales. El modelo más sencillo que solo tiene dos variables puede preferirse si, por ejemplo, es difícil medir el potencial de mercado (Poten). Sin embargo, si ya se cuenta con los datos y se requiere gran precisión en la predicción de las ventas es claro que se preferirá el modelo con las cuatro variables [que tiene un coeficiente de determinación ajustado de 88,1%] (p. 725).

7.2.13.3. Variables confusoras

Conforme a de Irala, J., Martínez-González, M.A., y Guillén Grima, F. (2001) en el caso de la regresión, en términos generales, se habla de confusión cuando existen diferencias importantes entre las estimaciones brutas de los coeficientes de regresión antes de hacer un ajuste del modelo y las ajustadas por los posibles factores de confusión. Según de Irala y otros (2001) en un modelo de regresión se recomienda escoger toda variable que sea responsable de un cambio de más del 10% entre el valor de los coeficientes de regresión antes de ajustar el modelo (sin dicha variable en el modelo) y dicho valor después de realizar el ajuste (con la mencionada variable en el modelo), siempre que la variable en cuestión tenga un nivel de significación (valor-*p*) conservador aproximadamente menor de 0,20.

De acuerdo a de Irala y otros (2001) cualquier estimación que produzca un modelo de regresión se puede considerar ajustada por las variables que constituyen dicho modelo. Así, según de Irala y otros (2001) para ajustar un modelo de regresión por una variable confusora basta con introducir dicho factor de confusión en el referido

modelo. Por esa razón si se detecta una variable confusora conviene conservarla en el modelo aunque no haya sido identificada como significativa, siempre que su valor-*p* sea menor que 0,2.

7.2.14. Estimación del intervalo de confianza en regresión múltiple

Según Levine y otros (2006) para determinar en un modelo de regresión múltiple la estimación del intervalo de confianza para una pendiente poblacional, por ejemplo $\beta_{y1.2}$ (el efecto de x_1 sobre y, manteniendo constante el efecto de x_2), se utiliza la fórmula siguiente (donde cada $j = y1.2$):

$$b_j \pm t_{n-k-1} s_{b_j}$$

7.2.14.1. Uso de la ecuación de regresión estimada para estimaciones y predicciones en regresión múltiple

Anderson y otros (2009) señalan

> Los procedimientos empleados en la regresión múltiple para estimar el valor medio de y y para predecir el valor de un solo valor de y son similares a los empleados en el análisis de regresión para una sola variable independiente. Recuérdese, primero, que en (…) [el marco de la regresión lineal simple] se mostró que la estimación puntual del valor (…) [medio] de y para un valor dado de x y la estimación puntual de un solo valor de y es la misma. En ambos casos se usó como estimación puntual $\hat{y} = b_0 + b_{y1}x_1$. En la regresión múltiple se emplea el mismo procedimiento, es decir, los valores dados de $x_1, x_2, \ldots, x_p$ se sustituyen en la ecuación de regresión y como estimación puntual se emplea el correspondiente valor de $\hat{y}$. (…) Para obtener estimaciones por intervalo para el valor medio de y y

para un solo valor de *y* se emplean procedimientos similares a los empleados en el análisis de regresión con una sola variable (p. 647).

Conforme a Levine y otros (2006) si se desea determinar la estimación por intervalo para el valor medio de *y* se debe calcular el intervalo de confianza a través de la fórmula:

$$\hat{y} \pm t_{n-2} s_{yx} \sqrt{\frac{1}{n} + \frac{(x_i - x)^2}{\sum (x_i - x)^2}}$$

En cambio, Levine y otros (2006) indican que si se quiere determinar la estimación por intervalo para un solo valor de *y* se debe calcular el intervalo de predicción por medio de la fórmula:

$$\hat{y} \pm t_{n-2} s_{yx} \sqrt{1 + \frac{1}{n} + \frac{(x_i - x)^2}{\sum (x_i - x)^2}}$$

7.2.15. Coeficientes de regresión parcial estandarizados en regresión múltiple

Anteriormente se estableció conforme a Tacq (1998) que los coeficientes de regresión parcial no podían compararse entre sí y que ese sería el caso solo cuando las dispersiones de las variables independientes fueran iguales. De lo contrario Tacq (1998) indica que sería necesario estandarizar las variables independientes y realizar un análisis de regresión, ya que los coeficientes de regresión resultantes serán automáticamente comparables y la importancia de los efectos se podrá apreciar. Como se indicó en el apartado *Método de mínimos cuadrados en regresión*

múltiple[305], estos coeficientes de regresión son llamados *Coeficientes de regresión parcial estandarizados*[306] o *Beta*[307], y se denotan b^*.

Asumiendo que los ejemplos que se darán a continuación se plantean en el marco de un análisis de regresión múltiple con dos variables independientes (x_1 y x_2) y una dependiente *y*, Tacq (1998) indica cómo calcular los beta. Así, conforme a Tacq (1998) los puntajes de x_1 se transforman en puntajes z_1, los puntajes de x_2 pasan a z_2 y los puntajes de *y* se transforman en z_y como se ilustra en la figura 39.

Figura 39. Fórmulas para el cálculo de z_1, z_2 y z_y[308]

$$z_1 = \frac{(x_1 - \bar{x}_1)}{s_1}$$

$$z_2 = \frac{(x_2 - \bar{x}_2)}{s_2}$$

$$z_y = \frac{(y - \bar{y})}{s_y}$$

Después según Tacq (1998) se realiza el análisis de regresión múltiple entre las variables independientes estandarizadas z_1 y z_2 y la variable dependiente estandarizada z_y. Así, resulta una ecuación sin intercepto (pues los puntajes *z* implican desviaciones de la media y las desviaciones de la media implican una traslación al origen) y con coeficientes beta como coeficientes de regresión. De acuerdo a Tacq (1998) dicha ecuación es la siguiente:

$$z_y = b^*_{y1.2} z_1 + b^*_{y2.1} z_2 + e$$

[305] La cursiva es nuestra.
[306] Ibidem.
[307] Ibidem.
[308] Adaptado de *Multivariate analysis techniques in social science research*, (p. 123), por J. Tacq, 1998, Inglaterra Londres. SAGE Publications Ltd. Copyright 1997 por Jacques Tacq.

Sin embargo, Tacq (1998) propone un método más corto para calcular los beta. En ese sentido Tacq (1998) señala que es sabido que:

$$b_{yx} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Igualmente Tacq (1998) indica que en el caso de variables estandarizadas la media es igual a 0. Por ende según Tacq (1998) se puede establecer que:

$$b^*_{yx} = \frac{\text{covariación } z_x \text{ , } z_y}{\text{variación } z_x}$$

$$= \frac{\sum(z_x - 0)(z_y - 0)}{\sum(z_x - 0)^2}$$

$$= \frac{\sum z_x z_y}{\sum z_x^2} = \frac{\left[\frac{(x - \bar{x})}{s_x}\right]\left[\frac{(y - \bar{y})}{s_y}\right]}{\left[\frac{(x - \bar{x})}{s_x}\right]^2}$$

$$= \frac{(1/s_x s_y)\sum(x - \bar{x})(y - \bar{y})}{(1/s_x^2)\sum(x - \bar{x})^2}$$

De ese modo, conforme a Tacq (1998) se puede concluir que:

$$b^*_{yx} = \frac{s_x}{s_y} b_{yx}$$

En ese sentido de acuerdo a Tacq (1998) es fácil inferir que:

$$b^*_{y1.2} = \frac{s_{x_1}}{s_y} b_{y1.2}$$

Y que:

$$b^*_{y2.1} = \frac{s_{x_2}}{s_y} b_{y2.1}$$

Por ende el coeficiente beta es igual a su correspondiente coeficiente de regresión parcial multiplicado por el cociente entre la desviación estándar de la variable independiente *x* y la desviación estándar de la variable dependiente *y*. Según Tacq (1998) estos beta son apropiados para determinar la importancia relativa de los predictores x_1 y x_2.

Igualmente Tacq (1998) señala que los beta no deben ser mayor que uno. Sin embargo a veces se obtiene un beta mayor que uno. Una razón que frecuentemente explica eso es una excesivamente fuerte multicolinealidad (asociación entre los factores x_1 y x_2).

7.2.16. Variables cualitativas independientes

"(…) En muchas situaciones se tiene que trabajar con variables independientes cualitativas como (…) [sexo] (masculino o femenino), modo de pago (efectivo, tarjeta de crédito, cheque), etc." (Anderson y otros, 2009, p. 649).

"La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias [también llamadas indicadoras según Anderson y otros (2009)] para su transformación en métricas" (Pérez López, 2005, p. 89). Así, las variables independientes se recodifican de manera binaria incorporándose al análisis como si fueran variables numéricas.

7.2.16.1. Interpretación de los parámetros ante la presencia de variables cualitativas independientes

Para comprender mejor la interpretación de los parámetros Anderson y otros (2009) plantean el siguiente escenario. Una compañía determinada da servicio de mantenimiento a los sistemas de filtración en una localidad dada. Por cada solicitud de servicio los administradores desean predecir el tiempo requerido para ejecutar la reparación.

"Se cree que el tiempo requerido para (…) [la] reparación está relacionado con dos factores, meses transcurridos desde el último servicio de mantenimiento y tipo (…) [de] problema (mecánico o eléctrico)" (Anderson y otros, 2009, p. 650). Conforme a Anderson y otros (2009) *y* es el tiempo necesario para la reparación, medido en horas,

x_1 son los meses transcurridos desde el último mantenimiento, y x_2 es el tipo de problema.

De acuerdo a Anderson y otros (2009) para incorporar en el modelo de regresión múltiple a x_2 dicha variable se define como sigue:

$$X_2 = \begin{matrix} 0 \text{ si el tipo de reparación es mecánica} \\ 1 \text{ si el tipo de reparación es eléctrica} \end{matrix}$$

Según Anderson y otros (2009) el modelo de regresión múltiple para el ejemplo de arriba es:

$$y = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}x_2 + \varepsilon$$

En función de los datos de la tabla 21 Anderson y otros (2009) indican que la ecuación de regresión múltiple estimada es:

$$\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2$$

Tabla 21. Datos del ejemplo Reparación de sistemas de filtración[309]

Cliente	Meses transcurridos desde el último mantenimiento	Tipo de reparación	Tiempo, en horas, necesarias para la reparación
1	2	1	2,9
2	6	0	3,0
3	8	1	4,8
4	3	0	1,8
5	2	1	2,9
6	7	1	4,9
7	9	0	4,2
8	8	0	4,8
9	4	1	4,4
10	6	1	4,5

Anderson y otros (2009) establecen

> Empleando 0,05 como nivel de significancia el valor-*p* correspondiente al estadístico de prueba *F* (…) es 0,001, lo cual indica que la relación de regresión es significativa. (…) [Igualmente, según] la prueba *t*, se observa que tanto meses transcurridos desde el último servicio (valor-*p* = 0,000) como tipo de reparación (valor-*p* = 0,005) son estadísticamente significativos. Además (…) [R^2] = 85,9% y (…) [R^2 ajustado] = 81,9% indican que la ecuación de regresión estimada explica adecuadamente la variabilidad en el tiempo necesitado para la reparación (p. 651).

Anderson y otros (2009) señalan: "(…) Considérese el caso en que $x_2 = 0$ (reparación mecánica). Usando $E(y \mid \text{mecánica})$ para denotar la media (…) del tiempo necesario para una reparación (…) [debido a] que se trata de una reparación mecánica, se tiene" (p. 651):

$$E(y \mid \text{mecánica}) = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}(0) = ß_0 + ß_{y1.2}x_1$$

[309] Adaptado de *Estadística para administración y economía*, (p. 651), por D. R. Anderson, D. J. Sweeney y T. A. Williams, 2009, México D.F.: Cengage Learning, Inc. Derechos reservados 2008 por Cengage Learning Editores, S.A. de C.V., una compañía de Cengage Learning, Inc.

Anderson y otros (2009) agregan

De manera similar en el caso de una reparación eléctrica ($x_2 = 1$) se tiene:

$$E(y \mid \text{eléctrica}) = ß_0 + ß_{y1.2}x_1 + ß_{y2.1}(1)$$

$$= ß_0 + ß_{y1.2}x_1 + ß_{y2.1}$$

$$= (ß_0 + ß_{y2.1}) + ß_{y1.2}x_1$$

Comparando las ecuaciones (…) [en que $x_2 = 1$ y $x_2 = 0$] se ve que la media del tiempo requerido para hacer una reparación es función lineal de x_1 tanto cuando se trata de reparaciones mecánicas como de reparaciones eléctricas. La pendiente en ambas ecuaciones es $ß_{y1.2}$, pero la intersección con el eje y varía. (…) En la ecuación (…) para (…) reparaciones eléctricas la intersección es $(ß_0 + ß_{y2.1})$. La interpretación de $ß_{y2.1}$ es que indica la diferencia entre la media del tiempo que se requiere para una reparación eléctrica y la media del tiempo que se requiere para una reparación mecánica.

Si $ß_{y2.1}$ es positiva la media del tiempo necesario para una reparación eléctrica será mayor que para una reparación mecánica; si $ß_{y2.1}$ es negativa la media del tiempo requerido para una reparación eléctrica será menor que para una reparación mecánica. Por último, si $ß_{y2.1} = 0$, no hay diferencia entre las medias del tiempo que necesita para reparaciones eléctricas y mecánicas y el tipo de reparación no está relacionado con el tiempo necesario para hacer una reparación.

Empleando la ecuación de regresión múltiple estimada $\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2$ se ve que (…) cuando $x_2 = 0$ (reparación mecánica):

$$\hat{y} = 0{,}93 + 0{,}388x_1$$

Y cuando $x_2 = 1$ (reparación eléctrica):

$$\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26(1)$$

$$= 2{,}19 + 0{,}388x_1$$

> De esta manera, el uso de una variable ficticia proporciona dos ecuaciones que sirven para predecir el tiempo requerido para una reparación, una ecuación corresponde a las reparaciones mecánicas y la otra a las reparaciones eléctricas. Además como $b_{y2.1} = 1{,}26$ se sabe que, en promedio, en las reparaciones eléctricas se necesitan 1,26 horas más que en las reparaciones mecánicas (pp. 652-653).

7.2.16.2. Variables cualitativas más complejas

Conforme a Anderson y otros (2009) en el caso en que la variable cualitativa tenga 2 niveles será fácil definirla empleando cero para indicar un nivel y uno para indicar el otro nivel. No obstante, Anderson y otros (2009) agregan

> (…) Cuando una variable cualitativa tiene más de dos niveles habrá que tener cuidado tanto al definir como al interpretar estas variables ficticias. (…) Si una variable ficticia tiene *k* niveles se necesitan *k* - 1 variables ficticias, cada una de las cuales tomará el valor 0 ó 1 (p. 653).

Por ejemplo, Barón López, F. J., y Téllez Montiel, F. (s. f.) plantean el siguiente escenario. Considérese el caso en el que se haya identificado a las variables Altura, Sexo (hombre o mujer) y Dieta (normal, alta en proteínas o vegetariana) como predictoras de la variable Peso. La equivalencia de cada variable sería:

> Peso = Peso corporal de una persona (en kg.)
>
> Altura = Estatura (en Cm).
>
> Sexo = 0 (si es hombre), y 1 (si es mujer).

En tanto, los niveles de la variable Dieta se expresarían en función de las variables

predictoras IndProteina e IndVegetal como muestra la tabla 22.

Tabla 22. Codificación variable Dieta[310]

Dieta	**IndProteina (x_3)**	**IndVegetal (x_4)**
Normal	0	0
Alta en proteínas	1	0
Vegetariana	0	1

Así, de acuerdo a Barón López y Téllez Montiel (s. f.) se halló la siguiente ecuación de regresión estimada:

$$Peso = -100 + 1.\text{Altura} - 5.\text{Sexo} + 4.ind\text{Proteína} - 6.ind\text{Vegetal}$$

En dicha ecuación se observa que por cada centímetro de estatura se aumenta un kilogramo de peso cuando se controlan las demás variables. Igualmente se aprecia que las mujeres, en promedio, pesan 5 kg. menos que los hombres con el resto de las variables controladas. Asimismo se destaca que las personas que consumen una dieta alta en proteínas pesan, en promedio, 4 kg. más que las personas que tienen una dieta normal. Aquellos que llevan una dieta vegetariana pesan 6 kg. menos que los que acostumbran una dieta normal. Mientras quienes guardan una dieta alta en proteínas pesan, en promedio, 10 kg. más que quienes hacen una dieta vegetariana. Esto se puede calcular como sigue. Por un lado el peso estimado en función de una dieta alta en proteínas corresponde a – 96 kg, es decir, $Peso = -100 + 4(1) = -96$. Por otro lado el peso estimado en base a una dieta vegetariana es de – 106 kg, o sea, $Peso = -100 - 6(1) = -106$ kg. Quienes realizan la dieta alta en proteínas logran, en promedio, reducir 96 kg. y quienes respetan la dieta vegetariana, en promedio, alcanzan a

[310] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

disminuir 106 kg. Por ende las personas del grupo vegetariano pesan 10 kg. menos que las del grupo alto en proteínas.

7.3. Ejemplos de análisis de regresión en SPSS

Hoy día uno de los paquetes estadísticos más utilizados es SPSS. Debido a su importancia actual en las siguientes líneas se presenta varios ejemplos de análisis de regresión en función de diversos escenarios en los que se puede utilizar esta herramienta. Debido a que se trata de ejemplos los casos presentados no necesariamente cumplan con cada uno de los criterios que se comentan en los apartados *Análisis de regresión lineal simple*[311] y *Análisis de regresión múltiple*[312]. Se analiza estudios que no consideraron, por ejemplo, el número de variables independientes del modelo cuando se eligió el tamaño de la muestra. Eso evidencia que dichos estudios analizan los datos en base a una muestra inadecuada. No obstante, todos los ejemplos mostrados se utilizan porque permiten una amplia ejemplificación de la teoría correspondiente al análisis de regresión. Cabe destacar que algunos estaban escritos en inglés, razón por la que en ciertos casos se conserva las variables en dicho idioma.

7.3.1. Caso de regresión múltiple en SPSS con análisis de supuestos y eliminación de observaciones influyentes

Conforme a The University of Texas at Austin (s. f.) una empresa denominada HATCO generaba un producto el cual, en este estudio, se llamará *Producto H*[313]. De acuerdo a The University of Texas at Austin (s. f.) en dicha empresa se pensaba que

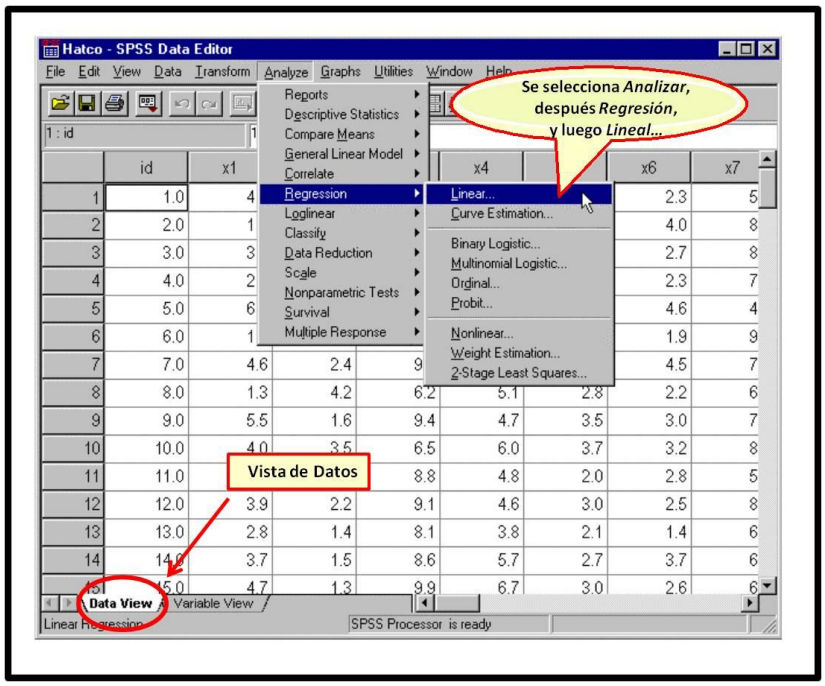
[311] La cursiva es nuestra.
[312] Ibidem.
[313] Ibidem.

el nivel de uso del producto por parte de sus consumidores (Usage level[314], en inglés) estaba vinculado a un conjunto de siete factores. Según The University of Texas at Austin (s. f.) dichos factores son: Velocidad de envío (Delivery speed), Nivel de precio (Price level), Flexibilidad de precio (Price flexibility), Imagen del manufacturador (Manufacturer's image), Servicio (Service), Imagen de fuerza de ventas (Sales force image), y Calidad del producto (Product quality). Dado a que conforme a The University of Texas at Austin (s. f.) HATCO quería determinar los factores que llevaban al incremento del nivel de uso del citado producto se solicitó un análisis de regresión múltiple en base a las referidas variables. El procedimiento empleado por los autores de la investigación para efectuar dicho análisis inicia teniendo en cuenta que según The University of Texas at Austin (s. f.) la variable dependiente es Usage level (x9). Mientras, conforme a The University of Texas at Austin (s. f.) las variables independientes son Delivery speed (x1), Price level (x2), Price flexibility (x3), Manufacturer's image (x4), Service (x5), Sales force image (x6), and Product quality (x7).

De acuerdo a The University of Texas at Austin (s. f.) se abre el archivo que contiene la base de datos sobre este estudio, se solicita el análisis de regresión desde la ventana *Vista de Datos*[315] a través de la secuencia: *Analizar*[316] → *Regresión*[317] → *Lineal…*[318]. La figura 40 muestra la aplicación de la referida secuencia.

[314] En un primer momento The University of Texas at Austin (s. f.) identifica la variable dependiente como Product usage pero luego en el desarrollo del caso ejemplificado se refiere a dicha variable como Usage level. Por ende se asume esta última denominación como única para nombrar la variable dependiente.
[315] La cursiva es nuestra.
[316] Ibidem.
[317] Ibidem.
[318] Ibidem.

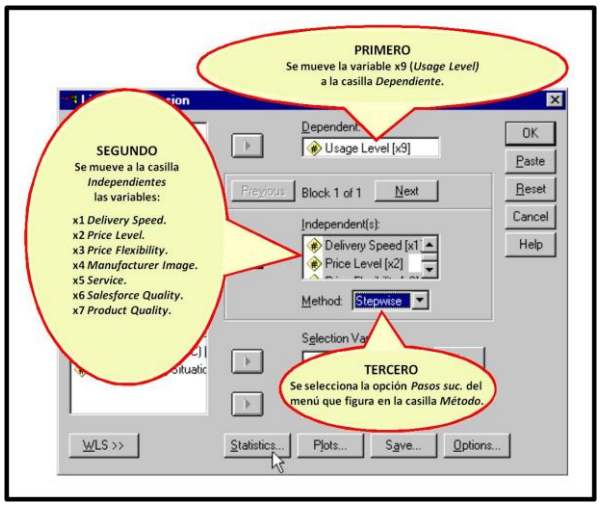
Figura 40. Aplicación de la secuencia para la solicitud del análisis de regresión en el caso HATCO[319]



A continuación conforme a The University of Texas at Austin (s. f.) se abre el cuadro de diálogo *Regresión lineal*[320]. Según The University of Texas at Austin (s. f.) del recuadro de la izquierda se mueve las variables a las casillas *Dependiente*[321] e *Independiente* (*s*)[322] según el rol que cumplan en el análisis. Igualmente de acuerdo a The University of Texas at Austin (s. f.), ya que se trata de un análisis exploratorio e interesa identificar el mejor conjunto de predictores, se elige el método de selección de variables *Por pasos*[323] el cual en la casilla *Método*[324] figura como *Pasos suc.*[325]. La figura 41 muestra dichas acciones.

[319] Adaptado de Request the regression analysis, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[320] La cursiva es nuestra.
[321] Ibidem.
[322] Ibidem.
[323] Ibidem.
[324] Ibidem.
[325] Ibidem.

Figura 41. Indicación del método y variables dependiente e independientes en el caso HATCO[326]

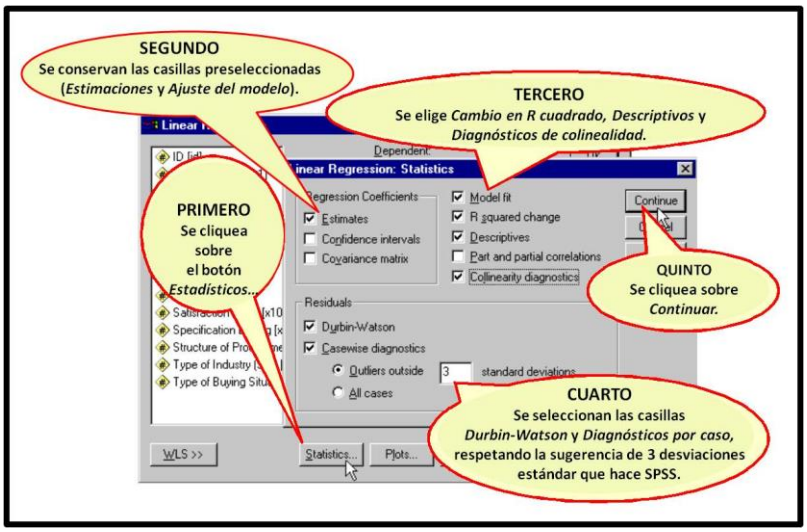


De acuerdo a The University of Texas at Austin (s. f.) en el cuadro de diálogo *Regresión lineal*[327] se cliquea sobre el botón *Estadísticos…*[328] para abrir el cuadro de diálogo *Regresión Lineal: Estadísticos*[329]. En el mismo aparece preseleccionada la opción *Estimaciones*[330], la cual se conserva seleccionada. En este caso según The University of Texas at Austin (s. f.) se ha seleccionado adicionalmente las alternativas *Ajuste del modelo*[331], *Cambio en R cuadrado*[332], *Descriptivos*[333], *Diagnósticos de colinealidad*[334], *Durbin-Watson*[335] y *Diagnósticos por caso*[336]. Según The University of Texas at Austin (s. f.) tras seleccionar esa última opción se genera otra llamada *Atípicos fuera de … Desviaciones estándar*[337]. En la misma conforme a The University of Texas at Austin (s. f.) se respeta la predeterminación de 3 desviaciones estándar que hace SPSS. De acuerdo a The University of Texas at

[326] Adaptado de Specify the dependent and independent variables and the variable selection method, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[327] La cursiva es nuestra.
[328] Ibidem.
[329] Ibidem.
[330] Ibidem.
[331] Ibidem.
[332] Ibidem.
[333] Ibidem.
[334] Ibidem.
[335] Ibidem.
[336] Ibidem.
[337] Ibidem.

Austin (s. f.) luego se cliquea sobre el botón *Continuar*[338] para cerrar el cuadro de diálogo *Regresión lineal: Estadísticos*[339]. La figura 42 muestra dicha selección de opciones.

Figura 42. Selecciones en el cuadro de diálogo *Regresión Lineal: Estadísticos*[340] del caso HATCO[341]

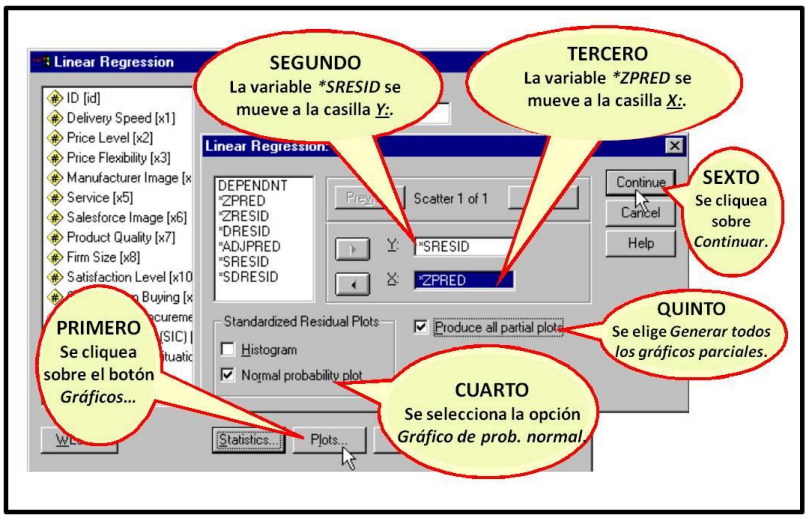


Asimismo según The University of Texas at Austin (s. f.) en el cuadro de diálogo *Regresión lineal*[342] se cliquea sobre el botón *Gráficos…*[343] para abrir el cuadro de diálogo *Regresión lineal: Gráficos*[344]. Conforme a The University of Texas at Austin (s. f.) del recuadro de la izquierda se mueve a la casilla *Y:* la variable que representará el eje *y*. De igual manera de acuerdo a The University of Texas at Austin (s. f.) se mueve la variable que constituirá el eje *x* a la casilla *X:*. En este caso conforme a The University of Texas at Austin (s. f.) **SRESID*[345] va a la casilla *Y:* y **ZPRED*[346] a la casilla *X:*. Igualmente según The University of Texas at Austin (s. f.) se selecciona

[338] La cursiva es nuestra.
[339] Ibidem.
[340] Ibidem.
[341] Adaptado de Specify the statistics options, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[342] La cursiva es nuestra.
[343] Ibidem.
[344] Ibidem.
[345] Ibidem.
[346] Ibidem.

las opciones *Gráfico de prob. normal*[347] y *Generar todos los gráficos parciales*[348]. Luego de acuerdo a The University of Texas at Austin (s. f.) se cliquea sobre el botón *Continuar*[349] para cerrar el cuadro de diálogo *Regresión lineal: Gráficos*[350]. La figura 43 muestra las acciones comentadas.

Figura 43. Elecciones varias en cuadro de diálogo Regresión Lineal: Gráficos del caso HATCO[351]

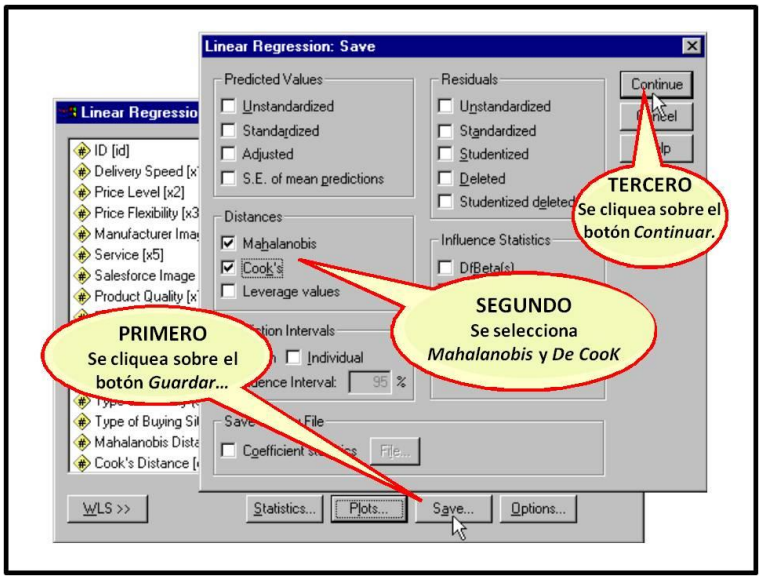


De igual manera conforme a The University of Texas at Austin (s. f.) en el cuadro de diálogo *Regresión lineal*[352] se cliquea sobre el botón *Guardar…*[353] para abrir el cuadro de diálogo *Regresión lineal: Guardar*[354]. En este caso de acuerdo a The University of Texas at Austin (s. f.) se selecciona las opciones *Mahalanobis*[355] y *Distancia de Cook*[356]. Luego según The University of Texas at Austin (s. f.) se

[347] La cursiva es nuestra.
[348] Ibidem.
[349] Ibidem.
[350] Ibidem.
[351] Adaptado de Specify the plots to include in the output, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[352] La cursiva es nuestra.
[353] Ibidem.
[354] Ibidem.
[355] Ibidem.
[356] Ibidem.

cliquea sobre el botón *Continuar*[357] para cerrar el cuadro de diálogo *Regresión lineal: Guardar*[358]. La figura 44 muestra estas ejecuciones.

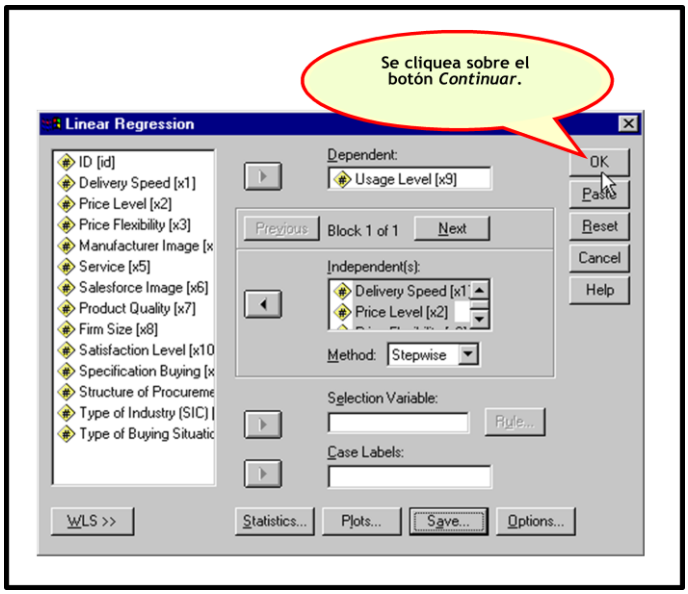
Figura 44. Selecciones en el cuadro de diálogo *Regresión Lineal: Guardar*[359] del caso HATCO[360]



Nuevamente en el cuadro de diálogo *Regresión lineal*[361] de acuerdo a The University of Texas at Austin (s. f.) se cliquea sobre el botón *Continuar*[362] para ejecutar las solicitudes que se han indicado. La figura 45 muestra esta acción.

[357] La cursiva es nuestra.
[358] Ibidem.
[359] Ibidem.
[360] Adaptado de Specify diagnostic statistics to save to the data set, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[361] La cursiva es nuestra.
[362] Ibidem.

Figura 45. Ejecución de solicitudes en el cuadro de diálogo Regresión Lineal del caso HATCO[363]

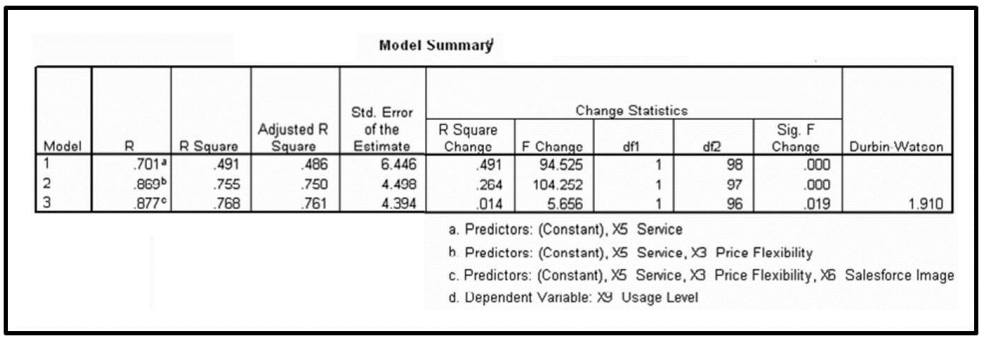


Tras requerir los resultados de las solicitudes especificadas se obtiene la tabla *Resumen del modelo*[364] mostrada en la tabla 23 y la tabla ANOVA presentada en la tabla 24. En el marco de este caso de estudio, se observa que el mejor modelo resultó ser el número tres (3). Dicho modelo conforme a The University of Texas at Austin (s. f.) incluye las variables Service (x5), Price flexibility (x3) e Image Sales Force (x6). Como se aprecia en la tabla 23, según The University of Texas at Austin (s. f.) el mismo explica aproximadamente un 76% de la varianza de la variable dependiente x9 o Usage level (*Adj.* R^2 = 0,761). Igualmente, como se distingue en la tabla 24, el modelo en cuestión (modelo 3) es significativo a un nivel de confianza de 95% (Sig. = 0,000).

[363] Adaptado de Complete the regression analysis request, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

[364] La cursiva es nuestra.

Tabla 23. Tabla Resumen del modelo según las especificaciones solicitadas en el caso HATCO[365]

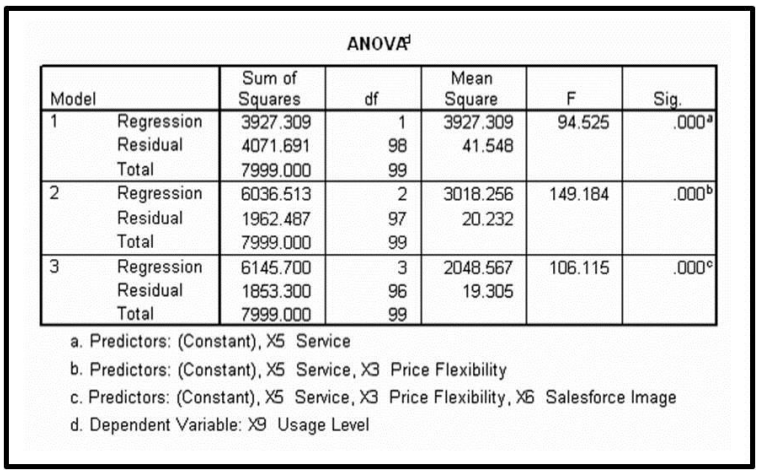


Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.701[a]	.491	.486	6.446	.491	94.525	1	98	.000	
2	.869[b]	.755	.750	4.498	.264	104.252	1	97	.000	
3	.877[c]	.768	.761	4.394	.014	5.656	1	96	.019	1.910

a. Predictors: (Constant), X5 Service
b. Predictors: (Constant), X5 Service, X3 Price Flexibility
c. Predictors: (Constant), X5 Service, X3 Price Flexibility, X6 Salesforce Image
d. Dependent Variable: X9 Usage Level

Tabla 24. Tabla ANOVA del caso HATCO[366]



ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3927.309	1	3927.309	94.525	.000[a]
	Residual	4071.691	98	41.548		
	Total	7999.000	99			
2	Regression	6036.513	2	3018.256	149.184	.000[b]
	Residual	1962.487	97	20.232		
	Total	7999.000	99			
3	Regression	6145.700	3	2048.567	106.115	.000[c]
	Residual	1853.300	96	19.305		
	Total	7999.000	99			

a. Predictors: (Constant), X5 Service
b. Predictors: (Constant), X5 Service, X3 Price Flexibility
c. Predictors: (Constant), X5 Service, X3 Price Flexibility, X6 Salesforce Image
d. Dependent Variable: X9 Usage Level

En la tabla de coeficientes de la tabla 25 conforme a The University of Texas at Austin (s. f.) se observa que el Uso del producto (x9) aumenta aproximadamente 7,62 unidades por cada unidad aumentada en el Servicio (x5). De igual modo The University of Texas at Austin (s. f.) señala que el Uso del producto (x9) aumenta aproximadamente 3,38 unidades por cada unidad aumentada en la Flexibilidad de precio (x3). Asimismo The University of Texas at Austin (s. f.) indica que el Uso del producto (x9) aumenta aproximadamente 1,41 unidades por cada unidad aumentada en la Imagen de fuerza de ventas (x6).

[365] Adaptado de Significance test of the coefficient of determination R Square, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

[366] Adaptado de Significance test of the coefficient of determination R Square, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

Tabla 25. Tabla Coeficientes del caso HATCO[367]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	21,653	2,596		8,341	,000		
	Servicio conjunto	8,384	,862	,701	9,722	,000	1,000	1,000
2	(Constante)	-3,489	3,057		-1,141	,257		
	Servicio conjunto	7,974	,603	,666	13,221	,000	,996	1,004
	Flexibilidad de precios	3,336	,327	,515	10,210	,000	,996	1,004
3	(Constante)	-6,520	3,247		-2,008	,047		
	Servicio conjunto	7,621	,607	,637	12,547	,000	,936	1,068
	Flexibilidad de precios	3,376	,320	,521	10,562	,000	,993	1,007
	Imagen de fuerza de ventas	1,406	,591	,121	2,378	,019	,939	1,064

Conforme a The University of Texas at Austin (s. f.) en la columna *Coeficientes estandarizados (beta)*[368] de la tabla 25 se nota que la variable más importante en el modelo es Service (x5), con un coeficiente estandarizado de aproximadamente 0,64. Según The University of Texas at Austin (s. f.) le sigue la variable Price flexibility (x3), con un coeficiente estandarizado de aproximadamente 0,52. Finalmente de acuerdo a The University of Texas at Austin (s. f.) aparece la variable Salesforce image (x6), con un coeficiente estandarizado de aproximadamente 0,12.

En la columna *Tolerancia*[369] de la tabla 25 los valores correspondientes al modelo 3 son mayores que 0,10 y en la columna *FIV*[370] los mismos no son mayores que 10. Conforme a The University of Texas at Austin (s. f.) eso indica que dicho modelo está libre de problemas de multicolinealidad.

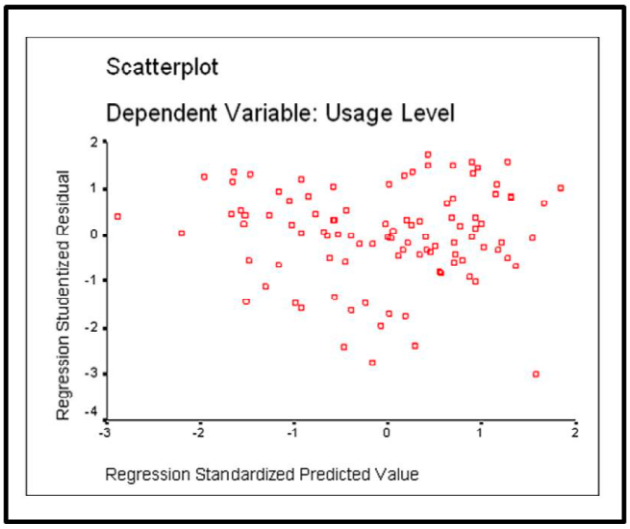
[367] Adaptado de Significance test of individual regression coefficients, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[368] La cursiva es nuestra.
[369] Ibidem.
[370] Ibidem.

De esa manera según The University of Texas at Austin (s. f.) la ecuación de regresión resultante es:

$$\hat{y} = -6{,}520 + 7{,}621\ x5 + 3{,}376\ x3 + 1{,}406\ x6$$

Luego de haber seleccionado el modelo de regresión se evalúa los supuestos del mismo. Conforme a The University of Texas at Austin (s. f.) la figura 46 muestra un gráfico residual en el que el eje *y* está representado por los residuos estudentizados y el eje *x* por los valores predichos estandarizados. De acuerdo a The University of Texas at Austin (s. f.) la nube de puntos que se presenta en el mismo indica que hay una relación lineal entre las variables dependiente e independientes, ya que no hay presencia de un patrón definido. Igualmente The University of Texas at Austin (s. f.) establece que hay homocedasticidad, pues los puntos se distribuyen a lo largo de una banda horizontal.

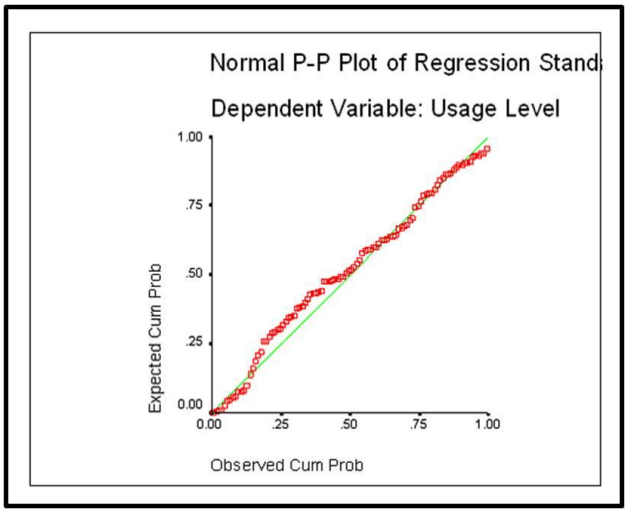
Figura 46. Gráfico residuos estudentizados y valores predichos estandarizados en caso HATCO[371]



[371] Adaptado de Linearity and constant variance for the dependent variable - residual plot, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

De acuerdo a The University of Texas at Austin (s. f.) para evaluar el supuesto de normalidad se genera un gráfico de probabilidad normal como se muestra en la figura 47. Conforme a The University of Texas at Austin (s. f.) en este caso se observa que los puntos se ajustan satisfactoriamente a la línea, por lo que se aprueba el supuesto de normalidad.

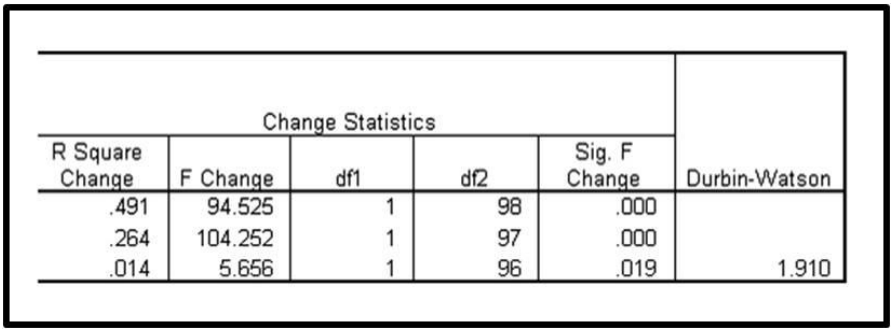
Figura 47. Gráfico de probabilidad normal del caso HATCO[372]



Asimismo de acuerdo a The University of Texas at Austin (s. f.) en la figura 48 se muestra el valor del estadístico Durbin-Watson. Dicho estadístico se encuentra en la tabla Resumen del modelo, pero se suele analizar una vez se analizan los demás supuestos. En lo referente al caso en cuestión según The University of Texas at Austin (s. f.) la columna Durbin-Watson presenta un valor de 1,910. Ya que el mismo es aproximadamente igual a 2 The University of Texas at Austin (s. f.) concluye que no existe autocorrelación o correlación serial. En ese sentido se aprecia que 1,910 supera el d_U = 1,74 que corresponde a la combinación K = 3 y n = 100 lo cual también conduce a señalar la ausencia de autocorrelación.

[372] Adaptado de Normal distribution of residuals - normality plot of residuals, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

Figura 48. Valor del estadístico Durbin-Watson[373]



Change Statistics					
R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
.491	94.525	1	98	.000	
.264	104.252	1	97	.000	
.014	5.656	1	96	.019	1.910

En tanto, según The University of Texas at Austin (s. f.) la tabla 26 muestra las estadísticas respecto a los residuos. Se observa que no todos los residuos oscilan entre +/- 2 desviaciones estándar. Esto así debido a que el valor máximo de los residuos estandarizados es igual a 1,724 y el valor mínimo es igual a -2,857. De ese modo se advierte que es posible que haya uno o varios casos atípicos en el modelo.

Tabla 26. Tabla Estadísticas respecto a los residuales[374]

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	23.373	60.592	46.100	7.879	100
Std. Predicted Value	-2.885	1.839	.000	1.000	100
Standard Error of Predicted Value	.467	1.429	.847	.236	100
Adjusted Predicted Value	23.180	60.388	46.104	7.915	100
Residual	-12.552	7.574	3.055E-15	4.327	100
Std. Residual	-2.857	1.724	.000	.985	100
Stud. Residual	-2.983	1.737	.000	1.004	100
Deleted Residual	-13.687	7.694	-4.29E-03	4.497	100
Stud. Deleted Residual	-3.115	1.756	-.004	1.017	100
Mahal. Distance	.129	9.485	2.970	2.185	100
Cook's Distance	.000	.201	.010	.022	100
Centered Leverage Value	.001	.096	.030	.022	100

a. Dependent Variable: Usage Level

Según The University of Texas at Austin (s. f.) SPSS guarda el valor de la distancia de Mahalanobis para cada caso, por lo que se debe solicitar la probabilidad para

[373] Adaptado de Independence of residuals - Durbin-Watson statistic, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

[374] Adaptado de Identifying dependent variable outliers - casewise plot of standardized residuals, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

identificar valores atípicos. A esos fines conforme The University of Texas at Austin (s. f.) se siguen los pasos indicados en las figuras 49 y 50.

Figura 49. Transformación de la variable p_mahal en el caso HATCO[375]

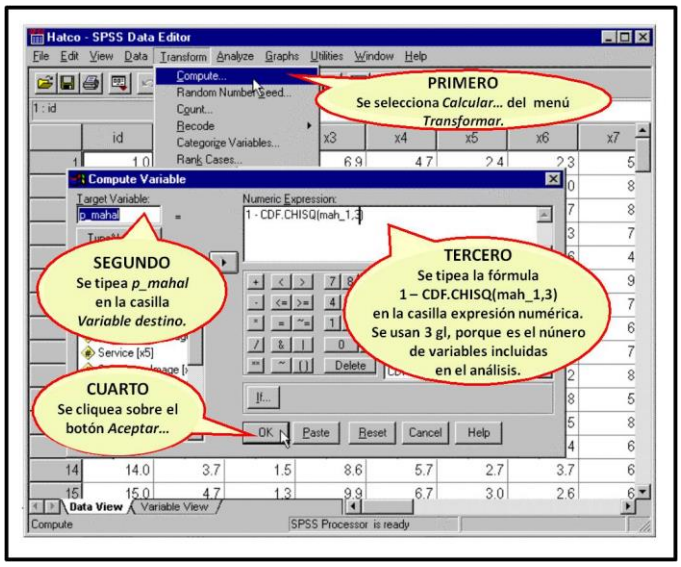
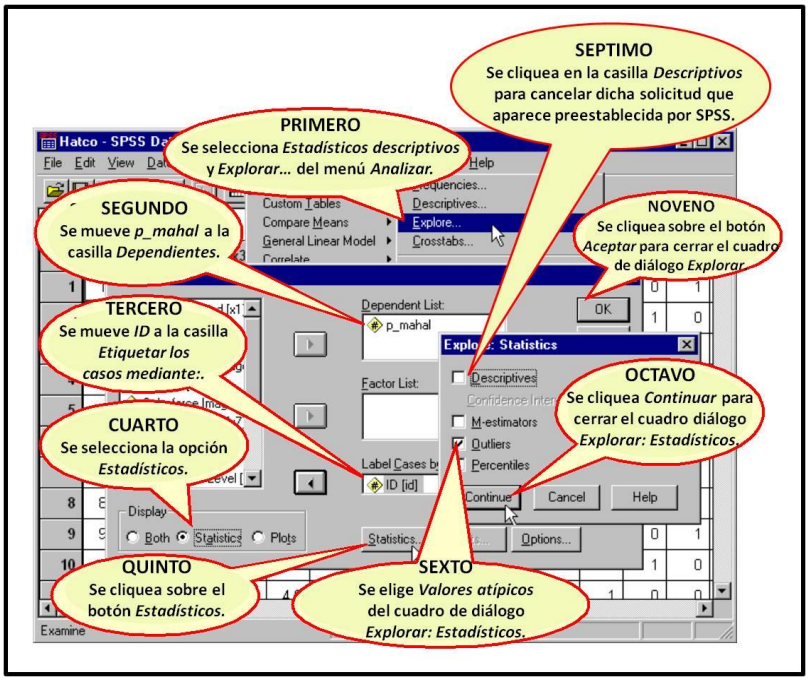


Figura 50. Identificación de valores atípicos vía la distancia de Mahalanobis en el caso HATCO[376]



[375] Adaptado de Computing probabilities for Mahalanobis distance, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

[376] Adaptado de Identifying Statistically Significant Mahalanobis Distance Scores, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

De acuerdo a The University of Texas at Austin (s. f.) los casos con una probabilidad de distancia de Mahalanobis menor que 0,05 se muestran en la mitad de la tabla 27 en la lista denominada *Más bajo*[377] (lowest, en inglés). Según The University of Texas at Austin (s. f.) en el ejemplo se observa que los casos 96, 82, 42, 5 y 93 constituyen valores atípicos conforme a este criterio.

Tabla 27. Valores Atípicos según probabilidad distancia de Mahalanobis en caso HATCO[378]

Extreme Values

			Case Number	ID	Value
P_MAHAL	Highest	1	84	84	.988
		2	59	59	.978
		3	41	41	.973
		4	87	87	.969
		5	85	85	.963
	Lowest	1	96	96	.023
		2	82	82	.031
		3	42	42	.045
		4	5	5.0	.045
		5	93	93	.064

El estado del arte no reveló ningún análisis que excluya dichos valores. Por ende se asume que al realizar un análisis de regresión sin dichos valores se obtuvieron resultados poco diferentes. Eso conduciría a conservar las observaciones atípicas en el análisis debido a la poca relevancia que tiene excluirlas del mismo.

De acuerdo a The University of Texas at Austin (s. f.) en adición al requerimiento del puntaje de la distancia de Mahalanobis se solicita también los puntajes de la distancia de Cook. Según The University of Texas at Austin (s. f.) la idea es identificar casos que tengan un puntaje mayor que el criterio computado usando la fórmula $4/(n - k - 1)$, siendo n el número de casos en el análisis y k el número de variables independientes. Conforme a The University of Texas at Austin (s. f.) en el marco de

[377] La cursiva es nuestra.
[378] Adaptado de Identifying potential outliers, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

este ejemplo n = 100 y k = 3, por lo que al aplicar la fórmula se obtiene: 4 / (100 - 3 - 1) = 0,042. De acuerdo a The University of Texas at Austin (s. f.) luego, como ilustra las figuras 51 y 52, se ordena la base de datos en función de la variable *coo_1* para buscar casos influyentes que tengan una distancia de Cook mayor a 0,042.

Figura 51. Ordenamiento base de datos conforme distancia de Cook en caso HATCO[379]

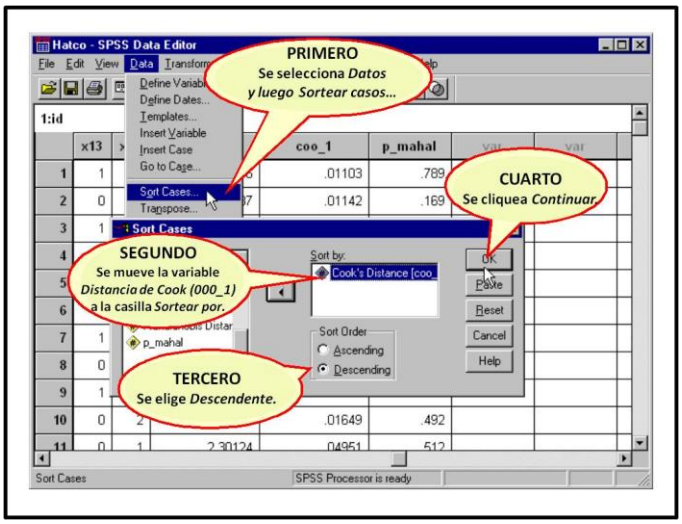
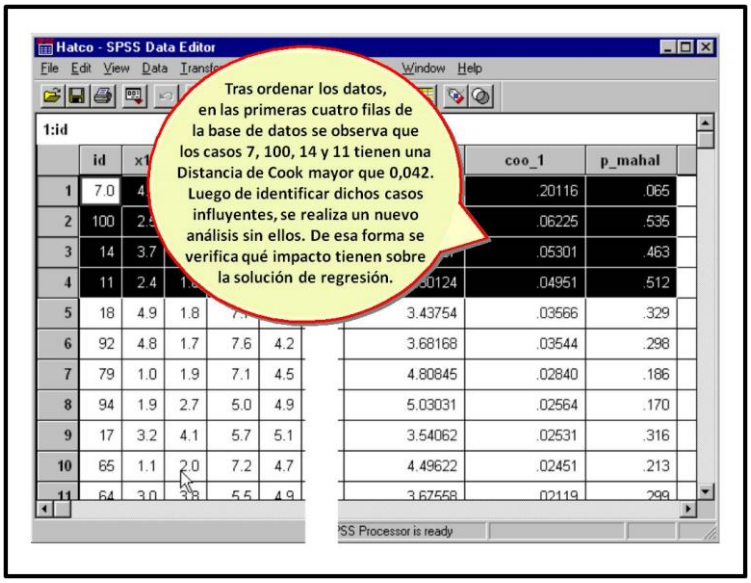


Figura 52. Casos con distancia de Cook mayor que 0,042 en caso HATCO[380]



[379] Adaptado de Sorting Cook's distance scores in descending order, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

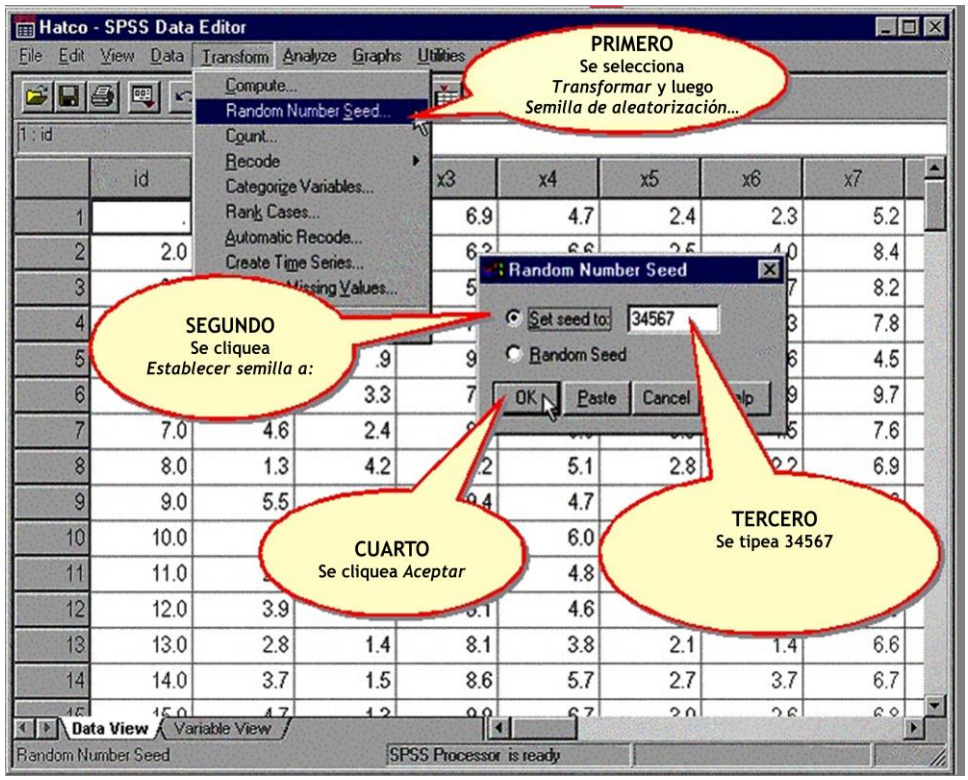
[380] Adaptado de cases with large Cook's distances, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

Según The University of Texas at Austin (s. f.) se realiza un nuevo análisis de regresión sin las observaciones influyentes encontradas en la figura 52. Si se constata que tras esa medida el modelo presenta una mejor explicación de la varianza del Uso del producto (x9) se adopta los nuevos resultados como los definitivos.

Si en este punto del análisis se decide aceptar el modelo obtenido entonces corresponde verificar si los resultados son generalizables a la población o no. De acuerdo a The University of Texas at Austin (s. f.) un indicativo de que se puede generalizar los resultados más allá de la muestra es el valor de R^2 ajustado. Conforme a The University of Texas at Austin (s. f.) si dicho valor es mucho menor que el valor de R^2 se considera que los resultados no son generalizables. The University of Texas at Austin (s. f.) señala que en este caso R^2 = 0,768 y R^2 ajustado = 0,761. Según The University of Texas at Austin (s. f.) estos valores están muy cerca uno del otro, por lo que se entiende que los resultados se pueden generalizar a la población. Igualmente de acuerdo a The University of Texas at Austin (s. f.) se puede recurrir a una estrategia más elaborada para validar el análisis de regresión la cual consiste en dividir aleatoriamente la muestra en dos grupos. En ese sentido The University of Texas at Austin (s. f.) señala que se selecciona del menú *Transformar*[381] la opción *Semilla de aleatorización…*[382]. Según The University of Texas at Austin (s. f.) después se cliquea *Establecer semilla a:* [383], se tipea 34567 el cual es el valor de la semilla aleatoria especificada por los autores del estudio, y se selecciona el botón *Aceptar*[384]. Las referidas acciones se presentan en la figura 53.

[381] La cursiva es nuestra.
[382] Ibidem.
[383] Ibidem.
[384] Ibidem.

Figura 53. Pasos para establecer la semilla aleatoria[385]

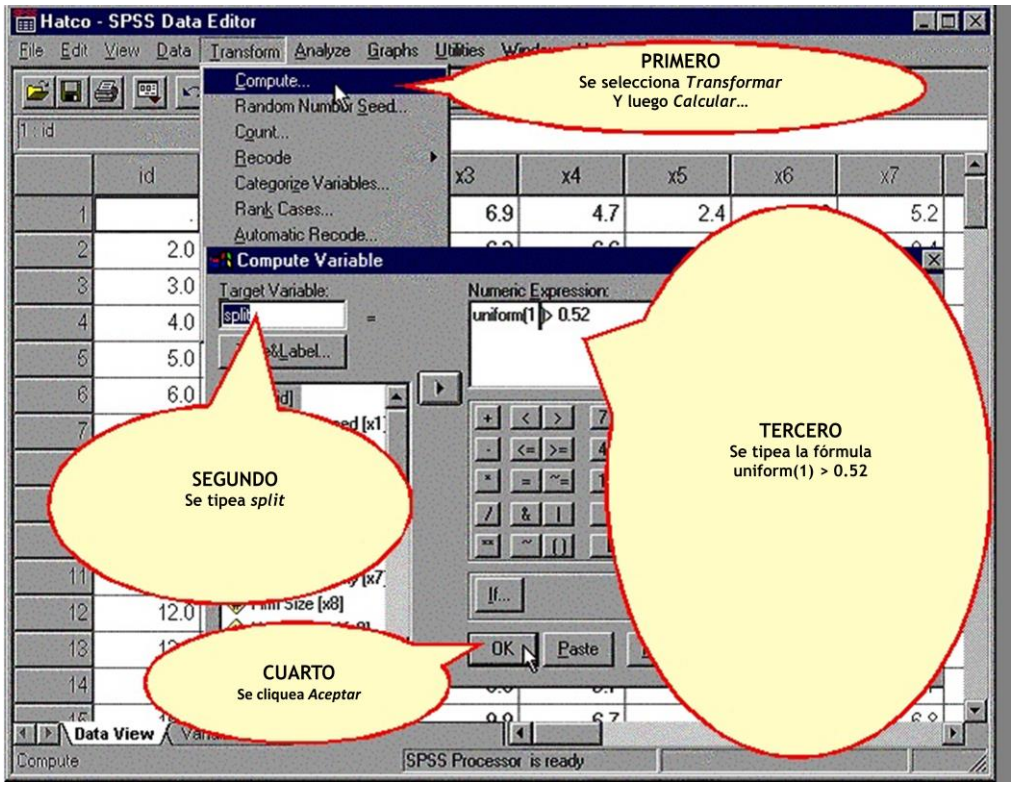


De acuerdo a The University of Texas at Austin (s. f.) luego se vuelve a elegir el menú *Transformar*[386] y se selecciona la opción *Calcular*[387]. Conforme a The University of Texas at Austin (s. f.) se crea una nueva variable tipeando *split*[388] en el cuadro de texto denominado *Variable destino*[389] e ingresando la fórmula uniform(1) > 0.52 en el cuadro de texto *Expresión numérica*[390]. Conforme a The University of Texas at Austin (s. f.) la función *uniform*[391] genera un número aleatorio entre 0 y 1 para cada caso. Así, según The University of Texas at Austin (s. f.) si el número aleatorio generado es mayor que 0,52 la expresión numérica será 1 pues se cumple el criterio especificado en la citada función. En cambio, de acuerdo a The University of

[385] Adaptado de Set the starting point for random number generation, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[386] La cursiva es nuestra.
[387] Ibidem.
[388] Ibidem.
[389] Ibidem.
[390] Ibidem.
[391] Ibidem.

Texas at Austin (s. f.) si el número aleatorio generado es menor o igual que 0,52 la referida expresión será 0 ya que se incumple el criterio establecido en la función *uniform*[392]. La mencionada secuencia se observa en la figura 54.

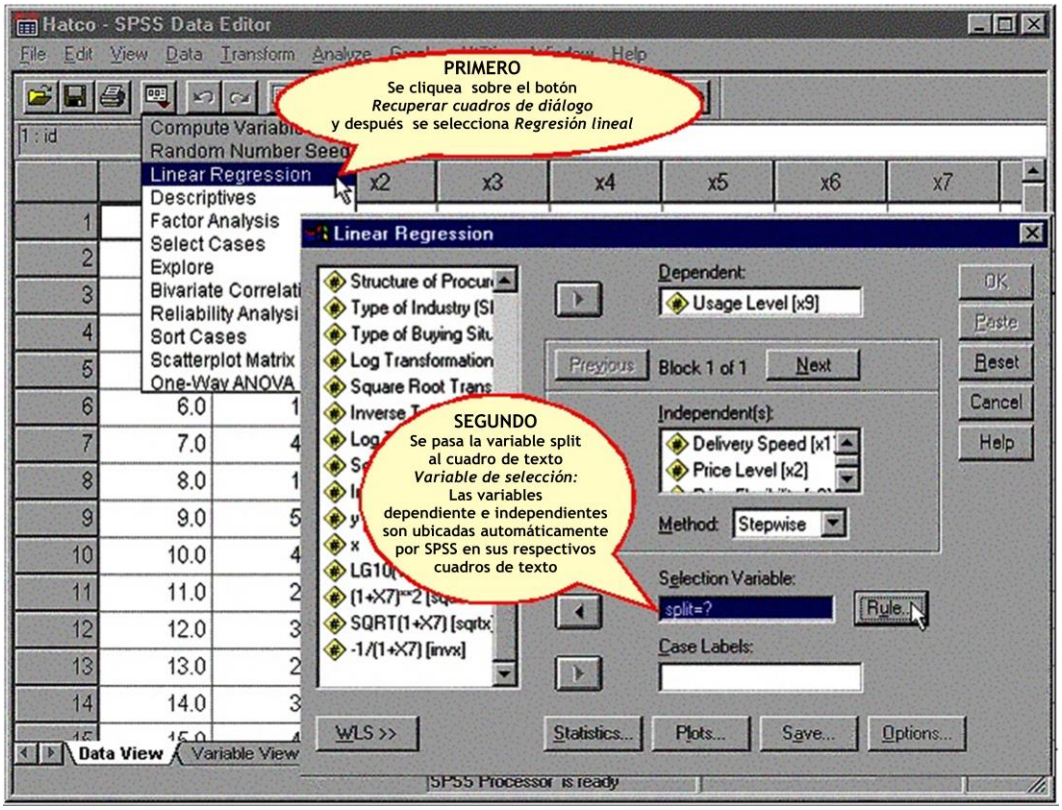
Figura 54. Empleo de la fórmula Uniform[393]



Según The University of Texas at Austin (s. f.) después se elige el botón *Recuperar cuadros de diálogo*[394] y se selecciona *Regresión lineal*[395]. Posteriormente The University of Texas at Austin (s. f.) señala que se pasa la variable split al cuadro de texto *Variable de selección:* [396]. Estas acciones se muestran en la figura 55.

[392] La cursiva es nuestra.
[393] Adaptado de Compute the variable to randomly split the sample into two halves, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[394] La cursiva es nuestra.
[395] Ibidem.
[396] Ibidem.

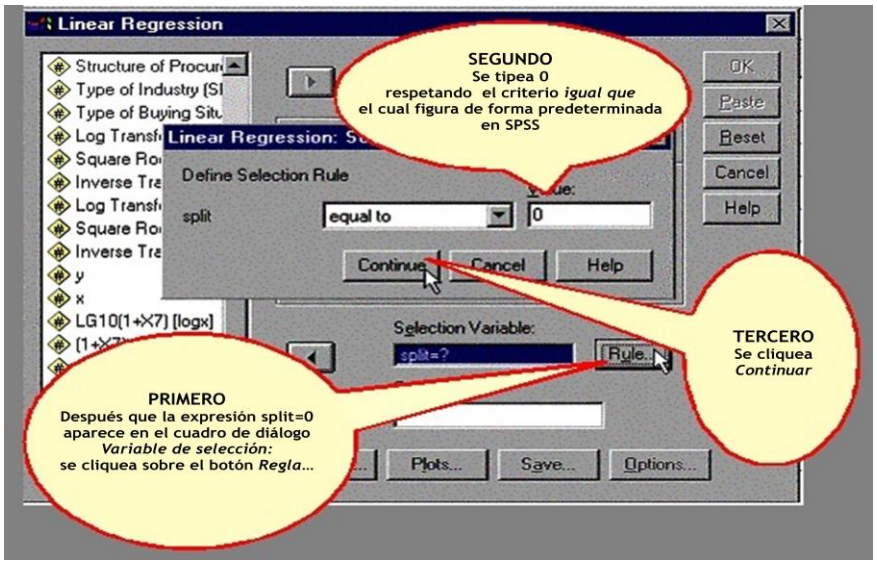
Figura 55. Adición de la variable split al cuadro de texto Variable de selección[397]



Conforme a The University of Texas at Austin (s. f.) luego que la variable split figura en *Variable de selección*:[398] se cliquea el botón *Regla…*[399]. De acuerdo a The University of Texas at Austin (s. f.) se respeta el criterio *igual que*[400] el cual está predeterminado por SPSS, se tipea 0 como valor y se elige la opción *Continuar*[401]. Esta secuencia se presenta en la figura 56.

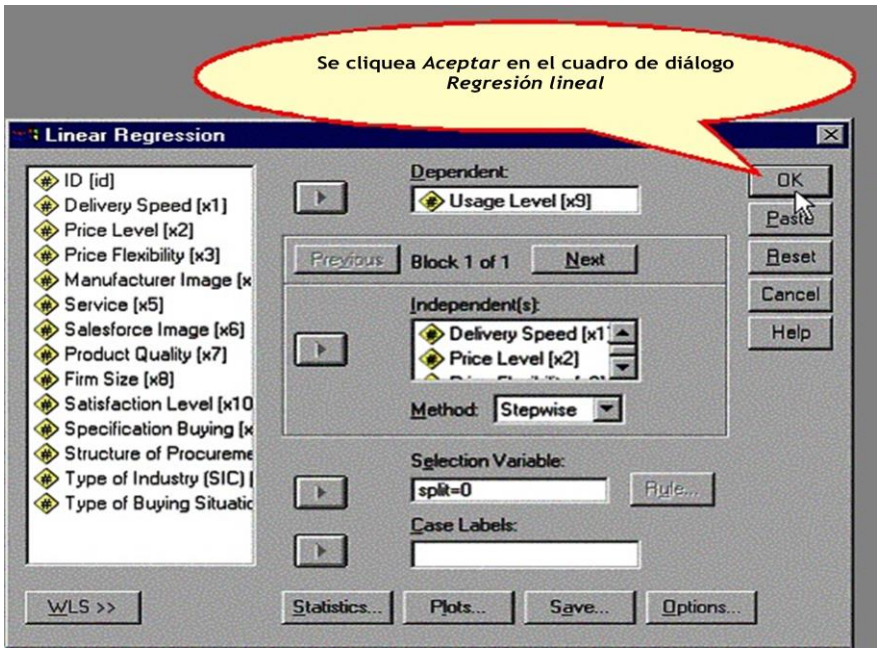
[397] Adaptado de Specify the cases to include in the first screening sample, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[398] La cursiva es nuestra.
[399] Ibidem.
[400] Ibidem.
[401] Ibidem.

Figura 56. Primera especificación del valor de la variable split[402]



Según The University of Texas at Austin (s. f.) después se cliquea el botón *Aceptar*[403] en el cuadro de diálogo *Regresión lineal*[404]. Esas acciones figuran en la figura 57.

Figura 57. Ejecución del análisis de regresión respecto al primer grupo[405]



[402] Adaptado de Define the case selection rule, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

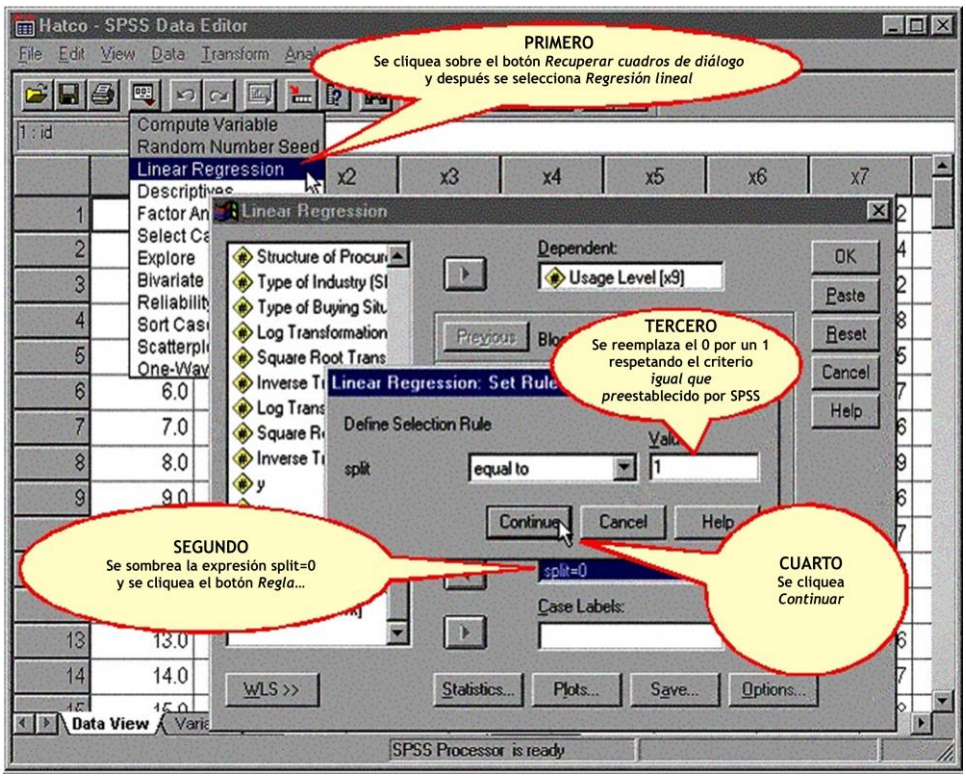
[403] La cursiva es nuestra.

[404] Ibidem.

[405] Adaptado de Complete the regression analysis request for the first screening sample, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

Conforme a The University of Texas at Austin (s. f.) posteriormente se elige el botón *Recuperar cuadros de diálogo*[406], se selecciona *Regresión lineal*[407], se sombrea la expresión split=0, se cliquea el botón *Regla…*[408], se reemplaza el valor 0 por 1, y se selecciona el botón *Continuar*[409]. Esta secuencia se presenta en la figura 58.

Figura 58. Segunda especificación del valor de la variable split[410]



Según The University of Texas at Austin (s. f.) luego se elige el botón *Aceptar*[411] en el cuadro de diálogo *Regresión lineal*[412]. Esta acción se muestra en la figura 59.

[406] La cursiva es nuestra.
[407] Ibidem.
[408] Ibidem.
[409] Ibidem.
[410] Adaptado de Specify the cases to include in the second screening sample, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).
[411] La cursiva es nuestra.
[412] Ibidem.

Figura 59. Ejecución del análisis de regresión respecto al segundo grupo [413]

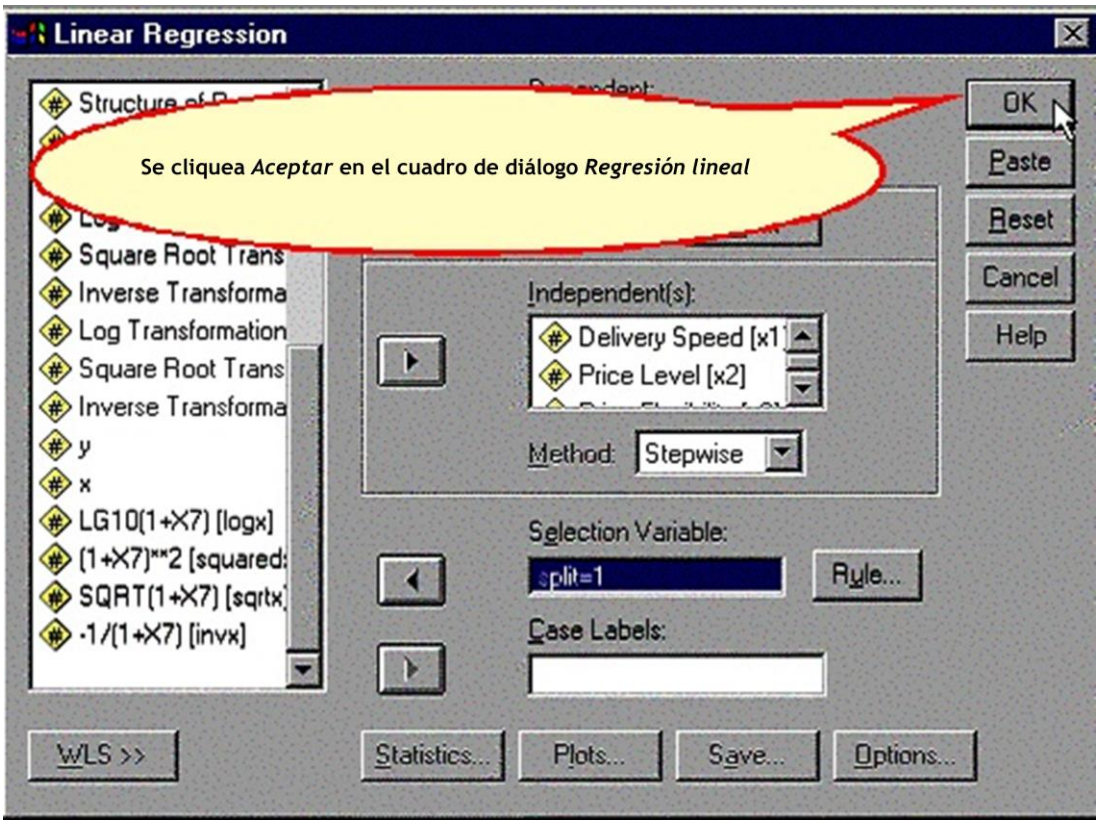


Tabla 28. Conjunto de resultados de los análisis de regresión efectuados[414]

	Modelo Completo	**Split = 0**	**Split = 1**
R	0,877	0,861	0,909
Coeficientes Significantes (p < 0,05)	Service Price flexibility Salesforce image	Service Price flexibility	Service Price flexibility Salesforce image
R^2	0,768	0,741	0,826
R^2 Ajustado	0,761	0,730	0,814

[413] Adaptado de Complete the regression analysis request for the second screening sample, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

[414] Adaptado de Summary table for validation analysis, en *Illustration of regression analysis*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html, por The University of Texas at Austin, (s. f.).

De acuerdo a The University of Texas at Austin (s. f.) se concluye que el modelo es generalizable pues, como se observa en la tabla 28, los *R* múltiple de ambas muestras son muy similares (0,861 y 0,909 en el caso de split=0 y split=1 respectivamente). Lo mismo se establece si se considera el denominado *Indice de reducción en la validación cruzada*[415]. Según Molinero, L. M. (2002) el mismo se obtiene a partir de la resta de los R^2 de las dos muestras. Conforme a Molinero (2002) si la diferencia entre los R^2 es menor que 0,1 el modelo se considera muy fiable pero si resulta mayor que 0,9 se asume como muy poco fiable. Dado a que el R^2 de la muestra split=1 es 0,83 y el R^2 de la muestra split=0 es 0,74 la diferencia de dichos coeficientes de determinación equivale a 0,09. Por ende se concluye que el modelo es generalizable ya que el *Indice de reducción en la validación cruzada*[416] es menor que 0,1. Sin embargo conforme a The University of Texas at Austin (s. f.) la variable Salesforce image no figura en la muestra split=0. Al respecto The University of Texas at Austin (s. f.) concluye que en este estudio hay una relación fuerte entre la variable dependiente Usage level y las variables independientes Service y Price flexibility. En ese sentido The University of Texas at Austin (s. f.) indica que probablemente Salesforce image mantenga cierta relación con la variable dependiente en algunos casos pero que la misma no es consistentemente evidente. Así, The University of Texas at Austin (s. f.) concluye que reportaría los resultados del modelo que incluye las variables Service y Price flexibility. De esa manera, The University of Texas at Austin (s. f.) señala que para obtener el correcto R^2 y otras estadísticas del modelo se debe generar un nuevo análisis que incluya solo el mencionado par de variables independientes.

[415] La cursiva es nuestra.
[416] Ibidem.

7.3.2. Caso de regresión múltiple en SPSS con análisis de componentes principales como solución de la multicolinealidad

Para ejemplificar la utilización del ACP en casos de regresión que presentan multicolinealidad grave a continuación se aborda un ejemplo presentado por The University of Texas at Austin (s. f.). Dicho caso se ha denominado en este estudio *Degree-Happiness*[417], haciendo alusión a los dos componentes principales que se genera como resultado final.

De acuerdo a The University of Texas at Austin (s. f.) el mencionado ejemplo representa un estudio en el que se tienen las siguientes variables independientes:

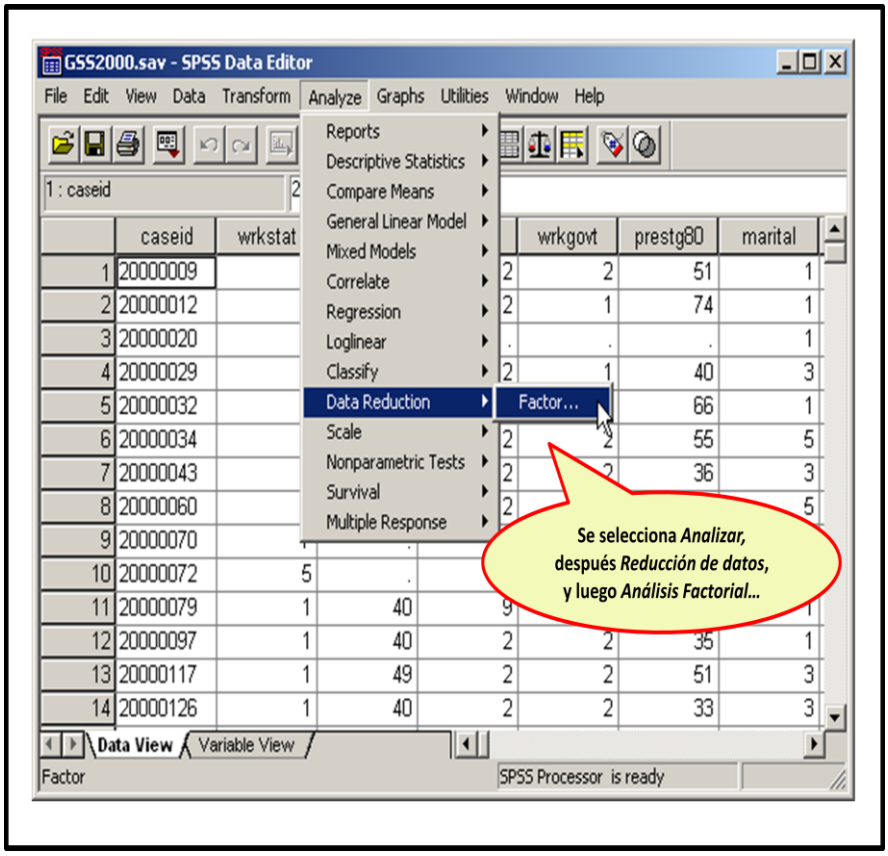
- Highest academic degree (degree).
- Father's highest academic degree (padeg).
- Mother's highest academic degree (madeg).
- General happiness (happy).
- Happiness of marriage (hapmar).
- Attitude toward life (life).
- Condition of health (health).
- Spouse's highest academic degree (spdeg).

Asumiendo que entre dichas variables dos o más están altamente correlacionadas, se puede proceder a generar un análisis de ACP para eliminar el problema de multicolinealidad. En ese sentido conforme a la University of Texas at Austin (s. f.)

[417] La cursiva es nuestra.

desde *Vista de Datos*[418] se selecciona *Analizar*[419], después *Reducción de datos*[420] y luego *Análisis factorial*[421] como se muestra en la figura 60.

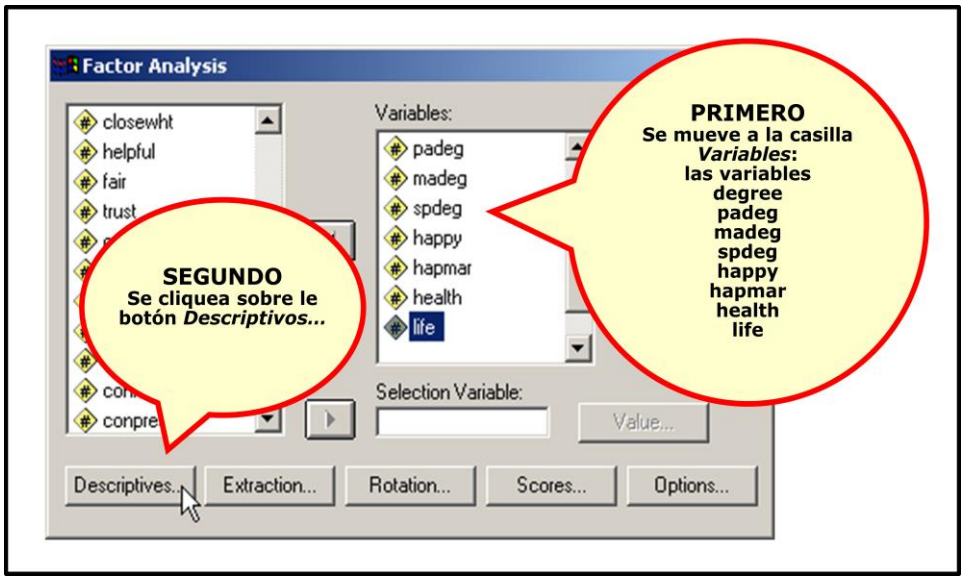
Figura 60. Solicitud de análisis factorial en el caso Degree-Happiness[422]



De acuerdo a la University of Texas at Austin (s. f.) en el cuadro de diálogo *Análisis factorial*[423] se mueve todas las variables independientes a la casilla *Variables*[424] y se cliquea sobre el botón *Descriptivos…*[425] La figura 61 presenta dicha secuencia.

[418] La cursiva es nuestra.
[419] Ibidem.
[420] Ibidem.
[421] Ibidem.
[422] Adaptado de Computing a principal component analysis, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[423] La cursiva es nuestra.
[424] Ibidem.
[425] Ibidem.

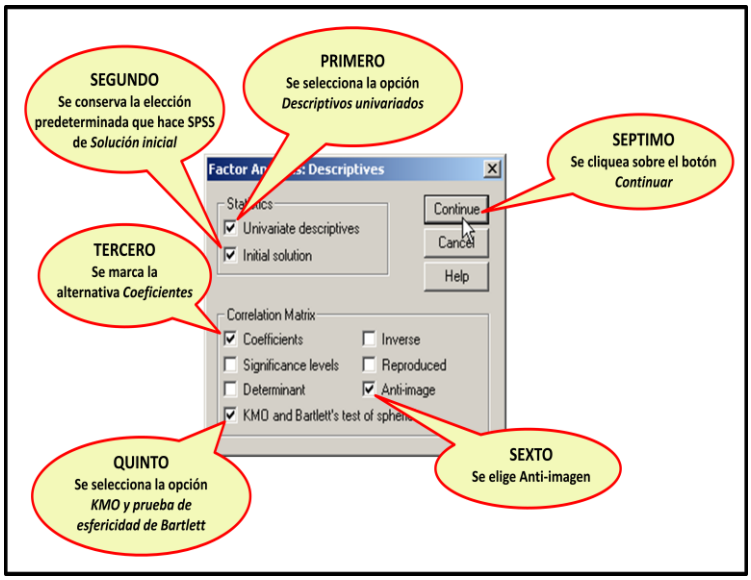
Figura 61. Elección variables independientes y estadísticos descriptivos caso Degree-Happiness[426]



Como se ilustra en la figura 62 la University of Texas at Austin (s. f.) señala que se elige las opciones *Descriptivos univariados*[427], *Coeficientes*[428], *KMO y Prueba de esfericidad de Bartlett*[429], así como *Anti-Imagen*[430]. Conforme la University of Texas at Austin (s. f.) se mantiene la alternativa *Solución inicial*[431] que SPSS establece de forma predeterminada y después se cliquea sobre el botón *Continuar*[432].

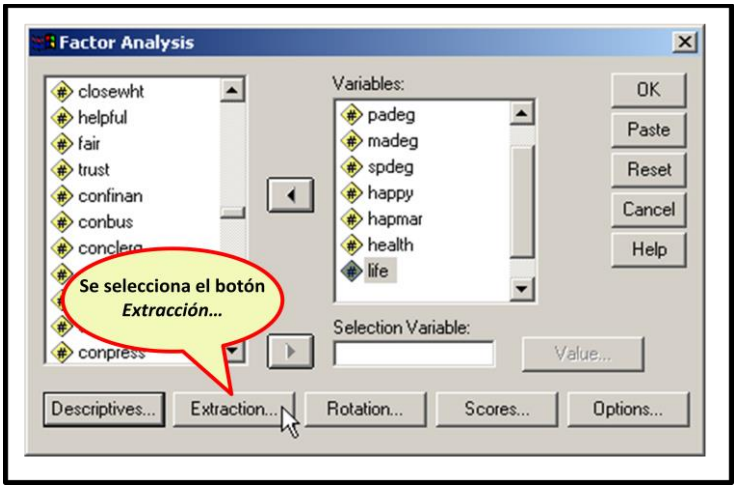
[426] Adaptado de Add the variables to the analysis, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[427] La cursiva es nuestra.
[428] Ibidem.
[429] Ibidem.
[430] Ibidem.
[431] Ibidem.
[432] Ibidem.

Figura 62. Selección de opciones en el cuadro de diálogo Análisis Factorial: Descriptivos[433]



De acuerdo a la University of Texas at Austin (s. f.) de vuelta en el cuadro de diálogo *Análisis factorial*[434] se cliquea sobre el botón *Extracción*[435]. Así, se solicita abrir el cuadro de diálogo *Análisis factorial: Extracción*[436] como se aprecia en la figura 63.

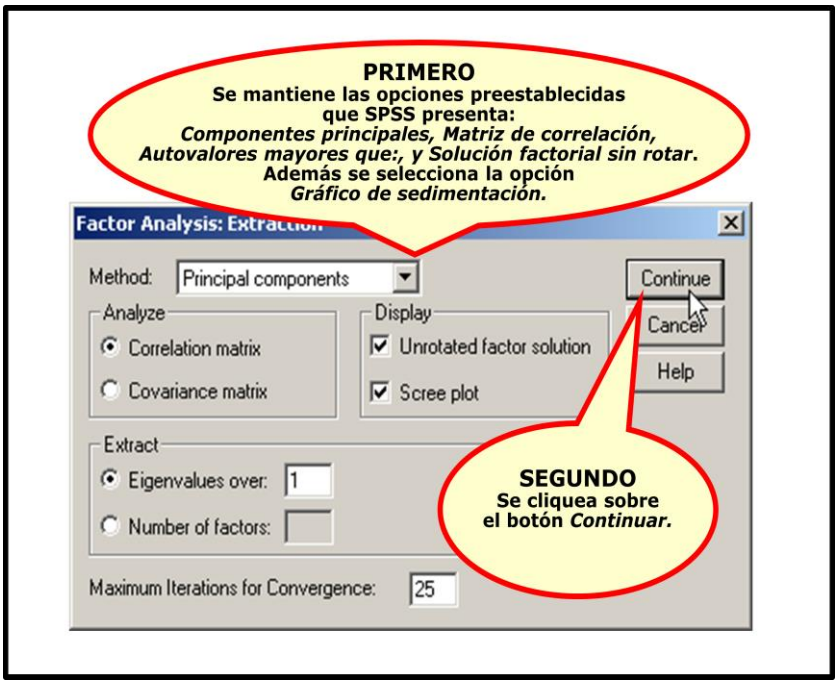
Figura 63. Solicitud cuadro de diálogo Análisis Factorial: Extracción en el caso Degree-Happiness[437]



[433] Adaptado de Compete the descriptives dialog box, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[434] La cursiva es nuestra.
[435] Ibidem.
[436] Ibidem.
[437] Adaptado de Select the extraction method, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

Como se muestra en la figura 64 según The University of Texas at Austin (s. f.) en el cuadro de diálogo *Análisis factorial: Extracción*[438] se conserva las opciones preestablecidas por SPSS, es decir, *Componentes principales*[439], *Matriz de correlación*[440], *Autovalores mayores que:*[441], *y Solución factorial sin rotar*[442]. Además se selecciona la opción *Gráfico de sedimentación*[443]. Finalmente, de acuerdo a The University of Texas at Austin (s. f.) se cliquea el botón continuar.

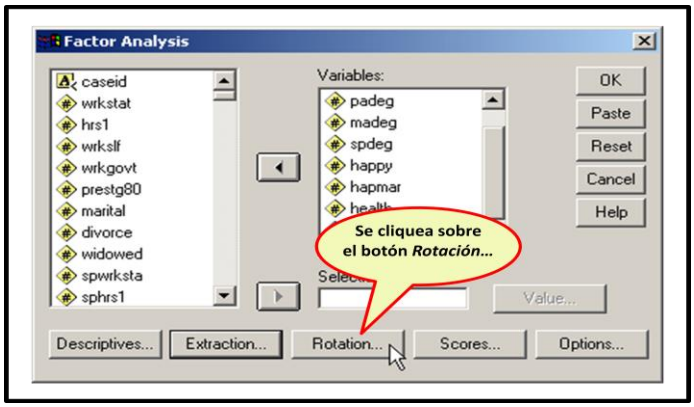
Figura 64. Solicitud gráfico de sedimentación y opciones preestablecidas caso Degree-Happiness[444]



Conforme a The University of Texas at Austin (s. f.) del mismo modo, como muestra la figura 65, se cliquea sobre el botón *Rotación*[445] para abrir el cuadro de diálogo *Análisis factorial: Rotación*[446].

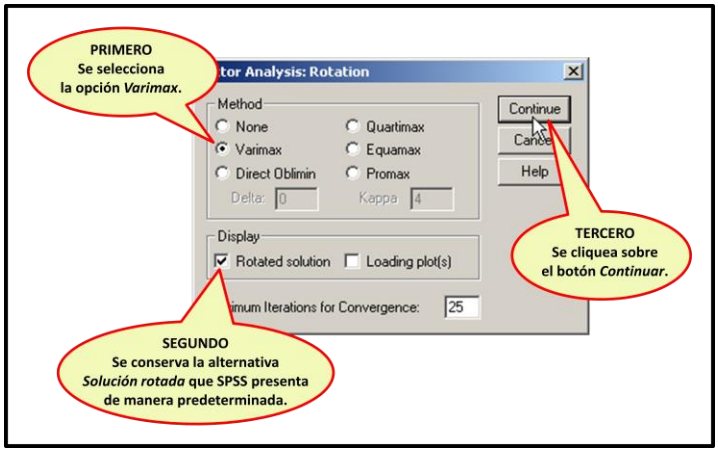
[438] La cursiva es nuestra.
[439] Ibidem.
[440] Ibidem.
[441] Ibidem.
[442] Ibidem.
[443] Ibidem.
[444] Adaptado de Compete the extraction dialog box, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

Figura 65. Elección del botón Rotación en el caso Degree-Happiness[447]



Como se aprecia en la figura 66 según The University of Texas at Austin (s. f.) en el cuadro de diálogo *Análisis factorial: Rotación*[448] se elige la alternativa *Varimax*[449] *y* se mantiene la opción *Solución rotada*[450], la cual viene seleccionada de modo predeterminado por SPSS. Posteriormente, conforme a The University of Texas at Austin (s. f.) se cliquea el botón continuar.

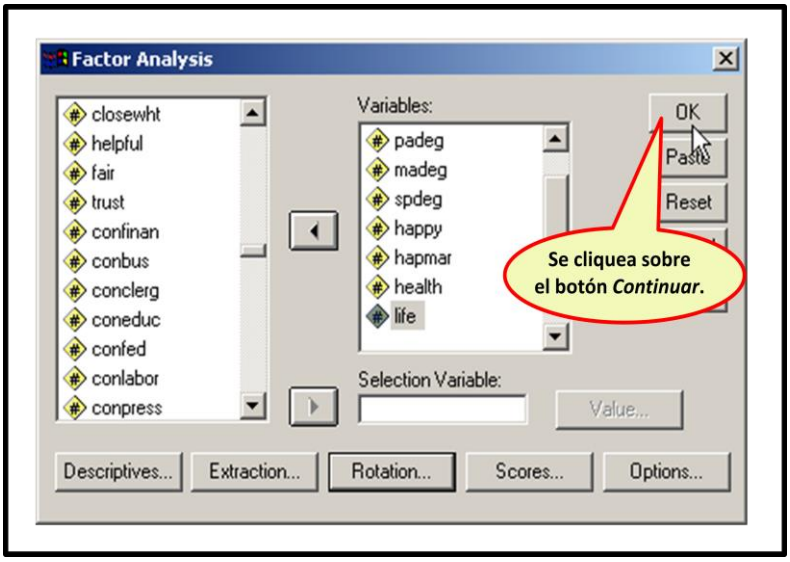
Figura 66. Selección de las opciones Varimax y Solución rotada en el caso Degree-Happiness[451]



[445] La cursiva es nuestra.
[446] Ibidem.
[447] Adaptado de Select the rotation method, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[448] La cursiva es nuestra.
[449] Ibidem.
[450] Ibidem.
[451] Adaptado de Compete the rotation dialog box, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

De acuerdo a The University of Texas at Austin (s. f.), como se aprecia en la figura 67, nuevamente en el cuadro de diálogo *Análisis factorial*[452] se cliquea sobre el botón *Continuar*[453] para obtener los resultados del análisis.

Figura 67. Solicitud de resultados del análisis componentes principales en caso Degrre-Happiness[454]



En el marco de este caso el primer elemento que se analiza es la tabla *Estadísticos descriptivos*[455]. Como se muestra en la tabla 29 conforme a la University of Texas at Austin (s. f.) en este análisis se ha considerado 68 casos válidos. Por eso, como se mencionó en el apartado *Multicolinealidad*[456], The University of Texas at Austin (s. f.) indica que se debe llegar a conclusiones cuidadosas pues se pueden realizar análisis de componentes principales con muestras mayores a 50 y menores a 100 casos siempre que las interpretaciones se hagan con cautela.

[452] La cursiva es nuestra.
[453] Ibidem.
[454] Adaptado de Complete the request for the analysis, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[455] La cursiva es nuestra.
[456] Ibidem.

Tabla 29. Estadísticos descriptivos en el caso Degree-Happiness[457]

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
RS HIGHEST DEGREE	1.68	1.085	68
FATHERS HIGHEST DEGREE	.96	.984	68
MOTHERS HIGHEST DEGREE	.85	.797	68
SPOUSES HIGHEST DEGREE	1.97	1.233	68
GENERAL HAPPINESS	1.65	.617	68
HAPPINESS OF MARRIAGE	1.47	.532	68
CONDITION OF HEALTH	1.76	.848	68
IS LIFE EXCITING OR DULL	1.53	.532	68

En la tabla *Matriz de correlaciones*[458], como se indicó en el apartado *Multicolinealidad*[459], de acuerdo a The University of Texas at Austin (s. f.) debe figurar algunas variables con una correlación mayor que 0,30. The University of Texas at Austin (s. f.) señala que en este estudio hay 7 correlaciones con esa característica las cuales aparecen sombreadas en amarillo en la tabla 30. Sin embargo es claro que en realidad hay 6 correlaciones de ese tipo ya que la correlación entre las variables Spouses highest degree y Condition of health es igual a -0,392 la cual representa una correlación menor que 0,30.

Tabla 30. Tabla Matriz de correlaciones del caso Degree-Happiness[460]

Correlation Matrix

		RS HIGHEST DEGREE	FATHERS HIGHEST DEGREE	MOTHERS HIGHEST DEGREE	SPOUSES HIGHEST DEGREE	GENERAL HAPPINESS	HAPPINESS OF MARRIAGE	CONDITION OF HEALTH	IS LIFE EXCITING OR DULL
Correlation	RS HIGHEST DEGREE	1.000	.490	.410	.595	-.017	-.172	-.246	-.138
	FATHERS HIGHEST DEGREE	.490	1.000	.677	.319	-.100	-.131	-.174	-.012
	MOTHERS HIGHEST DEGREE	.410	.677	1.000	.208	.105	-.046	-.008	.151
	SPOUSES HIGHEST DEGREE	.595	.319	.208	1.000	-.053	-.138	-.392	-.090
	GENERAL HAPPINESS	-.017	-.100	.105	-.053	1.000	.514	.267	.214
	HAPPINESS OF MARRIAGE	-.172	-.131	-.046	-.138	.514	1.000	.282	.161
	CONDITION OF HEALT	-.246	-.174	-.008	-.392	.267	.282	1.000	.214
	IS LIFE EXCITING OR DULL	-.138	-.012	.151	-.090	.214	.161	.214	1.000

[457] Adaptado de Sample size requirement: minimum number of cases, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

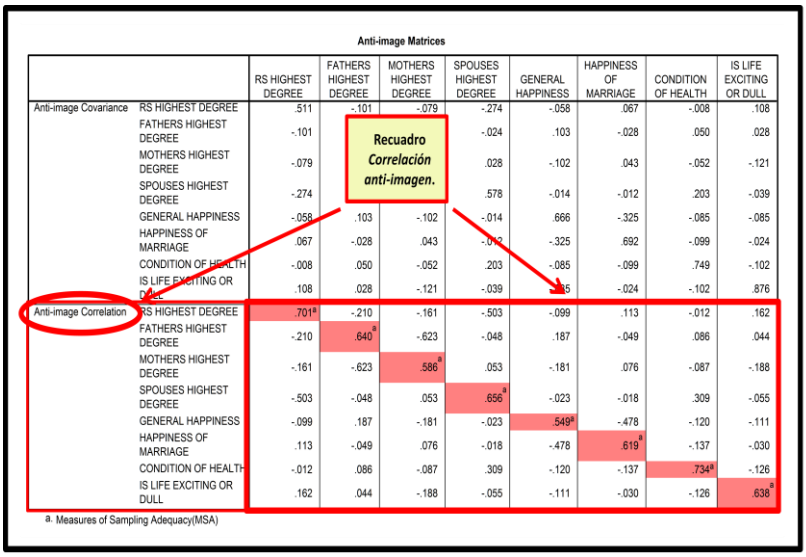
[458] La cursiva es nuestra.

[459] Ibidem.

[460] Adaptado de Appropriateness of factor analysis: Presence of substantial correlations, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

De acuerdo a The University of Texas at Austin (s. f.) en la tabla 31 se aprecia la tabla *Matrices anti-imagen*[461] en la cual destaca que los valores de la diagonal del recuadro *Correlación anti-imagen*[462] son mayores que 0,50 lo que refuerza la conservación en el estudio del conjunto de variables analizadas.

Tabla 31. Tabla Matrices anti-imagen en el caso Degree-Happiness[463]



Anti-image Matrices

		RS HIGHEST DEGREE	FATHERS HIGHEST DEGREE	MOTHERS HIGHEST DEGREE	SPOUSES HIGHEST DEGREE	GENERAL HAPPINESS	HAPPINESS OF MARRIAGE	CONDITION OF HEALTH	IS LIFE EXCITING OR DULL
Anti-image Covariance	RS HIGHEST DEGREE	.511	-.101	-.079	-.274	-.058	.067	-.008	.108
	FATHERS HIGHEST DEGREE	-.101	.103	-.024	.103	-.028	.050	.028	
	MOTHERS HIGHEST DEGREE	-.079	.028	-.102	.043	-.052	-.121		
	SPOUSES HIGHEST DEGREE	-.274	.578	-.014	-.012	.203	-.039		
	GENERAL HAPPINESS	-.058	.103	-.102	-.014	.666	-.325	-.085	-.085
	HAPPINESS OF MARRIAGE	.067	-.028	.043	-.012	-.325	.692	-.099	-.024
	CONDITION OF HEALTH	-.008	.050	-.052	.203	-.085	-.099	.749	-.102
	IS LIFE EXCITING OR DULL	.108	.028	-.121	-.039	-.025	-.024	-.102	.876
Anti-image Correlation	RS HIGHEST DEGREE	.701[a]	-.210	-.161	-.503	-.099	.113	-.012	.162
	FATHERS HIGHEST DEGREE	-.210	.640[a]	-.623	-.048	.187	-.049	.086	.044
	MOTHERS HIGHEST DEGREE	-.161	-.623	.586[a]	.053	-.181	.076	-.087	-.188
	SPOUSES HIGHEST DEGREE	-.503	-.048	.053	.656[a]	-.023	-.018	.309	-.055
	GENERAL HAPPINESS	-.099	.187	-.181	-.023	.549[a]	-.478	-.120	-.111
	HAPPINESS OF MARRIAGE	.113	-.049	.076	-.018	-.478	.619[a]	-.137	-.030
	CONDITION OF HEALTH	-.012	.086	-.087	.309	-.120	-.137	.734[a]	-.126
	IS LIFE EXCITING OR DULL	.162	.044	-.188	-.055	-.111	-.030	-.126	.638[a]

a. Measures of Sampling Adequacy(MSA)

Recuadro *Correlación anti-imagen.*

De hecho la tabla KMO y Prueba de Bartlett, mostrada en la tabla 32, refuerza la concepción de que la muestra es adecuada al análisis y que se debe conservar en el estudio las variables consideradas (KMO = 0,64). Igualmente es importante notar que la prueba de Bartlett resultó significativa (Sig. = 0,00). Así, se demuestra que existe correlación entre las variables analizadas.

[461] La cursiva es nuestra.
[462] Ibidem.
[463] Adaptado de Appropriateness of factor analysis: Sampling adequacy of individual variables, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

Tabla 32. Tabla KMO y Prueba de Bartlett en el caso Degree-Happiness[464]

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.640
Bartlett's Test of Sphericity	Approx. Chi-Square	137.823
	df	28
	Sig.	.000

Como se aprecia en la tabla 33 la tabla varianza total explicada indica la identificación de tres componentes. Según The University of Texas at Austin (s. f.) dichos componentes, todos, sobrepasan el valor 1 en la columna *Total*[465] de los autovalores iniciales y explican un 60% o más de la varianza total explicada. Conforme a The University of Texas at Austin (s. f.) en conjunto explican aproximadamente un 68% de la varianza de la variable dependiente. Ese porcentaje aparece en la tercera fila de la columna *% Acumulado*[466] en la tabla 33. Cabe destacar que The University of Texas at Austin (s. f.) no identifica cuál es la variable dependiente en este ejemplo).

Tabla 33. Detección del número de componentes en el caso Degree-Happiness[467]

Total Variance Explained

	Initial Eigenvalues			action Sums of Squared Loa			ation Sums of Squared Loadi		
Compon	Total	of Varianc	umulative %	Total	of Varianc	umulative %	Total	of Varianc	umulative %
1	2.600	32.502	32.502	2.600	32.502	32.502	2.070	25.877	25.877
2	1.772	22.149	54.651	1.772	22.149	54.651	1.770	22.119	47.996
3	1.079	13.486	68.137	1.079	13.486	68.137	1.611	20.141	68.137
4	.827	10.332	78.469						
5	.631	7.888	86.358						
6	.487	6.087	92.445						
7	.333	4.161	96.606						
8	.272	3.394	100.000						

Extraction Method: Principal Component Analysis.

[464] Adaptado de Appropriateness of factor analysis: Sampling adequacy for set of variables, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

[465] La cursiva es nuestra.

[466] Ibidem.

[467] Adaptado de Number of factors to extract: Latent root criterion, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

No obstante, en la tabla 34 de acuerdo a The University of Texas at Austin (s. f.) la tabla *Comunalidades*[468] revela que la variable *Is life exciting or dull*[469] debe ser eliminada del análisis pues el valor de su comunalidad tras la extracción de los factores es menor que 0,50.

Tabla 34. Variable Is life exciting or dull tras la extracción de factores caso Degree-Happiness[470]

Communalities

	Initial	Extraction
RS HIGHEST DEGREE	1.000	.717
FATHERS HIGHEST DEGREE	1.000	.768
MOTHERS HIGHEST DEGREE	1.000	.815
SPOUSES HIGHEST DEGREE	1.000	.715
GENERAL HAPPINESS	1.000	.763
HAPPINESS OF MARRIAGE	1.000	.711
CONDITION OF HEALTH	1.000	.548
IS LIFE EXCITING OR DULL	1.000	.415

Extraction Method: Principal Component Analysis.

Conforme a The University of Texas at Austin (s. f.) una vez retirada del análisis la variable *Is life exciting or dull*[471] se genera nuevamente el análisis de componentes principales. Tras realizar nuevamente el análisis según The University of Texas at Austin (s. f.) se detecta que la variable *Condition of health*[472] también debe eliminarse por el mismo motivo que la anterior lo cual se observa en la tabla 35.

[468] La cursiva es nuestra.
[469] Ibidem.
[470] Adaptado de Communality requiring variable removal, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[471] La cursiva es nuestra.
[472] Ibidem.

Tabla 35. Variable Condition of health tras la extracción de factores en el caso Degree-Happiness[473]

Communalities

	Initial	Extraction
RS HIGHEST DEGREE	1.000	.642
FATHERS HIGHEST DEGREE	1.000	.623
MOTHERS HIGHEST DEGREE	1.000	.592
SPOUSES HIGHEST DEGREE	1.000	.516
GENERAL HAPPINESS	1.000	.638
HAPPINESS OF MARRIAGE	1.000	.594
CONDITION OF HEALTH	1.000	.477

Extraction Method: Principal Component Analysis.

Conforme a The University of Texas at Austin (s. f.) después de retirar del análisis la variable *Condition of health*[474] se conduce otra vez el análisis de componentes principales. Según The University of Texas at Austin (s. f.) tras realizar una nueva vez el análisis se detecta que la variable *Spouses highest degree*[475] igualmente debe ser eliminada por la misma razón que se eliminó las variables anteriores. Esto se observa en la tabla 36.

Tabla 36. Variable Spouses highest degree tras la extracción de factores caso Degree-Happiness[476]

Communalities

	Initial	Extraction
RS HIGHEST DEGREE	1.000	.674
FATHERS HIGHEST DEGREE	1.000	.640
MOTHERS HIGHEST DEGREE	1.000	.577
SPOUSES HIGHEST DEGREE	1.000	.491
GENERAL HAPPINESS	1.000	.719
HAPPINESS OF MARRIAGE	1.000	.741

Extraction Method: Principal Component Analysis.

[473] Adaptado de Communality requiring variable removal, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[474] La cursiva es nuestra.
[475] Ibidem.
[476] Adaptado de Communality requiring variable removal, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).

Al realizar de nuevo el análisis sin la variable *Spouses highest degree*[477] se encuentra que todas las variables tienen comunalidades mayores que 0,50 tras la extracción de los factores. Así lo muestra la tabla 37.

Tabla 37. Comunalidades mayores que 0,5 tras la extracción de factores en caso Degree-Happiness[478]

Communalities

	Initial	Extraction
RS HIGHEST DEGREE	1.000	.577
FATHERS HIGHEST DEGREE	1.000	.720
MOTHERS HIGHEST DEGREE	1.000	.684
GENERAL HAPPINESS	1.000	.745
HAPPINESS OF MARRIAGE	1.000	.782

Extraction Method: Principal Component Analysis.

No obstante, de acuerdo a The University of Texas at Austin (s. f.) previo a considerar este conjunto de variables en el análisis se debe determinar si no hay variables con estructura compleja. En ese sentido The University of Texas at Austin (s. f.) señala que en la tabla *Matriz de componentes rotados*[479] de la tabla 38 se observa que no hay variables con cargas igual o mayor a 0,40 en ambos componentes.

Como se estableció en el apartado *Multicolinealidad*[480] según The University of Texas at Austin (s. f.) tras constatar que toda variable tiene estructura simple se confirma que cada comunalidad sea mayor que 0,50 y así corroborar que se explica una porción suficiente de la varianza de las variables originales. En este caso según The University of Texas at Austin (s. f.) todas las variables cumplen con los

[477] La cursiva es nuestra.
[478] Adaptado de Communality satisfactory for all variables, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[479] La cursiva es nuestra.
[480] Ibidem.

requisitos de tener una comunalidad mayor que 0,50 y estructura simple, por ende se acepta dicho conjunto de variables en el análisis.

Tabla 38. Matriz de componentes rotados en el caso Degree-Happiness[481]

Rotated Component Matrix[a]

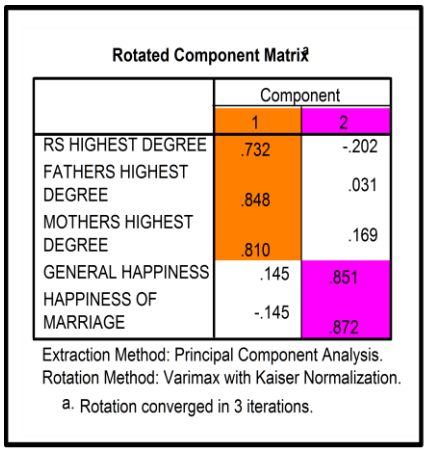
	Component	
	1	2
RS HIGHEST DEGREE	.732	-.202
FATHERS HIGHEST DEGREE	.848	.031
MOTHERS HIGHEST DEGREE	.810	.169
GENERAL HAPPINESS	.145	.851
HAPPINESS OF MARRIAGE	-.145	.872

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

La tabla 39 presenta la tabla *Matriz de componentes rotados*[482], en la que se aprecia la composición de cada componente. Conforme a The University of Texas at Austin (s. f.) el primero está constituido por las variables *Highest degree*[483], *Fathers highest degree*[484] y *Mothers highest degree*[485]. Del mismo modo de acuerdo a The University of Texas at Austin (s. f.) el segundo está conformado por las variables *General happiness*[486] y *Happiness of marriage*[487].

[481] Adaptado de Identifying complex structure, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[482] La cursiva es nuestra.
[483] Ibidem.
[484] Ibidem.
[485] Ibidem.
[486] Ibidem.
[487] Ibidem.

Tabla 39. Matriz de componentes rotados en el caso Degree-Happiness[488]



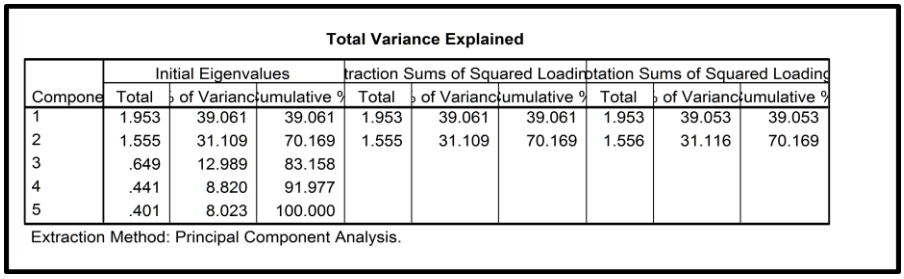
Rotated Component Matrix[a]

	Component	
	1	2
RS HIGHEST DEGREE	.732	-.202
FATHERS HIGHEST DEGREE	.848	.031
MOTHERS HIGHEST DEGREE	.810	.169
GENERAL HAPPINESS	.145	.851
HAPPINESS OF MARRIAGE	-.145	.872

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Conforme a The University of Texas at Austin (s. f.) como se aprecia en la tabla 40 ambos componentes explican prácticamente un 70% de la varianza total de las variables incluidas en los distintos componentes.

Tabla 40. Tabla Varianza total explicada del caso Degree-Happiness[489]



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loading			Rotation Sums of Squared Loading		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.953	39.061	39.061	1.953	39.061	39.061	1.953	39.053	39.053
2	1.555	31.109	70.169	1.555	31.109	70.169	1.556	31.116	70.169
3	.649	12.989	83.158						
4	.441	8.820	91.977						
5	.401	8.023	100.000						

Extraction Method: Principal Component Analysis.

En tanto, según The University of Texas at Austin (s. f.) las variables *Attitude toward life*[490], *Condition of health*[491] y *Spouse's highest academic degree*[492] fueron eliminadas de los componentes pero fueron retenidas como variables individuales. Según Tacq (1998) cuando quedan conformados los componentes principales se

[488] Adaptado de Variable loadings on components, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[489] Adaptado de Total variance explained, en *Principal component analysis: Validation, outliers, and reliability*, recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHomeworkProblems.htm, por The University of Texas at Austin, (s. f.).
[490] La cursiva es nuestra.
[491] Ibidem.
[492] Ibidem.

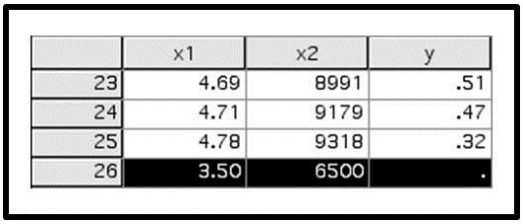
conduce un nuevo análisis de regresión considerando a dichas variables individuales y a los componentes como variables independientes. Así, conforme a Tacq (1998) el análisis de regresión estaría libre de problemas de multicolinealidad.

7.3.3. Caso de regresión múltiple en SPSS con determinación de intervalos de confianza y predicción

Para ilustrar cómo determinar los intervalos de confianza y predicción en un análisis de regresión múltiple se presenta un ejemplo propuesto por Virginia Commonwealth University (s. f.). Dicho ejemplo ha sido denominado Intervalos, haciendo referencia al objetivo del mismo. Según Virginia Commonwealth University (s. f.) imagínese que se quiere predecir a partir de un modelo de regresión debidamente construido el margen de ganancia de un banco (o varios bancos) que cuenta con 3,5 ingresos netos y 6.500 sucursales. Conforme a Virginia Commonwealth University (s. f.) en *Vista de datos*[493] y debajo del último caso en la base de datos se introduce el valor 3,5 en la casilla de la variable x_1. Igualmente de acuerdo a Virginia Commonwealth University (s. f.) se ingresa el valor 6.500 en la casilla de la variable x_2. Según Virginia Commonwealth University (s. f.) asimismo se tipea un punto (.) en la casilla de la variable dependiente *y*. Conforme a Virginia Commonwealth University (s. f.) eso le indica a SPSS que se solicita una predicción considerando dichos valores y que los mismos no deben ser incluidos en otros cálculos. De acuerdo a Virginia Commonwealth University (s. f.) la figura 68 ilustra la *Vista de datos*[494] al agregar el caso que se quiere analizar.

[493] La cursiva es nuestra.
[494] Ibidem.

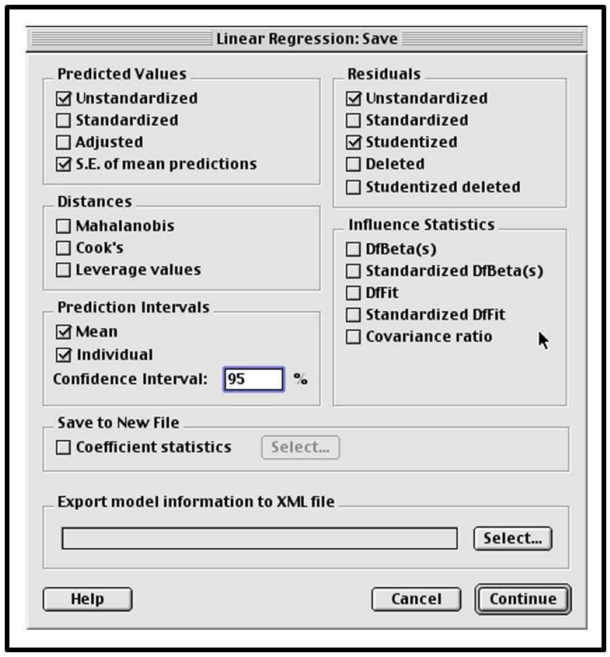
Figura 68. Vista de datos y caso agregado para analizar en el caso Intervalos[495]



	x1	x2	y
23	4.69	8991	.51
24	4.71	9179	.47
25	4.78	9318	.32
26	3.50	6500	.

Luego de acuerdo a Virginia Commonwealth University (s. f.) en el cuadro de diálogo *Regresión lineal*[496] se cliquea sobre el botón *Guardar*[497]. Después, en el cuadro de diálogo *Regresión lineal: Guardar*[498] conforme a Virginia Commonwealth University (s. f.) se selecciona *Media*[499] e *Individual*[500] y se conserva el 95% de confianza que SPSS presenta de modo predeterminado. Según Virginia Commonwealth University (s. f.) en la figura 69 se aprecia estas acciones.

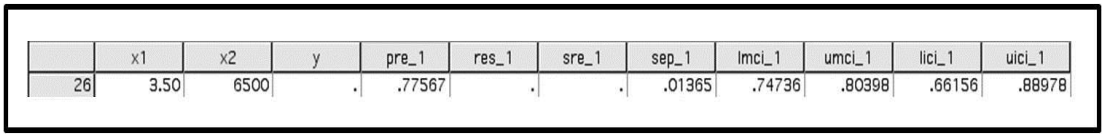
Figura 69. Solicitud de intervalos de confianza y predicción en el caso Intervalos[501]



[495] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[496] La cursiva es nuestra.
[497] Ibidem.
[498] Ibidem.
[499] Ibidem.
[500] Ibidem.
[501] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

De acuerdo a Virginia Commonwealth University (s. f.) la figura 70 presenta los resultados que emite SPSS.

Figura 70. Resultados de SPSS ante la solicitud de intervalos en el caso Intervalos[502]



	x1	x2	y	pre_1	res_1	sre_1	sep_1	lmci_1	umci_1	lici_1	uici_1
26	3.50	6500	.	.77567	.	.	.01365	.74736	.80398	.66156	.88978

Conforme a Virginia Commonwealth University (s. f.) en la casilla pre_1 de la figura 70 aparece la estimación puntual del promedio del margen de ganancia de los bancos que tienen 3,5 ingresos netos y 6.500 sucursales, que en este caso equivale a 0,77567. Según Virginia Commonwealth University (s. f.) en las casillas lmci_1 y umci_1 figura los límites inferior y superior, respectivamente, a un nivel de confianza de 95%. Conforme a Virginia Commonwealth University (s. f.) de ese modo hay 95% de confianza de que la media del margen de ganancia de los bancos con 3,5 ingresos netos y 6,500 sucursales oscila entre 0,74736 (lmci_1) y 0,80398 (umci_1).

En tanto, según a Virginia Commonwealth University (s. f.) el margen de ganancia predicho para un banco X que cuente con 3,5 ingresos netos y 6.500 sucursales también es 0,77567. Igualmente de acuerdo a Virginia Commonwealth University (s. f.) hay 95% de confianza de que el margen de ganancia para un banco X que tenga 3,5 ingresos netos y 6.500 sucursales, oscilará entre 0,66156 (lici_1) y 0,88978 (uici_1).

[502] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

7.3.4. Caso de regresión múltiple con presencia de efecto cuadrático en modelo de regresión simple inicial

Hay relaciones que se expresan mejor a través de una regresión cuadrática. Para reflejar esa situación se muestra un ejemplo presentado por Virginia Commonwealth University (s. f.). Se trata de un caso que está basado en un análisis de regresión simple que mejora el modelo cuando incluye el efecto cuadrático de la variable independiente, convirtiéndose en un modelo múltiple. Ya que el caso representa la forma más simple en la que puede expresarse el efecto cuadrático, se ha denominado al caso *SimpleC*[503].

En el modelo del caso SimpleC la variable dependiente es *Age*[504] y la variable independiente es *Price of Nissan Zs*[505]. Según Virginia Commonwealth University (s. f.) *Age*[506] está medida en *Años*[507] y *Price*[508] está medida en *Cientos de dólares*[509]. Conforme a Virginia Commonwealth University (s. f.) los valores de estas variables se presentan en la tabla 41, en función de una muestra de 31 casos.

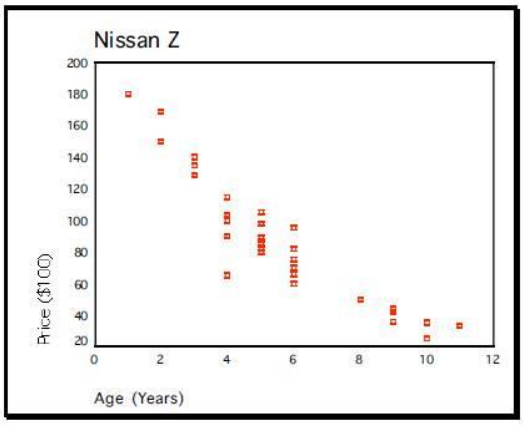
Tabla 41. Valores de las variables Age y Price en función de $n = 31$[510]

Age x	Price y	Age x	Price y	Age x	Price y	Age x	Price y
5	85	4	103	10	25	3	135
6	70	4	100	5	82	9	44
4	90	6	75	10	35	9	36
2	150	3	140	5	89	11	33
5	98	6	66	6	95	1	180
6	95	2	169	4	65	5	80
3	129	6	60	6	82	5	105
4	115	8	50	9	42		

[503] La cursiva es nuestra.
[504] Ibidem.
[505] Ibidem.
[506] Ibidem.
[507] Ibidem.
[508] Ibidem.
[509] Ibidem.
[510] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

De acuerdo a Virginia Commonwealth University (s. f.) a primera vista la relación entre *Age*[511] y *Price of Nissan Zs*[512] parece ser lineal en el rango de edad de 2 a 7 años. Sin embargo conforme a Virginia Commonwealth University (s. f.) definitivamente no lo es en el rango de 2 a 11 años. Según Virginia Commonwealth University (s. f.) ambos criterios se observan en la figura 71.

Figura 71. Diagrama de dispersión entre *x* = Age y *y* = Price en el caso SimpleC[513]



Para Virginia Commonwealth University (s. f.) los puntos del diagrama no están distribuidos conforme a una línea recta, al contrario, están distribuidos en forma de curva. De hecho, conforme a Virginia Commonwealth University (s. f.) luce que una parábola puede ajustarse bien a los datos por lo que se necesitaría una ecuación de regresión cuadrática. Así:

$$\hat{y} = b_0 + b_{1.2}x_1 + b_{2.1}x_2$$

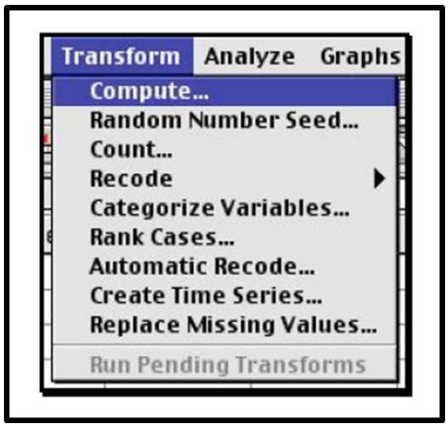
De acuerdo a Virginia Commonwealth University (s. f.) si se considera $x_1 = x_1$ y $x_2 = x_1$ cuadrado, entonces la ecuación de arriba se convierte en una ecuación de regresión

[511] La cursiva es nuestra.
[512] Ibidem.
[513] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

múltiple con dos variables predictoras (la variable *Age*[514] y la variable *Age al cuadrado*[515]): $\hat{y} = b_0 + b_{y1.2}x_1 + b_{y2.1}x_2$.

Conforme a Virginia Commonwealth University (s. f.) para lograr ese modelo se necesita antes crear la variable *Age al cuadrado*[516]. De acuerdo a Virginia Commonwealth University (s. f.) en ese sentido se elige *Calcular*[517] del menú *Transformar*[518] de SPSS. Según Virginia Commonwealth University (s. f.) la figura 72 muestra dicha secuencia.

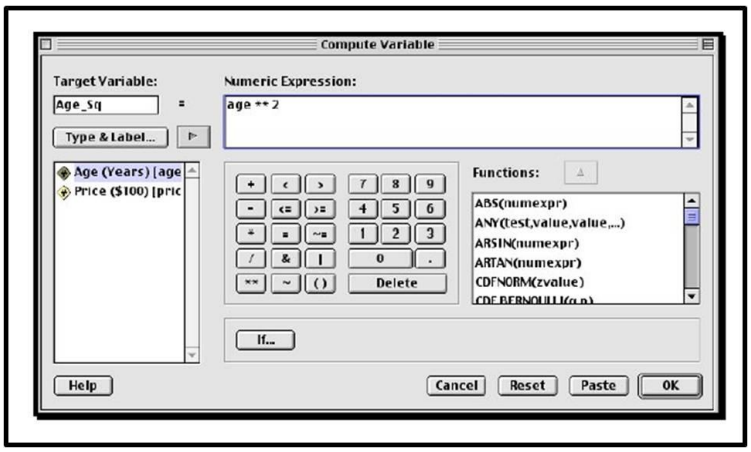
Figura 72. Selección de la opción Calcular desde el menú Transformar en el caso SimpleC[519]



En la casilla *Variable destino*[520] se tipea *Age_Sq*[521] y en *Expresión numérica*[522] se tipea *age ** 2*[523]. Luego conforme a Virginia Commonwealth University (s. f.) se cliquea sobre el botón *Aceptar*[524] para crear la nueva variable. De acuerdo a Virginia Commonwealth University (s. f.) estas acciones se presentan en la figura 73.

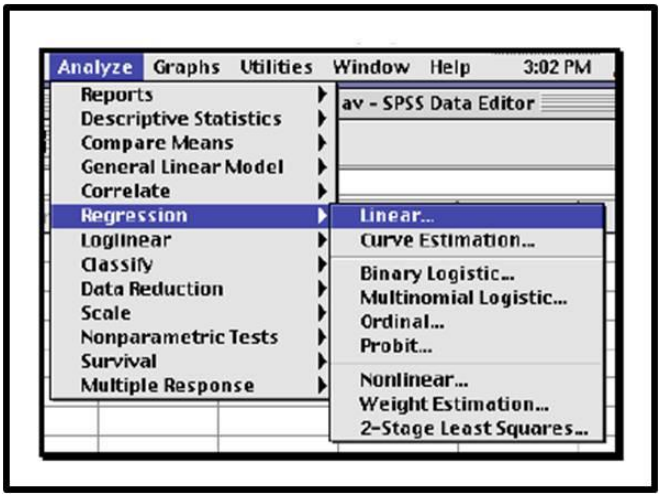
[514] La cursiva es nuestra.
[515] Ibidem.
[516] Ibidem.
[517] Ibidem.
[518] Ibidem.
[519] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[520] La cursiva es nuestra.
[521] Ibidem.
[522] Ibidem.
[523] Ibidem.
[524] Ibidem.

Figura 73. Creación de la variable Age_Sq[525]



Según Virginia Commonwealth University (s. f.) posteriormente se sigue la secuencia *Analizar*[526] → *Regresión*[527] → *Lineal…*[528] como muestra la figura 74.

Figura 74. Selección Analizar → Regresión → Lineal… en el caso SimpleC[529]



De acuerdo a Virginia Commonwealth University (s. f.) en el cuadro de diálogo *Regresión lineal*[530] se selecciona *Price*[531] como variable dependiente y *Age*[532] y *Age-*

[525] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[526] La cursiva es nuestra.
[527] Ibidem.
[528] Ibidem.
[529] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[530] La cursiva es nuestra.
[531] Ibidem.
[532] Ibidem.

sq[533] como variables independientes. Igualmente conforme a Virginia Commonwealth University (s. f.) en el marco de este caso se ha elegido el método *Introducir*[534]. Según Virginia Commonwealth University (s. f.) se selecciona *Estadísticos*[535], abriendo el cuadro de diálogo *Regresión lineal: Estadísticos*[536]. De acuerdo a Virginia Commonwealth University (s. f.) en el mismo se elige *Estimaciones*[537] e *Intervalos de confianza*[538] para los coeficientes de regresión. Asimismo de acuerdo a Virginia Commonwealth University (s. f.) se selecciona *Ajuste del Modelo*[539] y *Continuar*[540]. Después en el cuadro de diálogo *Regresión lineal*[541] según Virginia Commonwealth University (s. f.) se elige *Gráficos…*[542], generando el cuadro de diálogo *Regresión lineal: Gráficos*[543]. De acuerdo a Virginia Commonwealth University (s. f.) en el mismo se elige *Gráfico de probabilidad normal*[544], y se cliquea *Continuar*[545]. De igual modo conforme a Virginia Commonwealth University (s. f.) en el cuadro de diálogo *Regresión lineal*[546] se selecciona *Guardar…*[547], abriendo el cuadro de diálogo *Regresión lineal: Guardar*[548]. Según Virginia Commonwealth University (s. f.) en el mismo se elige *Valores predichos no estandarizados*[549], *Error estándar de las predicciones de la media*[550], *Residuales no estudentizados*[551] y *Residuales estudentizados*[552], *Intervalos de*

[533] La cursiva es nuestra.
[534] Ibidem.
[535] Ibidem.
[536] Ibidem.
[537] Ibidem.
[538] Ibidem.
[539] Ibidem.
[540] Ibidem.
[541] Ibidem.
[542] Ibidem.
[543] Ibidem.
[544] Ibidem.
[545] Ibidem.
[546] Ibidem.
[547] Ibidem.
[548] Ibidem.
[549] Ibidem.
[550] Ibidem.
[551] Ibidem.
[552] Ibidem.

predicción individual[553] *y Intervalos de predicción media*[554] *al nivel de* 90%, y se cliquea *Continuar*[555]. Conforme a Virginia Commonwealth University (s. f.) las figuras 75, 76, 77 y 78 muestran estas acciones.

Figura 75. Especificación de variables dependiente e independientes en el caso SimpleC[556]

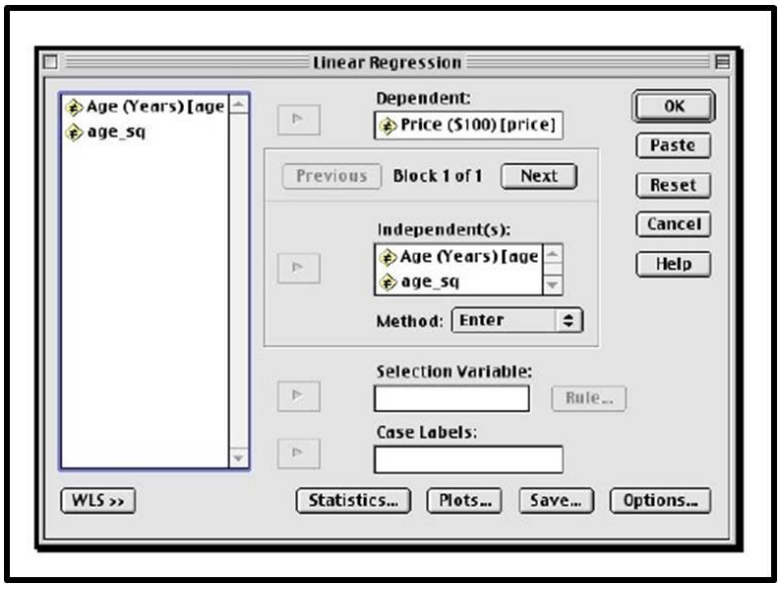
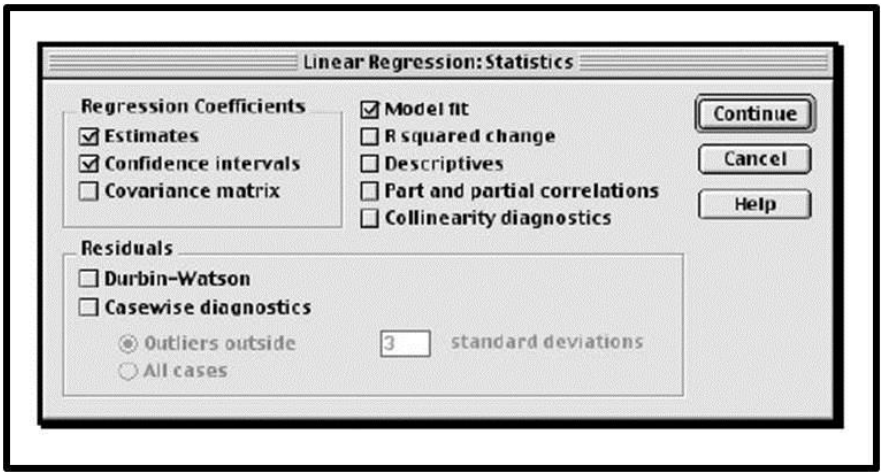


Figura 76. Elección opciones en cuadro de diálogo Regresión Lineal: Estadísticos del caso SimpleC[557]



[553] La cursiva es nuestra.
[554] Ibidem.
[555] Ibidem.
[556] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[557] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

Figura 77. Solicitud de gráficos residuales y de probabilidad normal en el caso SimpleC[558]

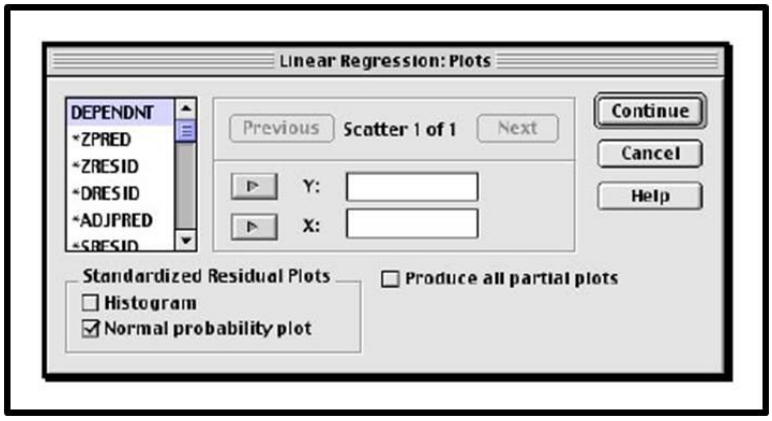
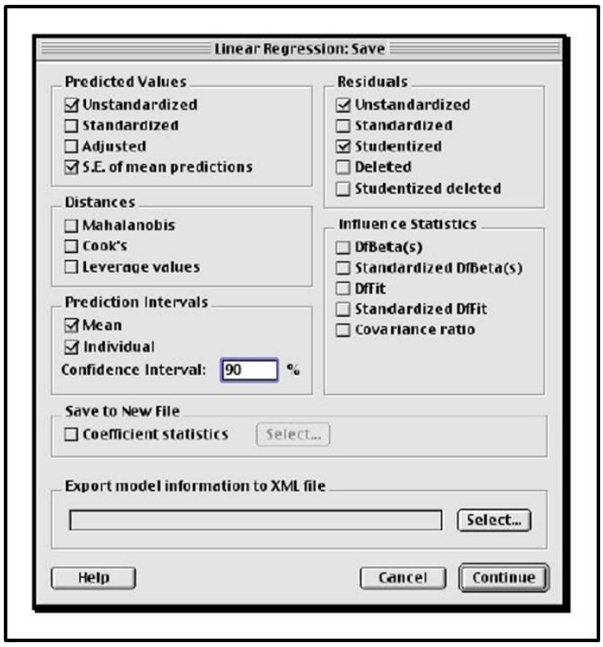


Figura 78. Solicitud de opciones en cuadro de diálogo Regresión lineal: Guardar del caso SimpleC[559]



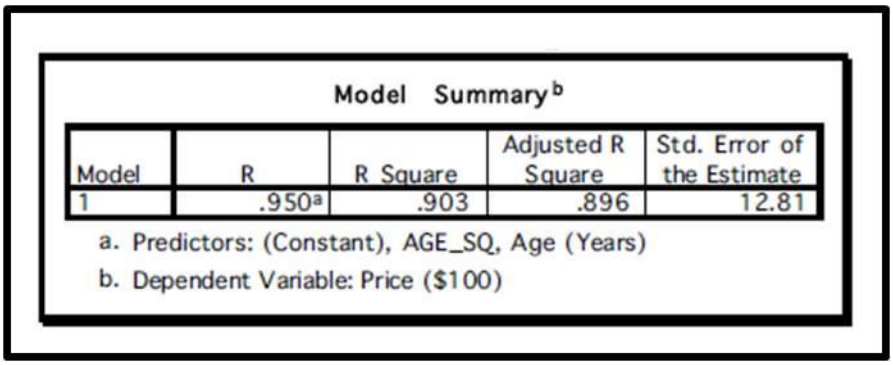
Una vez especificadas las solicitudes arriba mencionadas, los resultados son los siguientes.

Según Virginia Commonwealth University (s. f.) en la tabla *Resumen del modelo*[560] de la tabla 42 se observa que la ecuación lograda explica aproximadamente el 90% de

[558] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[559] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[560] La cursiva es nuestra.

la varianza de la variable dependiente *Price*[561], con *Adj.* R^2 = 0,896.

Tabla 42. Tabla Resumen del modelo en el caso SimpleC[562]



Model Summary[b]

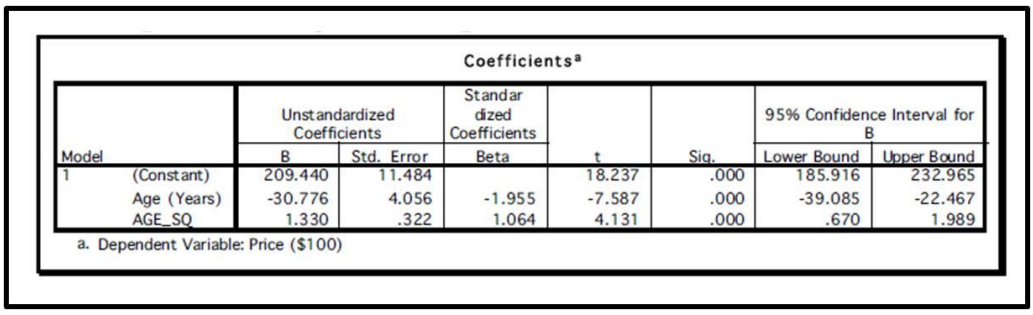
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.950[a]	.903	.896	12.81

a. Predictors: (Constant), AGE_SQ, Age (Years)

b. Dependent Variable: Price ($100)

De acuerdo a Virginia Commonwealth University (s. f.) la tabla de coeficientes que aparece en la tabla 43 muestra que la ecuación de regresión cuadrática que se obtuvo fue $\hat{y}$ = 209,44 - 30,776 x_1 + 1,330x^2. Según Virginia Commonwealth University (s. f.) en la misma tanto el intercepto como los coeficientes de regresión resultaron significativos con p = 0,00.

Tabla 43. Tabla Coeficientes del caso SimpleC[563]



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	209.440	11.484		18.237	.000	185.916	232.965
	Age (Years)	-30.776	4.056	-1.955	-7.587	.000	-39.085	-22.467
	AGE_SQ	1.330	.322	1.064	4.131	.000	.670	1.989

a. Dependent Variable: Price ($100)

Asimismo conforme a Virginia Commonwealth University (s. f.) la tabla ANOVA de la tabla 44 señala que el modelo es significativo con p = 0,00.

[561] La cursiva es nuestra.

[562] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

[563] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

Tabla 44. Tabla ANOVA del caso SimpleC[564]

ANOVA[b]

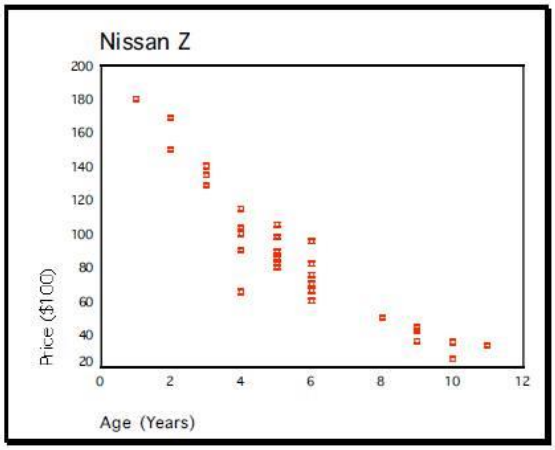
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42895.358	2	21447.679	130.698	.000[a]
	Residual	4594.836	28	164.101		
	Total	47490.194	30			

a. Predictors: (Constant), AGE_SQ, Age (Years)

b. Dependent Variable: Price ($100)

Analizando los supuestos de acuerdo a la Virginia Commonwealth University (s. f.) se observa que el diagrama de dispersión entre las variables *Age*[565] y *Price*[566] genera un patrón curvilíneo, por lo que en ese sentido el modelo cuadrático se ajustará bien a los datos. De ese modo según Virginia Commonwealth University (s. f.) se satisface el supuesto de linealidad. Igualmente conforme a Virginia Commonwealth University (s. f.) el mencionado diagrama no parece mostrar casos atípicos u observaciones influyentes. De acuerdo a Virginia Commonwealth University (s. f.) la figura 79 muestra el diagrama en cuestión.

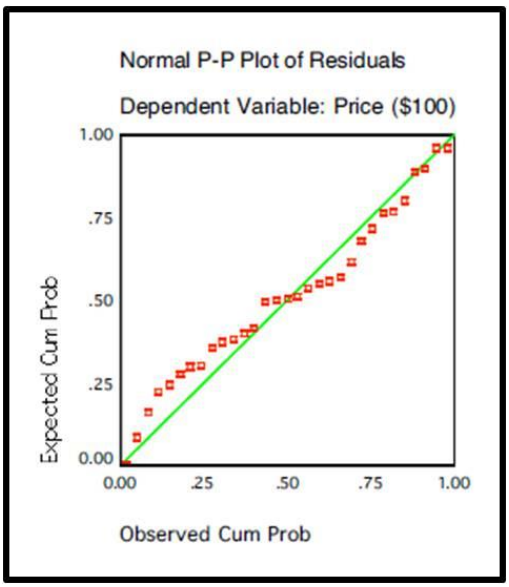
Figura 79. Diagrama de dispersión entre Age y Price en el caso SimpleC[567]



[564] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[565] La cursiva es nuestra.
[566] Ibidem.
[567] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

Mientras, conforme a Virginia Commonwealth University (s. f.) la figura 80 presenta el gráfico de probabilidad normal. Según Virginia Commonwealth University (s. f.) en el mismo se observa que los puntos se ajustan bien a la recta, por lo que se satisface el supuesto de normalidad.

Figura 80. Gráfico de probabilidad normal del caso SimpleC[568]



Según Virginia Commonwealth University (s. f.) la figura 81 representa el análisis residual entre *Age*[569] y los residuales estudentizados, así como la figura 82 muestra el análisis residual entre *Age_Sq*[570] y los residuales estudentizados. Conforme a Virginia Commonwealth University (s. f.) en ambos casos se nota que no hay un patrón aparente y que los puntos se distribuyen a lo largo de una banda horizontal. Por eso de acuerdo a Virginia Commonwealth University (s. f.) se establece que se cumplen los supuestos de linealidad y homocedasticidad.

[568] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[569] La cursiva es nuestra.
[570] Ibidem.

Figura 81. Análisis residual entre Age y los residuales estudentizados[571]

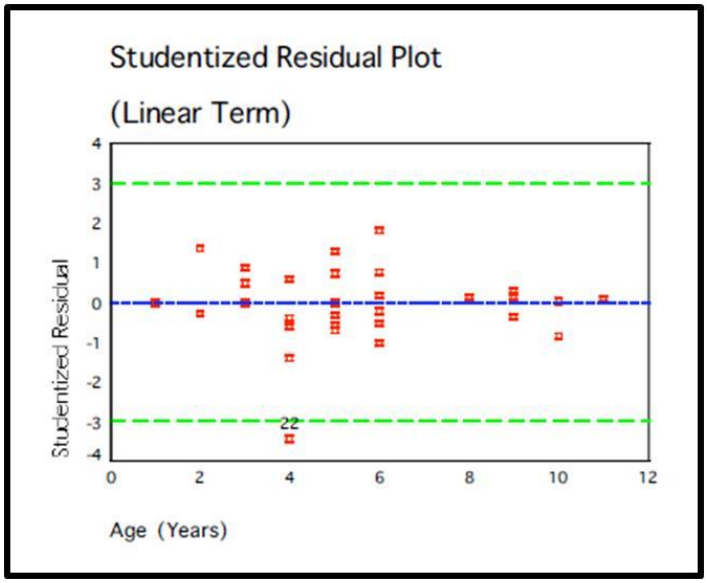
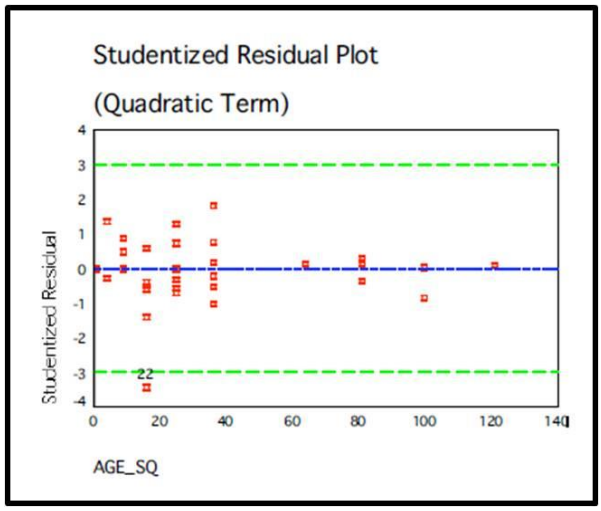


Figura 82. Análisis residual entre Age_Sq y los residuales estudentizados[572]



Como se aprecia en las figuras 81 y 82 de arriba según Virginia Commonwealth University (s. f.) la observación número 22 es una observación inusual en ambos gráficos. Su residual conforme a Virginia Commonwealth University (s. f.) está más allá de las 3 desviaciones estándar, lo que señala que puede ser una observación atípica.

[571] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).
[572] Adaptado de *Multiple regression in SPSS*, recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf, por Virginia Commonwealth University, (s. f.).

En tanto, a continuación se presenta un resumen (cuya gráfica no se muestra) de los niveles de confianza y de predicción que se obtuvieron considerando un Nissan Z de 8 años. De acuerdo a Virginia Commonwealth University (s. f.) se establece que el precio medio esperado para los Nissan Z que tengan 8 años es de USD$4.833,21. Sin embargo según Virginia Commonwealth University (s. f.) con un 90% la media del precio de los Nissan Z de 8 años oscila entre USD$4.258,85 (lmci_1) y USD$5.407,58 (umci_1).

Mientras, conforme a Virginia Commonwealth University (s. f.) el precio predicho para el Nissan Z de 8 años que está en venta de Bob Smith (por mencionar un nombre cualquiera) es USD$4.833,21. No obstante, de acuerdo a Virginia Commonwealth University (s. f.) con un 90% de confianza se establece que el precio del Nissan Z de 8 años que está en venta de Bob Smith oscila entre USD$2.579,61 (lici_1) y USD$7.086,82 (uici_1).

7.3.5. Caso de regresión múltiple con variables categóricas

Para ilustrar cómo lidiar con la presencia de variables categóricas en el análisis de regresión Barón López y otros (s. f.) proponen un ejemplo al que en este estudio se denomina Categ. Conforme a Barón López y otros (s. f.) se trata de un experimento en el que se compara tres métodos de aprendizaje de lectura denominados Técnica I, Técnica II y Control. Según Barón López y otros (s. f.) cada método representó un grupo al que se le asignó aleatoriamente 22 estudiantes. De acuerdo a Barón López y otros (s. f.) la capacidad de comprensión de los mismos se evaluó antes y después del experimento. Así, Barón López y otros (s. f.) señalan que la capacidad de

comprensión después del experimento se codificó como indica la tabla 45. En ese sentido Barón López y otros (s. f.) indican que las variables independientes son la capacidad de comprensión antes del experimento, la técnica I y la técnica II. Igualmente para Barón López y otros (s. f.) la variable dependiente está constituida por la diferencia entre la capacidad de comprensión después de participar en el experimento (tras la aplicación de una o ninguna técnica) y la capacidad de comprensión antes de hacerlo.

Tabla 45. Códigos según grupo en el caso Categ[573]

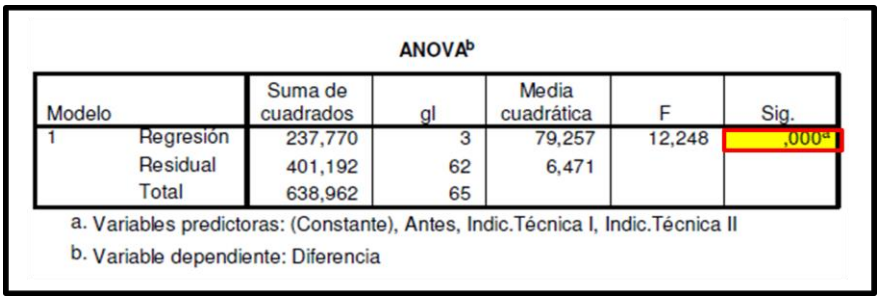
Grupo	**Indicador Técnica I**	**Indicador Técnica II**
Control	0	0
Técnica I	1	0
Técnica II	0	1

Según Barón López y otros (s. f.) en un análisis de varianza (ANOVA) realizado anteriormente se había encontrado una relación estadísticamente significativa entre las variables. A fin de obtener mayor información al respecto, Barón López y otros (s. f.) indican que se realiza un análisis de regresión.

Así, de acuerdo a Barón López y otros (s. f.) en la tabla ANOVA de la tabla 46 se observa que el modelo es significativo.

[573] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

Tabla 46. Tabla ANOVA del caso Categ[574]



ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	237,770	3	79,257	12,248	,000[a]
	Residual	401,192	62	6,471		
	Total	638,962	65			

a. Variables predictoras: (Constante), Antes, Indic.Técnica I, Indic.Técnica II

b. Variable dependiente: Diferencia

En la tabla 47 se muestra la tabla Coeficientes, en la que Barón López y otros (s. f.) establecen que todos los coeficientes también son significativos. La interpretación de los mismos se hace de igual forma como en el apartado *Variables cualitativas más complejas*[575]. Igualmente Barón López y otros (s. f.) señalan que el intercepto indica que los estudiantes del grupo control obtendrían un promedio de 13,56 puntos de diferencia entre su puntuación después del experimento y su puntuación inicial antes del mismo siempre que ésta sea 0. Asimismo, según Barón López y otros (s. f.) el coeficiente de regresión de x_1 señala que, en promedio, los estudiantes a los que se les aplique la técnica 1 obtendrían por cada punto alcanzado en la misma 3,41 puntos de diferencia más que los del grupo control. De igual modo conforme a Barón López y otros (s. f.) el coeficiente de regresión de x_2 establece que, en promedio, a los alumnos que se les aplique la técnica 2 obtendrían por cada punto alcanzado en la misma 2,83 puntos de diferencia más que los del grupo control. Igualmente Barón López y otros (s. f.) establecen que el coeficiente de regresión de x_3 indica que sin importar la técnica de aprendizaje que se aplique a los alumnos se espera que por cada punto logrado en la prueba inicial, en promedio, los estudiantes obtengan 0,47 puntos de diferencia menos. En tanto, también se puede observar que por cada unidad alcanzada en la post-prueba los integrantes de la técnica 1, en promedio, obtendrían

[574] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

[575] La cursiva es nuestra.

0,58 puntos de diferencia más que los integrantes de la técnica 2. Esto se hace evidente al realizar los cálculos que figuran abajo a partir de la ecuación de regresión estimada $\hat{y} = 13{,}56 + 3{,}41\ x_1 + 2{,}83\ x_2 - 0{,}47\ x_3$.

Con técnica 1: $\hat{y} = 13{,}56 + 3{,}41(1) - 0{,}47 = 16{,}50$

Con técnica 2: $\hat{y} = 13{,}56 + 2{,}83(1) - 0{,}47 = 15{,}92$

Así: $16{,}50 - 15{,}92 = 0{,}58$

Tabla 47. Tabla Coeficientes del caso Categ[576]

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
		B	Error típ.	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	13,557	1,292		10,494	,000	10,975	16,140
	Indic.Técnica I	3,406	,770	,516	4,422	,000	1,866	4,945
	Indic.Técnica II	2,827	,777	,428	3,637	,001	1,273	4,380
	Antes	-,467	,149	-,321	-3,144	,003	-,765	-,170

a. Variable dependiente: Diferencia

En la tabla Resumen del modelo de la tabla 48 se observa que el modelo explica aproximadamente un 34% de la varianza de la variable dependiente.

Tabla 48. Tabla Resumen del modelo del caso Categ[577]

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,610[a]	,372	,342	2,54378

a. Variables predictoras: (Constante), Antes, Indic.Técnica I, Indic.Técnica II

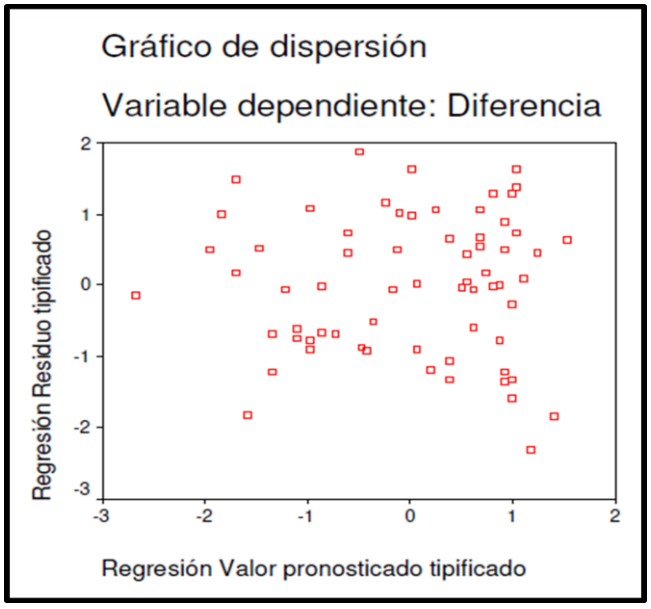
b. Variable dependiente: Diferencia

[576] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

[577] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

En tanto, según Barón López y otros (s. f.) la figura 83 muestra la gráfica entre residuos tipificados y valores pronosticados tipificados. De acuerdo a Barón López y otros (s. f.) en la misma no se aprecia un patrón en los puntos por lo que se satisface el supuesto de linealidad. Conforme a Barón López y otros (s. f.) se observa también que los puntos se distribuyen a lo largo de una banda horizontal por lo que se satisface el supuesto de homocedasticidad. La gráfica de probabilidad normal no se presenta en el ejemplo. No obstante, Barón López y otros (s. f.) establecen que en la misma los puntos se ciñen a la recta los suficiente como para dar por cumplido el supuesto de normalidad. Asimismo, tampoco se mostró el valor del estadístico de Durbin-Watson pero se asume que la muestra cumple con el supuesto de independencia de los errores.

Figura 83. Gráfico entre residuos tipificados y valores pronosticados tipificados del caso Categ[578]



[578] Adaptado de Variables confusoras, en *Apuntes de bioestadística*, recuperado de http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf, por F. J. Barón López y F. Téllez Montiel, (s. f.).

8. Propuesta de un procedimiento superador que en un momento específico determine el éxito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina

1- Hacer entrevistas en profundidad a expertos en política y Web 2.0 para determinar en la principal cuenta Facebook de un candidato a Diputado cuáles son las variables que pueden potencialmente influir en la intención de voto de los electores argentinos en un momento dado de una campaña legislativa general nacional. Si se tiene el conocimiento propio de un experto en política y Web 2.0 dichas variables pueden ser establecidas conforme al propio criterio.

2- Establecer las variables que el (los) experto (s) sugiere (n) como potenciales indicadores de la referida influencia.

3- Identificar puntualmente los elementos que compondrán la población.

4- Determinar el tamaño de muestra mínimo siguiendo las indicaciones especificadas en el apartado *Tamaño mínimo muestral*[579]. Si no se alcanza dicho tamaño mínimo debido a que la cantidad de candidatos con cuenta Facebook que participan en la contienda electoral no es suficiente se concluye que no se puede realizar la investigación porque no se satisfacen los requerimientos del tamaño muestral mínimo.

[579] La cursiva es nuestra.

5- Determinar el tamaño de muestra según las indicaciones especificadas en el apartado *Tamaño de la muestra*[580].

6- Calcular la muestra conforme se señala en el apartado *Tipo de muestra*[581].

7- Seleccionar los elementos muestrales de acuerdo a lo establecido en el apartado *Selección de los elementos muestrales*[582].

8- Especificar los elementos muestrales seleccionados.

9- Hacer un relevamiento de las variables identificadas como potenciales indicadores en la principal cuenta Facebook de los candidatos que formen parte de la muestra.

10- Tomar el valor de la intención de voto de dichos candidatos a partir de los datos que arroje un estudio que a los fines realice una consultora de investigación de mercados en la cual se confíe o un especialista en este tipo de medición en el que se confíe.

11- Realizar un análisis de regresión simple o múltiple. Eso dependerá fundamentalmente del tamaño de la población, el cual establece el tamaño mayor de muestra que se puede pretender. Eso a su vez permite definir la cantidad tope de variables independientes que podrán considerarse en el análisis. Si se puede considerar una sola variable independiente se permite realizar un análisis de regresión lineal simple. En cambio, si se puede

[580] La cursiva es nuestra.
[581] Ibidem.
[582] Ibidem.

considerar dos o más variables se permite efectuar un análisis de regresión múltiple.

12- En caso de construir un modelo de regresión que explique en cierta medida la intención de voto en un momento determinado de unas elecciones legislativas generales nacionales, se detalla cuáles variables conforman dicho modelo y a la vez se establece el modelo en cuestión. En caso contrario se concluye que en el momento de la elección en que se realizó el estudio las variables consideradas en no influyen en la intención del voto. Dicho de otro modo, se concluye que en ese momento de la elección legislativa general nacional la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido no influye en la intención de voto por lo cual no es exitosa.

9. Conclusión

Esta tesis establece que el Procedimiento Superador supera al Procedimiento de Medición Face2.0 Existente en la Argentina. De hecho, se indica que este último constituye una metodología inapropiada para determinar el éxito de la e-campaña de un candidato electoral en particular. Mientras, se demuestra que el Procedimiento Superador es adecuado para determinar en un momento dado el éxito de la e-campaña vía Facebook de un candidato a Diputado que encabece la lista de su partido en unas elecciones legislativas generales nacionales de la Argentina. Cabe destacar que el Procedimiento Superador puede aplicarse a todas las redes sociales. En cualquier caso para obtener resultados fiables en base al mismo es preciso respetar sus criterios

de aplicación y generar el análisis a partir de relevamientos debidamente realizados.

10. Anexos

Tabla A1. Tabla *t* de una cola[583].

Valores t de Student y probabilidad P asociada en función de los grados de libertad gl.

	P (de una cola)									
gl	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.596
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.706
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
Infinito	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

[583] Adaptado de *Valores t de Student y probabilidad P asociada en función de los grados de libertad gl*, recuperado de http://www.uv.es/~meliajl/Docencia/Tablas/TablaT.PDF, por Universidad de Valencia, (s. f.).

Tabla A2. Tabla *t* de dos colas[584].

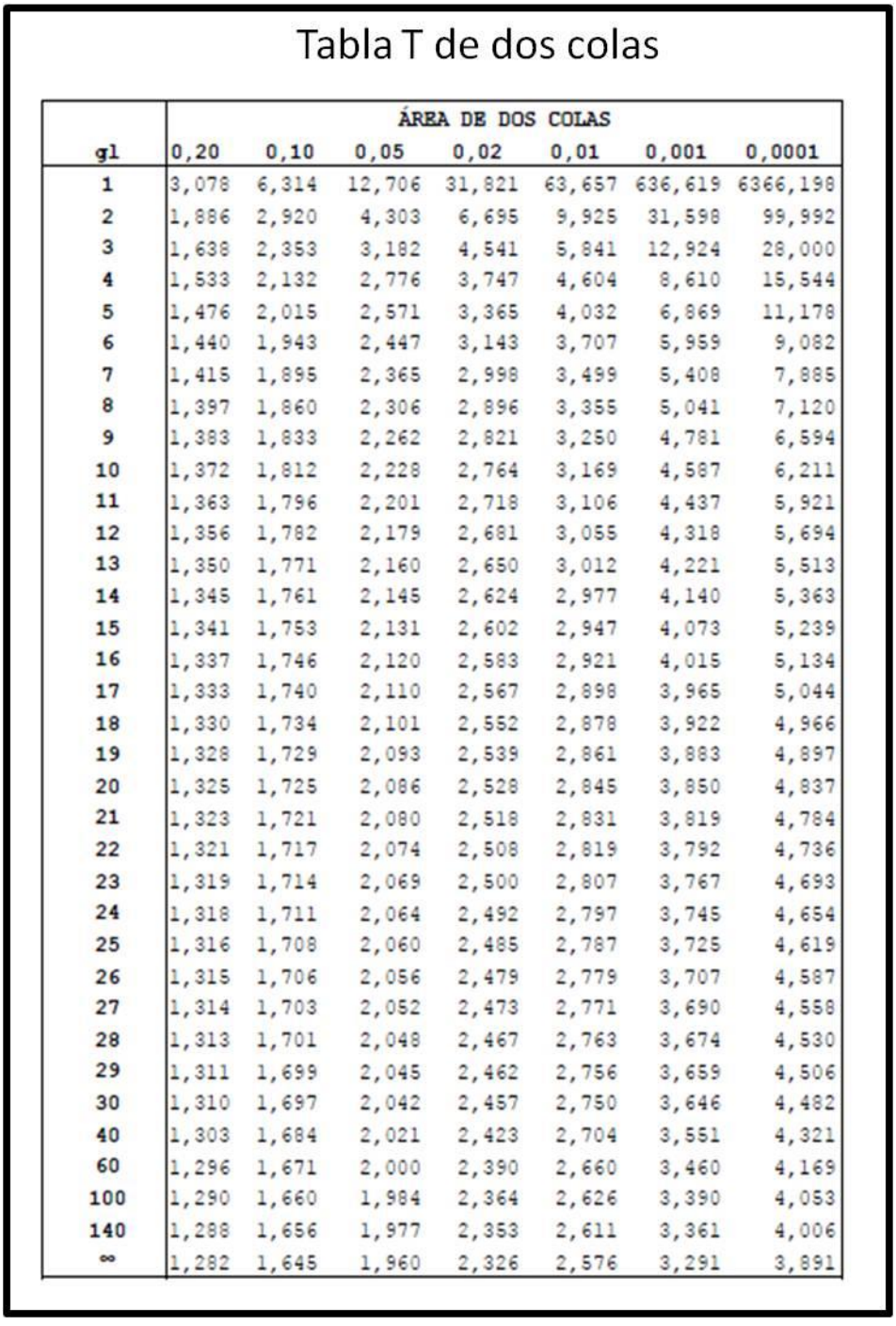


Tabla T de dos colas

	ÁREA DE DOS COLAS						
gl	0,20	0,10	0,05	0,02	0,01	0,001	0,0001
1	3,078	6,314	12,706	31,821	63,657	636,619	6366,198
2	1,886	2,920	4,303	6,695	9,925	31,598	99,992
3	1,638	2,353	3,182	4,541	5,841	12,924	28,000
4	1,533	2,132	2,776	3,747	4,604	8,610	15,544
5	1,476	2,015	2,571	3,365	4,032	6,869	11,178
6	1,440	1,943	2,447	3,143	3,707	5,959	9,082
7	1,415	1,895	2,365	2,998	3,499	5,408	7,885
8	1,397	1,860	2,306	2,896	3,355	5,041	7,120
9	1,383	1,833	2,262	2,821	3,250	4,781	6,594
10	1,372	1,812	2,228	2,764	3,169	4,587	6,211
11	1,363	1,796	2,201	2,718	3,106	4,437	5,921
12	1,356	1,782	2,179	2,681	3,055	4,318	5,694
13	1,350	1,771	2,160	2,650	3,012	4,221	5,513
14	1,345	1,761	2,145	2,624	2,977	4,140	5,363
15	1,341	1,753	2,131	2,602	2,947	4,073	5,239
16	1,337	1,746	2,120	2,583	2,921	4,015	5,134
17	1,333	1,740	2,110	2,567	2,898	3,965	5,044
18	1,330	1,734	2,101	2,552	2,878	3,922	4,966
19	1,328	1,729	2,093	2,539	2,861	3,883	4,897
20	1,325	1,725	2,086	2,528	2,845	3,850	4,837
21	1,323	1,721	2,080	2,518	2,831	3,819	4,784
22	1,321	1,717	2,074	2,508	2,819	3,792	4,736
23	1,319	1,714	2,069	2,500	2,807	3,767	4,693
24	1,318	1,711	2,064	2,492	2,797	3,745	4,654
25	1,316	1,708	2,060	2,485	2,787	3,725	4,619
26	1,315	1,706	2,056	2,479	2,779	3,707	4,587
27	1,314	1,703	2,052	2,473	2,771	3,690	4,558
28	1,313	1,701	2,048	2,467	2,763	3,674	4,530
29	1,311	1,699	2,045	2,462	2,756	3,659	4,506
30	1,310	1,697	2,042	2,457	2,750	3,646	4,482
40	1,303	1,684	2,021	2,423	2,704	3,551	4,321
60	1,296	1,671	2,000	2,390	2,660	3,460	4,169
100	1,290	1,660	1,984	2,364	2,626	3,390	4,053
140	1,288	1,656	1,977	2,353	2,611	3,361	4,006
∞	1,282	1,645	1,960	2,326	2,576	3,291	3,891

[584] Adaptado de *Tabla de cuantiles de la distribución t de Student*, recuperado de http://oromeroeii.blogcindario.com/ficheros/t-studentdoscolas.pdf, por Universidad de Valencia, (s. f.).

Tabla A3. Tabla *F* con alfa = 0,05 y con grados de libertad del numerador[585] entre 1 y 15[586]

		GLN												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**15**
GLD	**1**	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	245.9
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.43
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.07
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.06
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.04
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.03
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84
	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.75

[585] En la tabla los grados de libertad del numerador y del denominador se expresan por las siglas GLN y GLD respectivamente

[586] Adaptado de *Distribución F 0.05*, recuperado de http://www.uaaan.mx/~jmelbos/tablas/distf.pdf, por J. A. Mellado Bosque, (s. f.).

Tabla A4. Tabla *F* con alfa = 0,05 y con grados de libertad del numerador[587] entre 20 y 120[588]

		GLN					
		20	**24**	**30**	**40**	**60**	**120**
	1	248.0	249.1	250.1	251.1	252.2	253.3
	2	19.45	19.45	19.46	19.47	19.48	19.49
	3	8.66	8.64	8.62	8.59	8.57	8.55
	4	5.80	5.77	5.75	5.72	5.69	5.66
	5	4.56	4.53	4.50	4.46	4.43	4.40
	6	3.87	3.84	3.81	3.77	3.74	3.70
	7	3.44	3.41	3.38	3.34	3.30	3.27
	8	3.15	3.12	3.08	3.04	3.01	2.97
	9	2.94	2.90	2.86	2.83	2.79	2.75
	10	2.77	2.74	2.70	2.66	2.62	2.58
	11	2.65	2.61	2.57	2.53	2.49	2.45
	12	2.54	2.51	2.47	2.43	2.38	2.34
	13	2.46	2.42	2.38	2.34	2.30	2.25
	14	2.39	2.35	2.31	2.27	2.22	2.18
	15	2.33	2.29	2.25	2.20	2.16	2.11
G	**16**	2.28	2.24	2.19	2.15	2.11	2.06
L	**17**	2.23	2.19	2.15	2.10	2.06	2.01
N	**18**	2.19	2.15	2.11	2.06	2.02	1.97
	19	2.16	2.11	2.07	2.03	1.98	1.93
	20	2.12	2.08	2.04	1.99	1.95	1.90
	21	2.10	2.05	2.01	1.96	1.92	1.87
	22	2.07	2.03	1.98	1.94	1.89	1.84
	23	2.05	2.01	1.96	1.91	1.86	1.81
	24	2.03	1.98	1.94	1.89	1.84	1.79
	25	2.01	1.96	1.92	1.87	1.82	1.77
	26	1.99	1.95	1.90	1.85	1.80	1.75
	27	1.97	1.93	1.88	1.84	1.79	1.73
	28	1.96	1.91	1.87	1.82	1.77	1.71
	29	1.94	1.90	1.85	1.81	1.75	1.70
	30	1.93	1.89	1.84	1.79	1.74	1.68
	40	1.84	1.79	1.74	1.69	1.64	1.58
	60	1.75	1.70	1.65	1.59	1.53	1.47
	120	1.66	1.61	1.55	1.50	1.43	1.35

[587] En la tabla los grados de libertad del numerador y del denominador se expresan por las siglas GLN y GLD respectivamente

[588] Adaptado de *Distribución F 0.05*, recuperado de http://www.uaaan.mx/~jmelbos/tablas/distf.pdf, por J. A. Mellado Bosque, (s. f.).

Tabla A5. Tabla *F* con alfa = 0,01 y con grados de libertad del numerador[589] entre 1 y 15[590]

		GLN												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**15**
	1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6157
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.87
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.20
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.72
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.82
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.66
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.52
G	**16**	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.41
L	**17**	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.31
D	**18**	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.23
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.15
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.03
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	2.98
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.93
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.85
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.81
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.78
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.75
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.73
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.70
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.35
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.19

[589] En la tabla los grados de libertad del numerador y del denominador se expresan por las siglas GLN y GLD respectivamente

[590] Adaptado de *Distribución F 0.01*, recuperado de http://www.uaaan.mx/~jmelbos/tablas/distf.pdf, por J. A. Mellado Bosque, (s. f.).

Tabla A6. Tabla *F* con alfa = 0,01 y con grados de libertad del numerador[591] entre 20 y 120[592]

		GLN					
		20	**24**	**30**	**40**	**60**	**120**
	1	6209	6235	6261	6287	6313	6339
	2	99.45	99.46	99.47	99.47	99.48	99.49
	3	26.69	26.60	26.50	26.41	26.32	26.22
	4	14.02	13.93	13.84	13.75	13.65	13.56
	5	9.55	9.47	9.38	9.29	9.20	9.11
	6	7.40	7.31	7.23	7.14	7.06	6.97
	7	6.16	6.07	5.99	5.91	5.82	5.74
	8	5.36	5.28	5.20	5.12	5.03	4.95
	9	4.81	4.73	4.65	4.57	4.48	4.40
	10	4.41	4.33	4.25	4.17	4.08	4.00
	11	4.10	4.02	3.94	3.86	3.78	3.69
	12	3.86	3.78	3.70	3.62	3.54	3.45
	13	3.66	3.59	3.51	3.43	3.34	3.25
	14	3.51	3.43	3.35	3.27	3.18	3.09
	15	3.37	3.29	3.21	3.13	3.05	2.96
G	**16**	3.26	3.18	3.10	3.02	2.93	2.84
L	**17**	3.16	3.08	3.00	2.92	2.83	2.75
D	**18**	3.08	3.00	2.92	2.84	2.75	2.66
	19	3.00	2.92	2.84	2.76	2.67	2.58
	20	2.94	2.86	2.78	2.69	2.61	2.52
	21	2.88	2.80	2.72	2.64	2.55	2.46
	22	2.83	2.75	2.67	2.58	2.50	2.40
	23	2.78	2.70	2.62	2.54	2.45	2.35
	24	2.74	2.66	2.58	2.49	2.40	2.31
	25	2.70	2.62	2.54	2.45	2.36	2.27
	26	2.66	2.58	2.50	2.42	2.33	2.23
	27	2.63	2.55	2.47	2.38	2.29	2.20
	28	2.60	2.52	2.44	2.35	2.26	2.17
	29	2.57	2.49	2.41	2.33	2.23	2.14
	30	2.55	2.47	2.39	2.30	2.21	2.11
	40	2.37	2.29	2.20	2.11	2.02	1.92
	60	2.20	2.12	2.03	1.94	1.84	1.73
	120	2.03	1.95	1.86	1.76	1.66	1.53

[591] En la tabla los grados de libertad del numerador y del denominador se expresan por las siglas GLN y GLD respectivamente

[592] Adaptado de *Distribución F 0.01*, recuperado de http://www.uaaan.mx/~jmelbos/tablas/distf.pdf, por J. A. Mellado Bosque, (s. f.).

11. Bibliografía

Agencia Nacional de Noticias [Télam]. (2013). *Candidatos nacionales*. Recuperado de http://www.telam.com.ar/elecciones/candidatos

Alonso, C. (s. f.). *El modelo de regresión lineal múltiple*. Recuperado de http://www.eco.uc3m.es/docencia/econometria/NotasdeClase/Tema3.pdf

Anderson, D. R., Sweeney, D. J., y Williams, T. A. (2009). *Estadística para administración y economía* (10a ed.). México: Cengage Learning.

Bono, R., y Arnau, J. (1995). Consideraciones generales en torno a los estudios de potencia. *Anales de Psicología*, *11*(2), 193-202. Recuperado de http://www.um.es/analesps/v11/v11_2/08-11_2.pdf

comScore. (2012). *Argentina es el país más involucrado con las redes sociales a nivel global consumiendo cerca de 10 horas por visitante al Mes*. Recuperado de http://www.comscore.com/lat/Insights/Press_Releases/2012/12/Argentina_Ranks_First_in_Worldwide_Desktop_Social_Networking_Engagement

Cosenza, V. (2013). *World map of social networks*. Recuperado de http://vincos.it/world-map-of-social-networks/

Crettaz, J. (2013). *Me gusta Facebook: la mitad de los argentinos tiene cuenta.* Recuperado de http://www.lanacion.com.ar/1558643-tapa-me-gusta-facebook-la-mitad-de-los-argentinos-ya-tiene-una-cuenta

de Irala, J., Martínez-González, M.A., y Guillén Grima, F. (2001). ¿Qué es una variable de confusión?. *Medicina Clínica, 117*(10), 380-383. Recuperado de *(ver PDF ¿Qué es una variable de confusión? - ResearchGate)* http://www.google.com.ar/search?q=de+irala++ajustar+un+modelo+de+regre

sión+por+una+variable+confusora+&hl=es-AR&gbv=2&oq=de+irala++ajustar+un+modelo+de+regresión+por+una+variable+confusora+&gs_l=heirloom-serp.3...65354.65354.0.65764.1.1.0.0.0.0.0.0..0.0....0...1ac.1.34.heirloom-serp..1.0.0.BP3S3nZ0kOM

Erbin, A. (30 de junio de 2009). Políticos que usaron la web 2.0 en la campaña. [Mensaje de Blog]. Recuperado de http://webpoliticas.blogspot.com/2009/06/los-politicos-que-mas-usaron-la-web-20.html

Fernández-Ardáiz, J. (21 de mayo de 2010). Cómo medir el éxito de una campaña 2.0 [Mensaje de Blog]. Recuperado de http://politicaweb.cicoa.com.ar/noticia.php?tipo_id=1&id=165

Fernández-Ardáiz, J., y Doria, A. (2010). *Indice 2.0 Diputados.* Recuperado de http://www.youblisher.com/p/63630-Indice-2-0-Diputados-Nacionales-2010/

Fernández-Ardáiz, J., y Doria, A. (2011). *Indice 2.0 Senadores Nacionales 2011*. Recuperado de http://www.youblisher.com/p/218061-Indice-2-0-de-los-Senadores-Nacionales-de-Argentina-2011/

Pita Fernández, S. (1996). Determinación del tamaño muestral. *Cadernos de atención primaria*, *3*(3), 138-141. Recuperado de http://www.fisterra.com/mbe/investiga/9muestras/9muestras2.asp

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, M. (2010). *Metodología de la investigación* (5a ed.). Perú: McGraw-Hill.

iRedes. (1 de marzo de 2013). Presentación de la tercera versión del Mapa iRedes [Mensaje de Blog]. Recuperado de http://www.iredes.es/2013/03/tercera-version-del-mapa-iredes/

Levine, D. M., Krehbiel, T. C., y Berenson, M. L. (2006). *Estadística para administración* (4a ed.). México: Prentice Hall Inc.

Marín Diazaraque, J. M. (s. f.). *Transformaciones de variables.* Recuperado de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema4.pdf

Molinero, L. M. (2002). *Construcción de modelos de regresión multivariantes*. Recuperado de http://www.seh-lelha.org/regresion1.htm

Montero Lorenzo, J. M. (2007). *Estadística descriptiva*. Recuperado de http://books.google.com.ar/books?id=D6sj2d0xTgUC&printsec=copyright&hl=es&source=gbs_pub_info_r#v=snippet&q=10%2C57%25&f=false

Morales Vallejo, P. (2012). *El tamaño del efecto (effect size): análisis complementarios al contraste de medias.* Recuperado de http://www.upcomillas.es/personal/peter/investigacion/Tama%F1oDelEfecto.pdf

Morales Vallejo, P. (2012). *Correlación y regresión, simple y múltiple.* Recuperado de http://web.upcomillas.es/personal/peter/investigacion/Regresion.pdf

Nadal, H. (26 de octubre de 2011). Elecciones argentinas 2011: ¿Y dónde está la campaña 2.0? [Mensaje de Blog]. Recuperado de http://www.listao.com.ar/2011/10/elecciones-2-0-argentina-2011/

Osorio, M. (2007). *Diccionario de Ciencias Jurídicas, Políticas y Sociales.* Buenos Aires: Editorial Heliasta.

Pérez López, C. (2005). *Métodos estadísticos avanzados con SPSS.* Madrid: Thomson.

Preacher, K. J. (2003). *A primer on interaction effects in multiple linear regression.* Recuperado de http://www.quantpsy.org/interact/interactions.htm

Ramírez, D. C. (s. f.). *Autocorrelación*. Recuperado de http://webdelprofesor.ula.ve/economia/dramirez/MICRO/FORMATO_PDF/Materialeconometria/Autocorrelacion.pdf

Rodríguez, M. (2012). *Regresión lineal*. Recuperado de http://cristian415.wikispaces.com/file/view/5+-+Regresión+Lineal+-+Curvilinea+-+Multiple+(sin+PH).pdf

Sánchez Mangas, R. (s. f.). *Funciones de regresión no lineales (SW Cap. 6)*. Recuperado de http://www.uam.es/personal_pdi/economicas/rsmanga/docs/Econometria1-Transp-tema5-1.pdf

Sapiensman.com. (s. f.). *Ecuaciones cuadráticas con una variable*. Recuperado de http://www.sapiensman.com/matematicas/matematicas46.htm

Spiegel, M. R. (1989). *Estadística*. España: McGraw-Hill.

Soper, D. (2006). *Statistics calculators.* Recuperado de http://www.danielsoper.com/statcalc3/calc.aspx?id=1

Supo, J. (12 de agosto de 2011). Transformar variables en el programa SPSS [Archivo de video]. Recuperado de http://bioestadistico.com/transformar-variables-en-el-programa-spss

Stats Direct. (s. f.). *P values*. Recuperado de http://www.statsdirect.com/help/default.htm#basics/p_values.htm

Tacq, J. (1998). *Multivariate Analysis Techniques in Social Science Research*. Londres, Inglaterra: Sage Publications Ltd.

The University of Texas at Austin. (s. f.). *Principal component analysis: Validation, outliers, and reliability.* Recuperado de http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/SolvingHo

meworkProblems.htm

The University of Texas at Austin. (s. f.). *Illustration of regression analysis.* Recuperado de http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/_StartTopic_Illustration_of_Regression_Analysis.html

Todo Noticias [TN]. (2011). *Elecciones porteñas.* Recuperado de http://tn.com.ar/elecciones-2011/redes-sociales/candidatos.html

Todo Noticias [TN]. (2011). *Elecciones 2011.* Recuperado de http://tn.com.ar/elecciones-2011/presidenciales/

Universidad Complutense Madrid. (s. f.). *Análisis de regresión lineal: El procedimiento regresión lineal*. Recuperado de http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf

Universidad de Alicante. (s. f.). *La reexpresión de las variables (ver .doc La reexpresión de las variables – RUA).* Recuperado de http://www.google.com.ar/search?q=supuesto+linealidad+transformaciones+y&hl=es-AR&gbv=2&oq=supuesto+linealidad+transformaciones+y&gs_l=heirloom-serp.3...18282.19479.0.24431.9.7.0.0.0.1.280.550.2-2.2.0....0...1ac.1.24.heirloom-serp..9.0.0.xWZtOVBP9wo

Universidad de Antioquía. (s. f.). *Manual simplificado de estadística aplicada vía SPSS.* Recuperado de http://aprendeenlinea.udea.edu.co/revistas/index.php/ceo/article/viewFile/6556/6006

Universidad Rafael Urdaneta. (s. f.). *Análisis de regresión no lineal*. Recuperado de http://www.uru.edu/fondoeditorial/libros/pdf/manualdestatistix/cap9.pdf

University of Delaware. (s. f.). *Multiple regression with categorical data.* Recuperado de www.udel.edu/htr/Statistics/Notes816/class14.PDF

Varguillas Carmona, C. S., y Ribot de Flores, S. (2007). Implicaciones conceptuales y metodológicas en la aplicación de la entrevista en profundidad. *Laurus, 13*(23), 249-262. Recuperado de http://www.redalyc.org/pdf/761/76102313.pdf

VCE Further Maths. (24 de marzo de 2011). Maths tutorial: Question on data transformations (statistics) [Archivo de Video]. Recuperado de http://www.youtube.com/watch?v=EJ6EhfenqNs

Virginia Commonwealth University. (s. f.). *Multiple regression in SPSS.* Recuperado de http://www.or.vcu.edu/help/SPSS/SPSS.MultiReg.pdf

Wilderdom. (s. f.). *Tutorial 4 multiple linear regression*. Recuperado de http://wilderdom.com/courses/surveyresearch/tutorials/4/