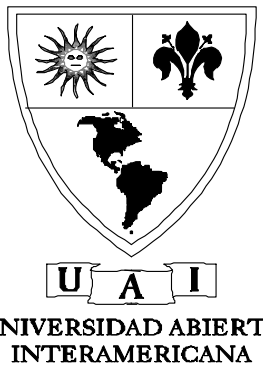
UNIVERSIDAD ABIERTA INTERAMERICANA
Facultad de Tecnología Informática



Carrera: Licenciatura en Matemática

REGRESIÓN LINEAL
Aplicación de la regresión lineal en un problema de pobreza.

Autor: Camila Jazmín Ageitos
Directora: Nora Sarti

TESIS PRESENTADA PARA OPTAR AL TÍTULO DE LICENCIADO EN MATEMÁTICA

Junio 2022

Índice

Capítulo 1

Introducción al análisis de regresión

El objetivo de esta tesis es realizar una lectura de los núcleos teóricos del artículo "*Aplicación de la regresión lineal en un problema de pobreza*", el cual realiza un modelo de regresión lineal de una situación particular en el campo de la economía, la pobreza, con datos de los años 2010 y 2011 de las trece principales ciudades de Colombia obtenidos del Departamento Administrativo Nacional de Estadística (DANE), profundizando en el estudio de residuos y observaciones influyentes, para luego replicar dicho análisis con datos de Argentina obtenidos de la encuesta EPH del Instituto Nacional de Estadísticas y Censos (INDEC) y realizar una conclusión con los valores obtenidos.

El procedimiento estadístico que se utiliza en el paper es el análisis de regresión. En el mismo hacen referencia a que el término regresión fue utilizado por primera vez como un concepto estadístico en 1877 por sir Francis Galton.

A continuación, se presenta un breve resumen del contexto histórico en el cual aparece el concepto de regresión.

Galton (1822-1911) estudió Medicina y Matemáticas en Londres y Cambridge, tenía una destacada facilidad para construir artificios mecánicos, habilidad que utilizó para construir sofisticados aparatos de medida. Tuvo obsesivo interés por medir, hacer recuentos y gráficos de fenómenos de antropología, biología, genética, sociología y psicología. A lo largo del periodo 1865-1890, su principal interés fueron los estudios empíricos de las leyes de la herencia por medio de métodos estadísticos.

Darwin, primo de la esposa de Galton, le propuso estudiar algún método que pudiese dar soporte a su teoría de la evolución, tratando de comparar las características físicas de los hijos con las de sus padres, pues si estos caracteres se heredan se confirmaría su hipótesis de que las características de los sujetos mejor adaptados pasarían de una a otra generación. Galton acepta el reto, pero, al no tener suficientes datos humanos diseñó un experimento con semillas de guisantes para estudiar la distribución de los pesos de las semillas en dos generaciones. Observó que la distribución de los pesos era normal, seleccionó siete grupos, conteniendo cada grupo 70 semillas del mismo peso. Pidió a siete amigos de diferentes partes del país que cultivaran un grupo de semillas y que le enviaran las semillas cosechadas. Sus conclusiones fueron:

- El peso medio de las semillas hijas era función lineal del peso de las semillas padres con una pendiente menor que la unidad, es decir, el peso medio de las semillas cosechadas se desvía menos de la población media que de los padres. Por tanto, los padres de peso $M + x$ producen hijos adultos de peso medio $M + r.x$, para $0 < r < 1$.
- Para cada grupo de semillas padres, el peso de las semillas cosechadas estaba normalmente distribuido. El peso de los hijos, debido a la variación aleatoria de entre hijos del mismo grupo de padres, llega a ser $M + r.x + y$.
- La desviación probable del peso de las semillas cosechadas es la misma para todos los grupos y más pequeña que la desviación probable del peso de las semillas

padres. Esta es la propiedad que hoy día se conoce como homocedasticidad u homogeneidad de las varianzas.

Galton continuó realizando estudios para concluir que la estatura de los hijos regresa hacia la media de la población y de ahí surge el término *regresión* que, desde entonces, se utiliza para designar cualquier relación estadística y a la mencionada recta de regresión que más se ajusta a una distribución otorgada.

El paper menciona la terminología de la regresión donde la variable que se va a predecir se llama dependiente, a explicar, o endógena; y las variables que se usan para predecir el valor de la variable dependiente se llaman independientes, explicativas o exógenas. Y remarca que, en general, existen cuatro posibles formas en que las variables se pueden relacionar: Relación lineal directa, relación lineal inversa, relación no lineal directa y relación no lineal inversa. De acuerdo con la estructura formal y funcional de la relación se puede decidir qué ecuación se debe emplear, cuál ha de ser la ecuación que mejor se ajusta a los datos y cómo debe validarse la significancia estadística de los pronósticos realizados.

En la aplicación del modelo de regresión lineal que desarrolla el paper, utiliza los datos de una muestra real, extraídos de un comunicado de prensa del Departamento Administrativo Nacional de Estadística (DANE) que revela el porcentaje de pobreza, pobreza extrema y el coeficiente de Gini (indicador de la desigualdad económica en una población) en los años 2010 y 2011 de trece de las principales ciudades de Colombia.

En el caso de argentina el instituto encargado de recolectar los datos es el INDEC (Instituto Nacional de Estadísticas y Censos), la encuesta que mide la pobreza tanto de las personas como de los hogares es la Encuesta Permanente de Hogares (EPH) la cual comprende una extensa muestra de hogares tomada en los principales aglomerados urbanos del país. En su contenido brinda datos sobre la situación laboral y social de la población urbana de la Argentina. Para realizar esta encuesta, trabajan 250 encuestadores previamente capacitados especialmente para realizar esa tarea a través de visitas personales en cada una de las viviendas seleccionadas en la muestra. Se utilizan tres cuestionarios: uno para obtener datos de la vivienda, otro sobre el hogar y otro para cada uno de los individuos de 10 años y más. Para este estudio se tomarán los datos recolectados del cuestionario individual el cual permite relevar atributos de las personas en cuanto a sus características ocupacionales y de ingreso. Una vez obtenida esta información, se compila la totalidad de los datos y difunde los resultados a nivel de "Total de Aglomerados Urbanos" y por cada uno de los "Aglomerados". Estos resultados se publican trimestralmente en informes de prensa y base de microdatos en la página web del INDEC.

Antes de desarrollar como se analizan los datos obtenidos de la encuesta mencionada anteriormente, es importante entender el concepto de pobreza, el paper no hace mención de este concepto, por este motivo brevemente se presentará la definición. El término pobreza hace referencia, en líneas generales a carencias o privaciones, refiere a la incapacidad de una persona para alcanzar un mínimo nivel de vida, un umbral conocido como línea de pobreza. Se pueden definir tres tipos de pobreza de acuerdo a las variables que se tomen para analizarla, pobreza de ingresos (se mide el nivel de vida a partir de una sola variable monetaria), pobreza multidimensional (se mide la combinación de variables que capten distintos aspectos del nivel de vida de la persona y su acceso a bienes, servicios y derechos) y pobreza crónica (condiciones de vida permanentemente bajas, carencias persistentes que no pueden ser superadas, incluso en contextos de alto empleo y mayor prosperidad económica general).

El caso de estudio refiere a la pobreza de ingresos, una persona se considera pobre si su ingreso no supera un determinado valor monetario conocido como línea de pobreza o indigencia. El ingreso relevante no es el propio, sino el ingreso total del hogar al que pertenece la persona en cuestión, dividido por algún factor que capte la estructura demográfica del hogar. En el caso argentino se divide al ingreso total familiar por la suma de adultos equivalentes del hogar. Esta metodología de estimación de pobreza e indigencia que aplica INDEC se basa en el "método de la línea": la situación de indigencia de cada hogar se determina comparando sus ingresos con el valor monetario de la Canasta Básica Alimentaria (CBA) que le corresponde de acuerdo con su composición demográfica. La CBA establece el costo mínimo de adquirir los alimentos que sirven al hogar para satisfacer un umbral mínimo de necesidades energéticas y proteicas. Los hogares con ingresos inferiores a ese valor monetario (Línea de indigencia) son identificados como indigentes, y la tasa de indigencia es el porcentaje de hogares/individuos identificados como indigentes entre el total de hogares/individuos en la población.

La Canasta Básica Alimentaria, en cantidades, que se tomó para definir la línea de pobreza que se utilizó para procesar los datos que aplicaremos a la regresión de este artículo es la siguiente:

Componente	Unidades	Productos que se incluyen
Pan	6.750 g	
Galletitas de agua	420 g	
Galletitas dulces	210 g	
Arroz	1.200 g	
Harina de trigo	1.080 g	
Otras harinas (maíz)	210 g	
Fideos	1.740 g	
Papa	6.510 g	
Batata	510 g	
Azúcar	1.230 g	
Dulces	330 g	Dulce de batata, mermelada, dulce de leche
Legumbres secas	240 g	Lentejas, arvejas
Hortalizas	5.730 g	Acelga, cebolla, lechuga, tomate perita, zanahoria, zapallo, tomate envasado
Frutas	4.950 g	Manzana, mandarina, naranja, banana, pera
Carnes	6.270 g	Asado, carnaza común, espinazo, paleta, carne picada, nalga, pollo, carne de pescado
Menudencias	270 g	Hígado
Fiambres	60 g	Paleta cocida, salame
Huevos	600 g	
Leche	9.270 g	
Queso	330 g	Queso crema, queso cuartirolo, queso de rallar
Yogur	570 g	
Manteca	60 g	
Aceite	1.200 g	
Bebidas no alcohólicas	3.450 cc	Gaseosas, jugos concentrados, soda
Bebidas alcohólicas	1.080 cc	Cerveza, vino
Sal fina	120 g	
Condimentos	120 g	Mayonesa, caldos concentrados
Vinagre	60 g	
Café	30 g	
Yerba	510 g	

Fuente: INDEC. "Actualización de la metodología oficial de cálculo de las líneas de pobreza". Documento en discusión. Argentina, 2004.

Por otro lado, se calcula además la Canasta Básica Total (CBT) usada para estimar la tasa de pobreza que se obtiene a partir de la CBA, agregando a la misma el valor de los bienes y servicios no alimentarios. La forma precisa en que el valor de la CBA se amplía para

obtener el valor de la CBT es multiplicándolo por la inversa del coeficiente de Engel, siendo dicho coeficiente la proporción del gasto total que la población de referencia destina al gasto alimentario. En la práctica no existe una canasta donde se especifiquen componentes no alimentarios y sus cantidades. En este sentido, el componente no alimentario de la CBT es más empírico que normativo.

En el capítulo 4 se realizará un análisis de regresión con los datos de pobreza del segundo semestre del 2018 contra los del segundo semestre del 2019. Si bien los porcentajes de pobreza ya se encuentran calculados por el INDEC, es interesante conocer los valores de la Canasta Básica Alimentaria y Canasta Básica Total utilizados para el cálculo de dicho porcentaje. A continuación, se presenta la tabla con los valores mensuales de la CBA y CBT por adulto.

Mes	Canasta básica alimentaria Línea de indigencia (Valor en $)	Inversa del coeficiente de Engel	Canasta básica total Línea de pobreza (Valor en $)
jun-18	2.537,45	2,50	6.343,62
jul-18	2.627,37	2,48	6.515,88
ago-18	2.701,48	2,50	6.753,70
sep-18	2.931,88	2,49	7.300,38
oct-18	3.150,62	2,49	7.845,04
nov-18	3.276,02	2,49	8.157,29
dic-18	3.300,17	2,50	8.250,42
jun-19	4.016,09	2,51	10.080,39
jul-19	4.133,91	2,50	10.334,77
ago-19	4.290,72	2,49	10.683,89
sep-19	4.502,88	2,50	11.257,20
oct-19	4.596,20	2,51	11.536,46
nov-19	4.886,34	2,49	12.166,99
dic-19	5.043,41	2,50	12.608,52

Fuente: INDEC

Capítulo 2

Regresión lineal simple

En este capítulo se leerán los conceptos desarrollados en el paper y se extenderán algunas definiciones en caso de ser necesario.

En primer lugar, en el paper se realiza una gráfica de dispersión de los datos observados. La misma sugiere que existe una relación lineal entre la variable independiente porcentaje de pobreza en 2010 y la variable dependiente porcentaje de pobreza en 2011, es por ello por lo que utiliza el modelo de regresión lineal simple.

Para ejemplificar se pueden observar algunas gráficas de dispersión en la siguiente figura:

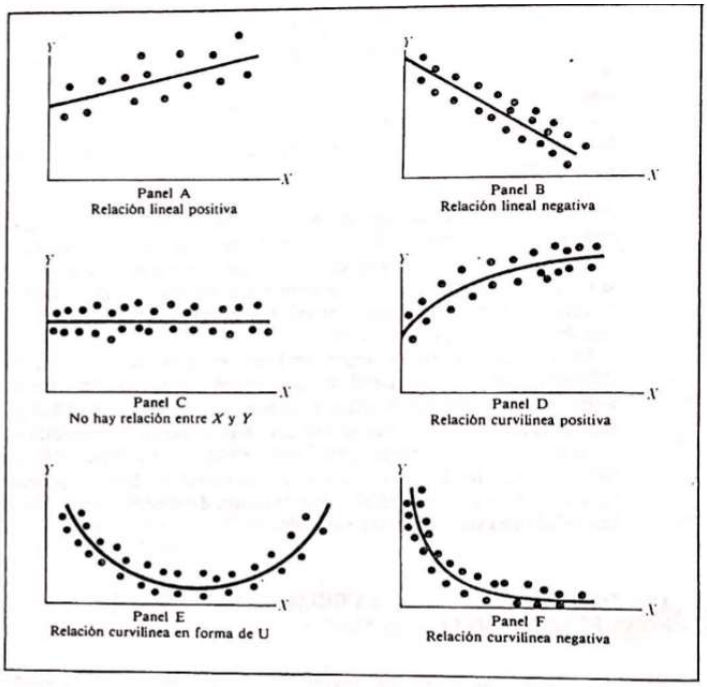


Figura 2.1. Gráficas de dispersión

Todos estos paneles hacen referencia a modelos diferentes que se podrían utilizar para representar la relación entre dos variables. Hay ocasiones en las que la relación que puede darse entre variables independientes y la variable dependiente no tenga un desarrollo lineal, sino que tengan, por ejemplo, un crecimiento exponencial.

Para el análisis de pobreza de las personas en Argentina, tema central de este documento, trabajaremos también sobre el modelo de regresión lineal simple. Se presenta la definición a continuación

***Definición 2.1**: Llamamos **Modelo de regresión lineal simple** a la ecuación*

$$\boldsymbol{Y = \beta_0 + \beta_1 x + \varepsilon}$$

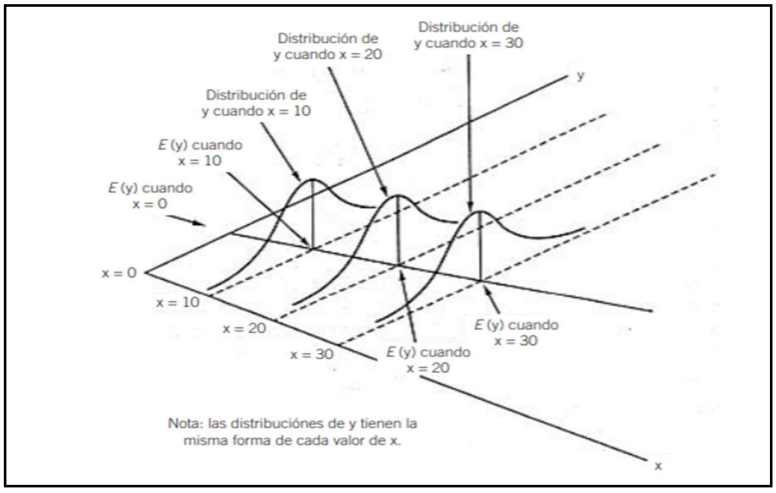
Donde β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y ε es una variable aleatoria distribuida con

$$E(\varepsilon) = 0 \; y \; Var(\varepsilon) = \sigma^2$$

Como se puede observar en la definición Y es una variable aleatoria, ya que ε es aleatoria, por el contrario, el valor x no es aleatorio.

En el paper se desarrolla que ε representa el término de error y explica la variabilidad en Y que no se puede explicar con la relación lineal, y menciona que a este término se le asocian los siguientes supuestos:

1. El término de error es una variable aleatoria con media o valor esperado igual a cero; $E(\varepsilon) = 0$
2. La varianza de ε, representada por σ^2, es igual para todos los valores de x. Esto implica que la varianza de Y es igual a σ^2 y es la misma para todos los valores de x.
3. Los valores de ε son independientes. El valor de ε para un determinado valor de x no se relaciona con el valor de ε para cualquier otro valor de x; así, el valor de Y para determinado valor de x no se relaciona con el valor de Y para cualquier otro valor de x
4. El término de error, ε , es una variable aleatoria con distribución normal.



Fuente: Sweeney y Williams, 2001.

Figura 2.2. Supuestos del modelo y sus implicaciones

El hecho de que $E(\varepsilon) = 0$ implica que, para una x específica, los valores de Y se distribuyen alrededor de la recta verdadera o recta de regresión de la población $Y = \beta_0 + \beta_1 x$

En la figura 2.3 se ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que sólo se dispone de $n = 5$ observaciones.

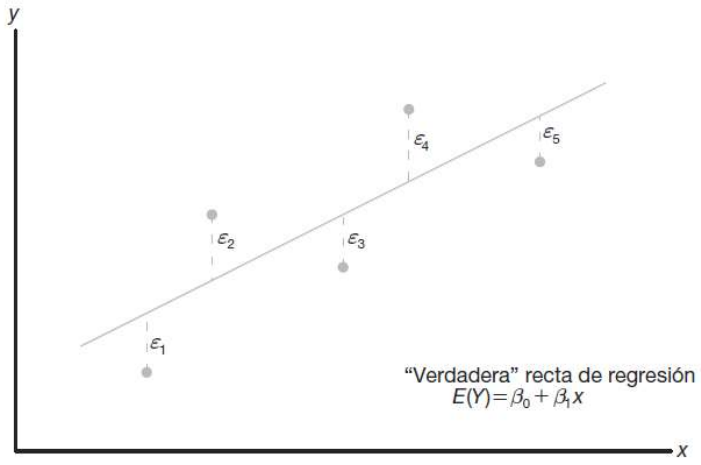


Figura 2.3: verdadera recta de regresión para $n = 5$

Tal como se hizo mención en la definición de modelo de regresión lineal simple, β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, el paper agrega que los mismos no se conocen y deben estimarse a partir de los datos de la muestra. Estos coeficientes que se calculan de la muestra son conocidos como regresores (b_0 y b_1).

Entonces, la recta de regresión estimada, o ajustada, es dada por:

$$\hat{y} = b_0 + b_1 x$$

donde $\hat{y}$ es el valor pronosticado o ajustado.

Es evidente que la recta ajustada es un estimado de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se dispone de una gran cantidad de datos. Para que la línea estimada de regresión se ajuste bien a los datos se desea que las diferencias entre los valores observados de y (y_i) y los valores estimados de y ($\hat{y}$) sean mínimas.

2.1. Método de mínimos cuadrados

El paper plantea que para calcular el valor de los regresores (b_0 y b_1) se utiliza el método de los mínimos cuadrados el cual emplea los datos de la muestra para determinar las características de la recta que hacen mínima la suma de los cuadrados de las desviaciones. Matemáticamente se minimiza

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

En donde y_i*= valor real de Y para la observación* i
$\hat{y}_i$*=valor predicho de Y para la observación* i

Si bien el paper no lo menciona en este punto, antes de desarrollar este método, es importante presentar el concepto de residual. En esencia, un residual es un error en el ajuste del modelo $\hat{y} = b_0 + b_1 x$.

Definición 2.2**: Dado un conjunto de datos de una regresión* $\{(x_i, y_i);\ i = 1,2,\ldots,n\}$ *y un modelo ajustado* $\hat{y} = b_0 + b_1 x$*, el i-ésimo* ***residual e_i *es dado por*

$$\boldsymbol{e_i = y_i - \hat{y}_i}$$

Con $i = 1,2,\ldots,n$.

Es evidente que, si un conjunto de n residuales es grande, entonces el ajuste del modelo no es bueno. Los residuales pequeños son indicadores de un ajuste adecuado. Por otro lado, mientras que los ϵ_i no se observan, los e_i no sólo se observan, sino que desempeñan un papel importante en el análisis total que se verá en el capítulo 3.

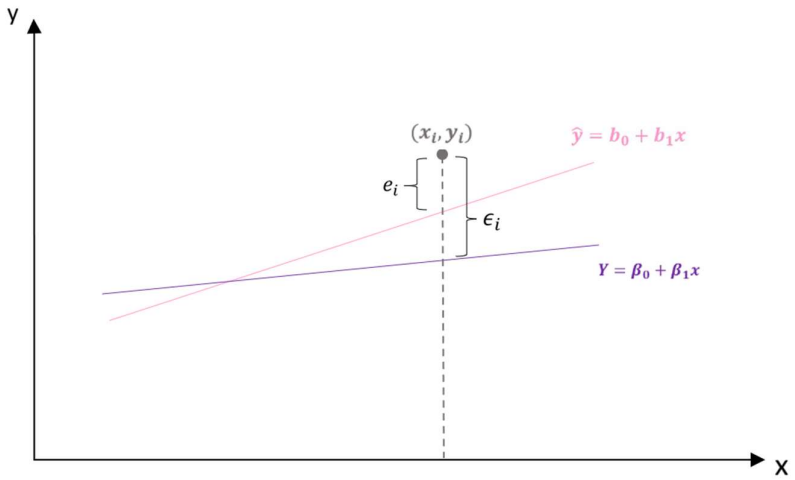


Figura 2.4.: Comparación de ϵ_i *con el residual* e_i

Ahora sí, se debe calcular b_0 y b_1, los estimados de β_0 *y* β_1, de manera que la suma de los cuadrados de los residuales sea mínima. La suma residual de los cuadrados con frecuencia se denomina suma de los cuadrados debidos al error y se denota como SSE. Cualquier valor para b_0 y b_1 que no sea el determinado por el método de mínimos cuadrados, darán por resultado una suma más grande de diferencias al cuadrado entre el valor real de Y y el valor predicho de Y.

Por lo tanto,

$$\boldsymbol{SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2}$$

El paper aclara que para minimizar SSE implica calcular las derivadas parciales de la expresión con respecto a los coeficientes de regresión e igualar a cero las dos derivadas y que al finalizar este procedimiento se llega a las siguientes ecuaciones, conocidas como ecuaciones normales:

$$\sum_{i=1}^{n} \boldsymbol{y_i} = \boldsymbol{nb_0} + \boldsymbol{b_1} \sum_{i=1}^{n} \boldsymbol{x_i} \qquad [\mathbf{1}]$$

$$\sum_{i=1}^{n} \boldsymbol{y_i x_i} = \boldsymbol{b_0} \sum_{i=1}^{n} \boldsymbol{x_i} + \boldsymbol{b_1} \sum_{i=1}^{n} \boldsymbol{x_i}^2 \qquad [\mathbf{2}]$$

Donde n es el número de observaciones.

A continuación, se realizan los cálculos mencionados anteriormente para obtener las Ecuaciones Normales

Partiendo de

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Al diferenciar la ecuación con respecto a b_0 y b_1, se obtiene

$$\frac{\partial (SSE)}{\partial b_0} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)^1 (-1) = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial (SSE)}{\partial b_1} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)^1 (-x_i) = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las ecuaciones normales

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 x_i = \frac{0}{-2}$$

$$\sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} \boldsymbol{y_i} = \boldsymbol{nb_0} + \boldsymbol{b_1} \sum_{i=1}^{n} \boldsymbol{x_i} \qquad [\mathbf{1}]$$

Luego,

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} (y_i x_i - b_0 x_i - b_1 x_i^2) = \frac{0}{-2}$$

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} b_0\, x_i - \sum_{i=1}^{n} b_1 {x_i}^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\boldsymbol{\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2} \qquad [\mathbf{2}]$$

Al resolver algebraicamente el sistema de ecuaciones normales se obtienen las soluciones para b_0 y b_1.

Teorema 2.1. *Dadas las ecuaciones normales*

$$\boldsymbol{\sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i} \qquad [\mathbf{1}]$$

$$\boldsymbol{\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2} \qquad [\mathbf{2}]$$

Para la muestra $\{(x_i, y_i)\}$; $i = 1,2,\dots,n$, *los estimados* b_0 *y* b_1 *de los mínimos cuadrados de los coeficientes de regresión* β_0 *y* β_1 *se calculan mediante las fórmulas*

$$\boldsymbol{b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$\boldsymbol{b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}}$$

Demostración 2.1:

Sea la ecuación normal obtenida del método de mínimos cuadrados

$$\sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i \qquad [1]$$

Despejando b_0

$$\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i = n b_0$$

$$\frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = b_0 \qquad [3]$$

Obsérvese que también se puede escribir como

$$\frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} = b_0$$

$$\bar{Y} - b_1 \bar{X} = b_0$$

Luego, reemplazando [3] en [2]

$$\sum_{i=1}^{n} y_i x_i = \left(\frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}\right) \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2$$

$$\sum_{i=1}^{n} y_i x_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - \frac{b_1 \sum_{i=1}^{n} x_i}{n}\right) \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2$$

$$\sum_{i=1}^{n} y_i x_i = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} - \frac{b_1 (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i)}{n} + b_1 \sum_{i=1}^{n} {x_i}^2$$

Multiplicando a ambos lados de la ecuación por $\frac{\boldsymbol{n}}{\boldsymbol{n}}$ *lo cual no cambia la igualdad se obtiene*

$$\frac{\boldsymbol{n}}{\boldsymbol{n}} \left(\sum_{i=1}^{n} y_i x_i\right) = \left(\frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} - \frac{b_1 (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i)}{n} + b_1 \sum_{i=1}^{n} {x_i}^2\right) \frac{\boldsymbol{n}}{\boldsymbol{n}}$$

$$\frac{n}{n} \left(\sum_{i=1}^{n} y_i x_i\right) = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} - \frac{b_1 (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i)}{n} + \frac{n}{n} b_1 \sum_{i=1}^{n} {x_i}^2$$

$$\frac{n}{n} \sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} = b_1 \left(-\frac{(\sum_{i=1}^{n} x_i)^2}{n} + \frac{n}{n} \sum_{i=1}^{n} {x_i}^2\right)$$

$$\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} = b_1 \left(\frac{-(\sum_{i=1}^{n} x_i)^2}{n} + \frac{n \sum_{i=1}^{n} {x_i}^2}{n}\right)$$

$$\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} = b_1 \left(\frac{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}{n}\right)$$

$$\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{\cancel{n}} * \frac{\cancel{n}}{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} = b_1$$

$$\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n \sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} = b_1$$

En este caso, se puede escribir también de la siguiente manera

Multiplicando a ambos lados de la igualdad por $\frac{n}{n}$

$$\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2} \frac{n}{n} = b_1$$

$$\frac{\frac{n(\sum_{i=1}^{n} y_i x_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\frac{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}{n}} = b_1$$

$$\frac{\frac{n(\sum_{i=1}^{n} y_i x_i)}{n} - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\frac{n\sum_{i=1}^{n} {x_i}^2}{n} - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = b_1$$

$$\frac{(\sum_{i=1}^{n} y_i x_i) - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = b_1$$

Fin de la demostración 2.1.

Como resultado, el procedimiento de mínimos cuadrados genera una recta que minimiza la suma de los cuadrados de las desviaciones verticales desde los puntos hasta la recta. El coeficiente b_1 también corresponde a la pendiente de la recta. En general, este coeficiente expresa la razón de cambio entre la variable dependiente con respecto a un cambio unitario en la variable independiente x.

2.2. Coeficientes de correlación y determinación

Siguiendo con el orden que plantea el paper, luego es necesario determinar la pertinencia de la ecuación de regresión hallada mediante un análisis de la bondad de ajuste de la recta, demostrar si la relación es estadísticamente significativa y validar los supuestos acerca del término de error.

El primer paso para validar esta recta es calcular el coeficiente de determinación.

Coeficiente de determinación

El paper plantea al coeficiente de determinación como una medida de la bondad de ajuste para una ecuación de regresión y define en este punto al el i-ésimo residual y la suma de los cuadrados debidos al error mencionada anteriormente cuando se demostró el método de los mínimos cuadrados.

Define a la SSE como una medida del error que se comete al usar la ecuación de regresión para calcular los valores de la variable dependiente en la muestra. Y agrega otras dos sumas de cuadrados que serán útiles para el análisis de regresión: SST y SSR.

La suma total de cuadrados (SST) representa desviaciones de los valores de y_i de los valores de la media $\overline{Y}$:

$$\boldsymbol{SST = \sum_{i=1}^{n} (y_i - \overline{Y})^2}$$

Y la suma de cuadrados debida a la regresión (SSR) que representa la diferencia entre $\overline{Y}$ (el valor promedio de y) y $\hat{y}_i$ (el valor de y que se predeciría con la relación de regresión). La misma sirve para saber cuánto se desvían los valores de $\hat{Y}_i$ medidos en la línea de regresión de los valores de $\overline{Y}$:

$$\boldsymbol{SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{Y})^2}$$

Para entender mejor los conceptos se muestra a continuación la gráfica de las distintas sumas de cuadrados.

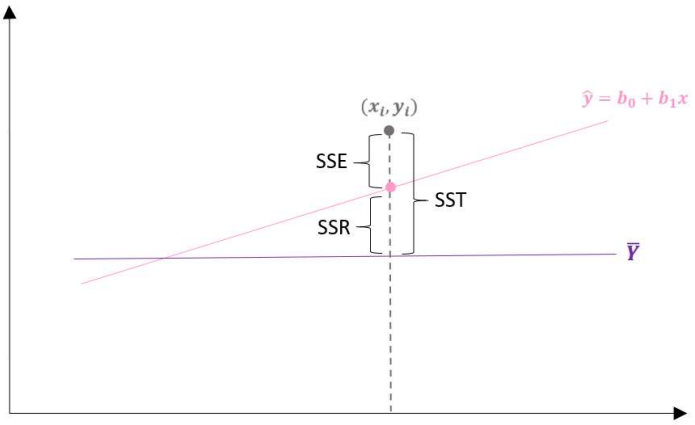


Figura 2.5. Sumas de cuadrados

Como se puede observar en la figura 2.5, existe una relación entre las tres sumas y es la siguiente:

$$\boldsymbol{SST = SSR + SSE}$$

Luego, tal como el paper menciona, esa ecuación tendría un ajuste perfecto si cada valor observado de la variable independiente estuviera sobre la línea de regresión.

Teorema 2.2

Sea la recta de regresión ajustada $\hat{y} = b_0 + b_1 x$, *con* $y_i = \hat{y}_i$ *para la muestra* $\{(x_i, y_i)\}$; $i = 1,2, \dots, n$, *entonces el máximo ajuste será*

$$\frac{SSR}{SST} = 1$$

Demostración 2.2.

Dada la recta de regresión ajustada $\hat{y} = b_0 + b_1 x$, *si* $y_i = \hat{y}_i$ *para todas las observaciones, entonces*

$$e_i = y_i - \hat{y}_i = y_i - y_i = 0$$

Luego,

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (0)^2 = 0$$

Por consiguiente

$$SST = SSR + SSE = SSR + 0 \Longrightarrow SST = SSR$$

Entonces, el máximo ajuste será

$$\frac{SSR}{SST} = 1$$

Fin de la demostración 2.2

De la demostración 2.2 se deduce que el máximo valor de SSE se tiene cuando SSR es cero.

Ahora sí, luego de las definiciones previas, la relación SSR/SST, que asume valores entre cero y uno, se le llama ***coeficiente de determinación*** y se representa por r^2.

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

La confiabilidad de r^2 depende del tamaño del conjunto de los datos de la regresión y del tipo de aplicación. Como se demostró en la demostración 2.2, el valor del coeficiente esta entre $0 \leq r^2 \leq 1$, y el límite superior se logra cuando el ajuste a los datos es perfecto, es decir, cuando todos los residuales son cero.

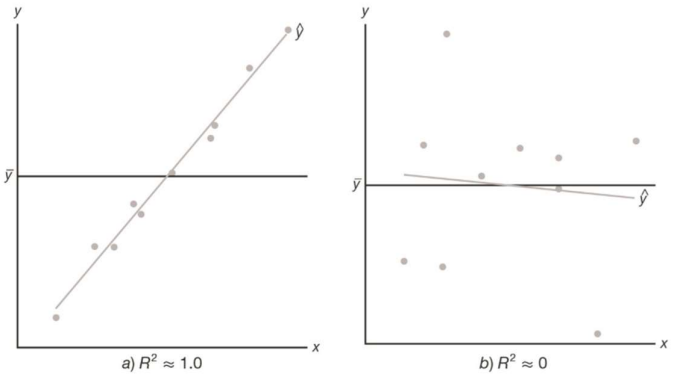


Figura 2.6. Gráficas que muestran un modelo muy bueno y otro deficiente

Además, Expresando este valor como un porcentaje, se puede interpretar a r^2 como el porcentaje de la variación de los valores de la variable dependiente que se puede explicar con la ecuación de regresión. (Levin & Rubin, 2004)

Coeficiente de correlación

La segunda medida de bondad de ajuste que utiliza el paper es el coeficiente de correlación, que describe qué tan bien explica una variable a la otra. El mismo se denota por $\boldsymbol{r}$ y es la raíz cuadrada del coeficiente de determinación:

$$\boldsymbol{r = (signo\ de\ b_1)\sqrt{(r)^2}}$$

El signo del coeficiente indica si la relación es directa o inversa.

En el caso de una relación lineal entre dos variables, el coeficiente de determinación y el de correlación permiten tener medidas de la intensidad de la relación. El coeficiente de determinación da una medida entre 0 y 1, mientras que el coeficiente de correlación da una medida entre -1 y 1.

El paper resalta que el coeficiente de correlación solo mide la fuerza de asociación en una relación lineal, el coeficiente de determinación se puede usar en relaciones no lineales (obviamente, teniendo como ecuación de regresión una función no lineal) y en relaciones con dos o más variables independientes. En este sentido, el coeficiente de determinación tiene mayor aplicabilidad.

Para entender mejor este coeficiente, se muestra a continuación la figura 2.7 con tres ejemplos de gráficas con el mismo $r = 0{,}75$. Lo cual indicaría una asociación fuerte entre las variables x e y.

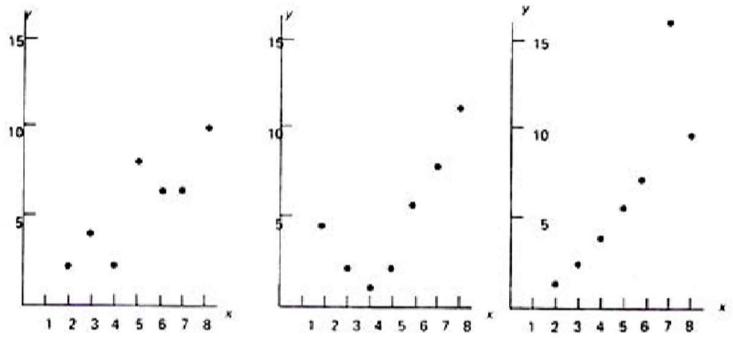


Figura 2.7. Errores en la interpretación de r

Sin embargo, es evidente que, ésta es una medida significativa de la fuerza de la relación solo en el primer caso. En el segundo hay una relación curvilínea muy evidente entre las dos variables y en el tercer caso, seis de los siete puntos en realidad caen en la línea recta, pero el séptimo punto está tan alejado que sugiere la posibilidad de un grave error de cálculo o un error en el registro de los datos. Así, antes de calcular r se deben graficar los datos para verificar si hay algún motivo para pensar que la relación es, de hecho, lineal.

Por último, el paper concluye que ambos coeficientes (determinación y correlación) no son suficientes para asumir si la relación es estadísticamente significativa. Esa conclusión se debe basar en consideraciones donde intervenga el tamaño de la muestra y las propiedades de las distribuciones muestrales adecuadas de los estimadores de los mínimos cuadrados.

2.3. Pruebas de significancia

Para estas pruebas de significancia se necesita un estimado de la varianza del error en el modelo de regresión y desviación estándar de la estimación.

Como menciona el paper, la varianza de ϵ, también representa la varianza de los valores de y respecto a la línea de regresión. Así, la suma de los residuales al cuadrado, SSE, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión. Cada suma de cuadrados tiene asociado un número que llamamos grados de libertad. Se ha demostrado que SSE tiene $n-2$ grados de libertad, porque se deben estimar dos parámetros β_0 y β_1.

El error cuadrado medio (s^2) es el estimado de σ^2. Se calcula mediante la ecuación:

$$\boldsymbol{s^2 = \frac{SSE}{n-2}}$$

Así, el error típico o ***desviación estándar del estimado*** se calcula como la raíz cuadrada de la varianza del estimado:

$$\boldsymbol{S = \sqrt{\frac{SSE}{n-2}}}$$

Así como la desviación estándar mide la variabilidad en torno a la media aritmética, el error estándar de estimación mide la variabilidad en torno a la recta ajustada de regresión.

Ahora sí con estas dos definiciones previas se pueden realizar las pruebas de significancia mencionadas anteriormente. El paper no desarrolla los pasos para realizarlas, los mismos se detallan a continuación.

Prueba t

El objetivo de esta prueba es determinar si el valor de β_1 es igual a cero, si esto sucede la ecuación de regresión será

$$\mathrm{Y} = \beta_0 + \beta_1 x \implies \mathrm{Y} = \beta_0 + 0x \implies \mathrm{Y} = \beta_0$$

En este caso el valor Y no depende del valor de x, y se concluye que no existe relación lineal entre las variables. En forma análoga, si el valor de β_1 no es igual a cero, se concluye que las dos variables se relacionan.

Las propiedades de la distribución muestral de b_1, el estimador de β_1 por mínimos cuadrados, son la base de esta prueba de hipótesis:

- Valor esperado: $\boldsymbol{E(b_1) = \beta_1}$
- Desviación estándar estimada:

$$\boldsymbol{S_b = \frac{S}{\sqrt{S_{XX}}}}$$

Siendo

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum {x_i}^2 - \frac{(\sum x_i)^2}{n}$$

- Forma de distribución: Normal

Por lo tanto, se usan los datos de la muestra para plantear las siguientes hipótesis:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Se rechaza H_0 si $t < -t_{\alpha/2}$ o si $t > t_{\alpha/2}$, siendo t el estadístico

$$\boldsymbol{t = \frac{b_1 - \beta_1}{S_b}}$$

con una distribución t con $n - 2$ grados de libertad, S_b la desviación estándar estimada de la distribución de b_1.

La conclusión será que $\beta_1 \neq 0$ y que hay una relación estadísticamente significativa entre las dos variables.

La distribución de t es simétrica alrededor de una media de cero y tienen forma de campana como se puede observar en la figura 2.8

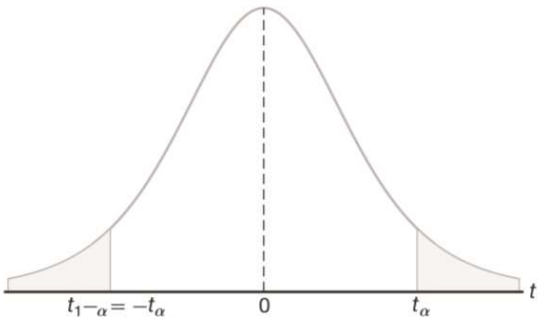


Figura 2.8. Propiedad de simetría (alrededor de 0) de la distribución t.

Prueba F

También se puede usar una prueba basada en la distribución F de probabilidades, para probar si la regresión es significativa. Como solo hay una variable independiente, la prueba F debe indicar la misma conclusión que la prueba t, pero cuando hay más de una variable independiente solo se puede usar la prueba F.

El objetivo de la prueba también es ver si se puede concluir que $\beta_1 \neq 0$. Para esto se plantean las hipótesis:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

El estadístico a tener en cuenta en este caso es F, con

$$\boldsymbol{F = \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}} = \frac{SSR}{S^2}}$$

Y se rechaza H_0 si $F > F_\alpha$

Donde F_α se basa en una distribución F con un grado de libertad en el numerador y $n-2$ grados de libertad en el denominador.

Cuando se rechaza la hipótesis nula, es decir, cuando el estadístico F calculado excede al valor crítico $F_\alpha\ (1, n-2)$, concluimos que hay una cantidad significativa de variación en la respuesta justificada por el modelo postulado, que es la función de la línea recta. Si el estadístico F está en la región de no rechazo, se concluye que los datos no reflejan evidencia suficiente para apoyar el modelo que se postula.

Capítulo 3

Residuales y Estadística inferencial

Luego de comentar acerca de las pruebas t y F el paper plantea el análisis de residuales. Para realizarlo comenta la existencia de las tres gráficas principales de los residuales y analiza los resultados obtenidos sin entrar en detalle de como manipular los datos para poder obtener estas graficas. Es por esto por lo que a continuación se analizaran en detalle los residuales.

3.1. Análisis de residuos

Las técnicas del diagnóstico en regresión se abocan a validar que los supuestos realizados por el modelo sean apropiados para los datos con los que se cuenta. Como menciona el paper, constan principalmente de técnicas gráficas, aunque también en la exhibición de algunas medidas de bondad de ajuste. Si el modelo propuesto, no proporciona residuos que parezcan razonables, entonces comenzamos a dudar de que algún aspecto del modelo sea apropiado para nuestros datos. Un tema relacionado es asegurarse que la estimación realizada no sea dependiente de un sólo dato (o un pequeño subconjunto de datos) en el sentido en que si no se contara con dicho dato las conclusiones del estudio serían completamente diferentes. La identificación de estos puntos influyentes forma parte relevante del diagnóstico, es por esto que sumaremos al análisis de residuales un análisis de outliers y observaciones influyentes.

Como se mencionó en el capítulo 2 los residuales ($e_i = y_i - \hat{y}_i$) corresponden al error en el ajuste de la recta de regresión, por lo que sirven para imitar a las ε_i. De manera ideal, los residuales deberían demostrar fluctuaciones aleatorias alrededor del valor de cero.

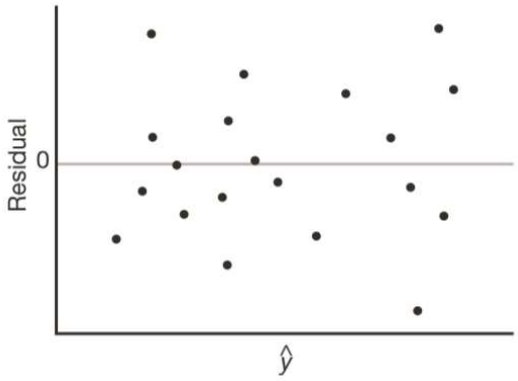


Figura 3.1: Gráfica ideal de los residuales

Antes de entrar en el análisis de residuales se desarrolla el concepto de *Leverage de una observación* que será utilizado para realizar dicho análisis.

El valor predicho de un dato puede escribirse como combinación lineal de las observaciones

$$\hat{y}_i = b_0 + b_1 x_i = \sum_{k=1}^{n} h_{ik} y_k$$

Donde,

$$h_{ik} = \frac{1}{n} + \frac{(x_i - \bar{X})(x_k - \bar{X})}{S_{XX}}$$

y como caso particular tenemos que

$$\boldsymbol{h_i = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{S_{XX}}}$$

La cantidad h_i se denomina ***leverage del dato i-ésimo***, o en español influencia. Es una medida que resume cuán lejos cae el valor de x_i de la media muestral de las X. Mide, de alguna manera, cuánto es el aporte de la observación i-ésima a la varianza muestral de las X. Observemos que es un concepto que no depende del valor y_i observado.

Volviendo entonces a los residuos puede probarse que:

$$\boldsymbol{E(e_i) = 0}$$

$$\boldsymbol{Var(e_i) = S^2(1 - h_i)}$$

donde h_i es el leverage de la observación i-ésima. En consecuencia, la varianza del residuo de un dato depende del valor de la covariable, y los residuos de distintos casos tienen diferentes varianzas. De la ecuación de varianza de e_i vemos que cuanto mayor sea h_i, menor será la varianza del e_i (mientras más cercano a uno sea h_i más cercana a cero será la varianza del residuo de la observación i-ésima). Esto quiere decir que para observaciones con gran h_i, $\hat{y}_i$ tenderá a estar cerca del valor observado y_i, sin importar cuánto sea el valor y_i observado. En el caso extremo e hipotético en que $h_i = 1$, la recta ajustada sería forzada a pasar por el valor observado (x_i, y_i).

Residuos estandarizados

Para hacer más comparables a los residuos entre sí, podemos dividir a cada uno de ellos por un estimador de su desvío estándar, obteniendo lo que se denominan residuos estandarizados:

$$\boldsymbol{e_{zi} = \frac{(y_i - \hat{y}_i)}{S_{ei}}}$$

Donde S_{ei} es un estimador de su desvío estándar

$$S_{ei} = \sqrt{Var(e_i)} = \sqrt{s^2(1 - h_i)}$$

$$\boldsymbol{S_{ei} = s\sqrt{1 - h_i}}$$

Puede probarse que los residuos estandarizados tienen media poblacional cero (igual que los residuos), y varianza constante.

Luego de estas definiciones previas, se detallan a continuación las gráficas de residuales que menciona el paper y como realizarlas.

3.2. Graficas de residuos

Como menciona el paper, las tres gráficas utilizadas para el análisis de residuos son:

- Gráfica de los residuales en función de la variable independiente
- Gráfica de residuales estandarizados
- Gráfica de probabilidad normal

Gráfica de los residuales en función de la variable x

Ésta es una gráfica en la que los valores de la variable independiente se representan en el eje horizontal y los valores de los residuales correspondientes en el eje vertical. Se grafica un punto para cada residual. También es usual presentar la gráfica de residuales con respecto a los valores de la variable dependiente ($\hat{y}_i$) estimados por la ecuación. Para la regresión lineal simple, la gráfica de residuales en función de x y la de residuales en función de $\hat{y}$ muestran la misma información; mientras que, para la regresión lineal múltiple, la gráfica de residuales en función de $\hat{y}$ se usa con más frecuencia, porque se maneja más de una variable independiente.

Como se mencionó anteriormente $E(e_i) = 0$ Esto quiere decir que el grafico de los residuos versus las x_i debe estar centrado alrededor del cero (de la recta horizontal de altura cero).

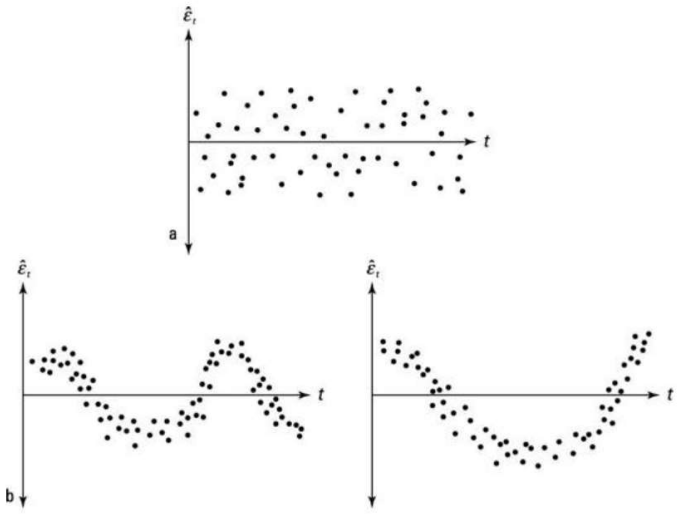


Figura 3.2. Gráficas de residuales en función de x

Gráfica de residuales estandarizados

Esta gráfica es similar a la gráfica de los residuales en función de la variable x, pero en este caso se trata de los residuales estandarizados en función de la variable independiente. La misma, debería mostrar menor variabilidad para los valores de x más alejados de la media muestral (serán los que tengan mayor leverage h_i).

La gráfica de residuales estandarizados nos brinda información acerca de la hipótesis de que el término de error tiene distribución normal. Si es cierta la hipótesis, cabe esperar que, aproximadamente, el 95% de los residuales estandarizados estén entre –2 y 2.

También es usual presentar la gráfica de residuales estandarizados con respecto a los valores de la variable dependiente ($\hat{y}$), esta es una combinación de las otras dos gráficas, mostrando implícitamente como varían los residuos con x y como se comparan los valores ajustados con los valores observados. Esta, es la más recomendada para el análisis de regresión múltiple.

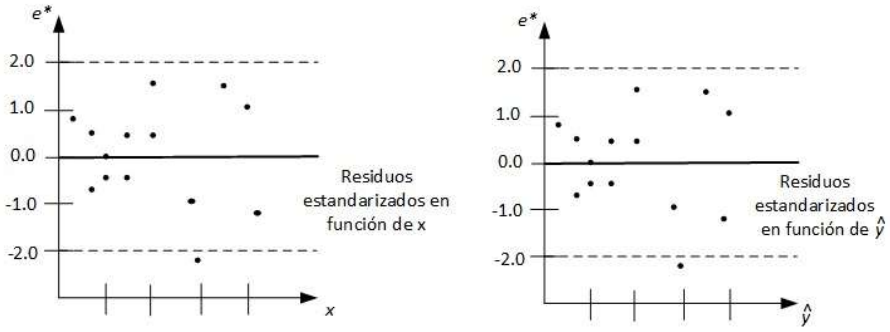


Figura 3.3. Residuos estandarizados en función de x y de $\hat{y}$

Grafica de probabilidad normal

Para realizar la gráfica de probabilidad normal es necesario calcular el porcentaje empírico de residuos menor que el residuo específico que se considera así:

$$\boldsymbol{P_e = \frac{i - 0,5}{n}}$$

Donde *i*: es el número de orden de cada dato y n es el total de datos.

Luego se calcula el porcentaje teórico de residuos menor que el residuo específico usando la tabla de distribución normal en función del porcentaje empírico de residuos menor que el residuo específico mencionado anteriormente.

Y por último se grafica la pareja $(F(residuo), P_e)$

Si los puntos parecen ajustarse a una línea recta (de la forma $y = x$), indicaría que los datos provienen de una distribución normal.

Ejemplos de gráficas de residuos y como interpretarlas

En el caso en el que el modelo es incorrecto, el gráfico de residuos (o de residuos estandarizados) versus la variable predictora (o versus los valores predichos) suele tener algún tipo de estructura. En la Figura 3.5 se ven varios de estos posibles gráficos de residuos.

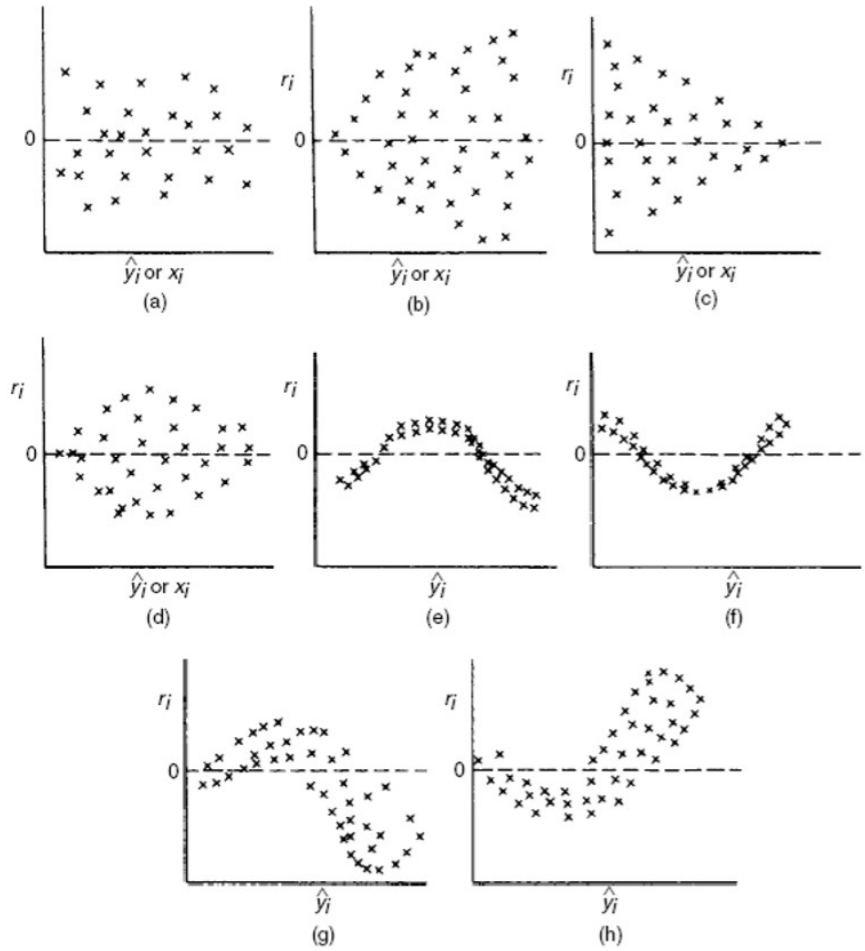


Figura 3.4. *Ejemplos de gráficos de residuos*

El primero de ellos es una nube de puntos sin estructura que indica que no hay problemas con el modelo ajustado.

De las Figuras 3.4 (b) a 3.4 (d) inferiríamos que el supuesto de homogeneidad de varianzas no se satisface: la varianza depende de la variable graficada en el eje horizontal.

Las Figuras 3.4 (e) a 3.4 (h) son indicadoras de que se viola el supuesto de linealidad de la esperanza condicional, lo cual nos lleva a pensar que el vínculo entre el valor esperado de la variable respuesta y y la covariable se ve mejor modelado por una función más complicada que la lineal (lo que genéricamente suele denominarse una curva).

Las dos últimas figuras, las 3.4 (g) y 3.4 (h) sugieren la presencia simultánea de curvatura y varianza no constante. En la práctica, los gráficos de residuos no son tan claros como estos.

Es útil recordar que aun cuando todos los datos satisficieran todos los supuestos, la variabilidad muestral podría hacer que el gráfico tuviera pequeños apartamientos de la imagen ideal.

Como se mencionó anteriormente sumaremos al análisis de residuales un análisis de outliers y observaciones influyentes para completar el análisis asegurándonos que la estimación realizada no sea dependiente de un sólo dato (o grupo de datos).

3.3. Outliers y observaciones influyentes

En algunos problemas, la respuesta observada para algunos pocos casos puede parecer no seguir el modelo que sí ajusta bien a la gran mayoría de los datos. Es decir, el modelo lineal puede ser correcto para la mayoría de los datos, pero uno o algunos de los casos está muy alejado de lo que el modelo ajustado le prescribe. Diremos que este dato alejado es un *outlier*. Observemos que el concepto de outlier es un concepto relativo al modelo específico en consideración. Si se modifica la forma del modelo propuesto a los datos, la condición de ser outlier de un caso individual puede modificarse. En otras palabras, un outlier es un caso que no sigue el mismo modelo que el resto de los datos.

La identificación de estos casos puede ser útil ya que, entre otras cosas, el método de mínimos cuadrados es muy sensible a observaciones alejadas del resto de los datos. De hecho, las observaciones que caigan lejos de la tendencia del resto de los datos pueden modificar sustancialmente la estimación.

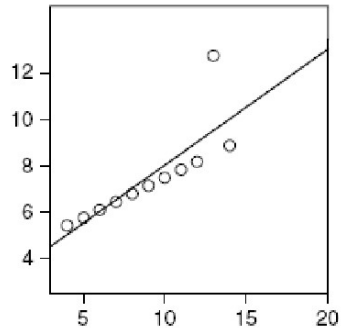


Figura 3.5. Ejemplo de el desajuste de una observación al modelo ajustado.

Un outlier se detecta por tener un residuo que es un valor inusual, muy grande o pequeño en relación con la distribución asociada a los residuos. Dado que los residuos estandarizados e_{zi} son una muestra aleatoria de una distribución normal con media cero y desviación estándar uno, se verifica que aproximadamente un 68% de los e_{zi} deben estar entre -1 y1, y alrededor del 95% entre -2 y 2 y prácticamente todos entre -3 y 3. Por ello, un residuo estandarizado que diste más de 3 o 4 unidades del 0 corresponde, potencialmente, con una observación atípica.

Si sospechamos que la observación i-ésima es un outlier podemos proceder del siguiente modo:

1. Eliminamos esa observación de la muestra, de modo que ahora tenemos una muestra con $n-1$ casos.
2. Usando el conjunto de datos reducidos volvemos a estimar los parámetros, obteniendo $\hat{b}_{0(i)}, \hat{b}_{1(i)}, \hat{S}_{(i)}{}^2$ donde el subíndice (i) está escrito para recordarnos que los parámetros fueron estimados sin usar la i-ésima observación.
3. Para el caso omitido, calculamos el valor ajustado $\hat{Y}_{i(i)} = \hat{b}_{0(i)} + \hat{b}_{1(i)} x_i$. Como el caso i-ésimo no fue usado en la estimación de los parámetros, Y_1 y $\hat{Y}_{i(i)}$ son independientes. La varianza de $Y_1 - \hat{Y}_{i(i)}$ puede calcularse y se estima usando $\hat{S}_{(i)}{}^2$.

4. Escribimos

$$t_i = \frac{Y_1 - \hat{Y}_{i(i)}}{\sqrt{\hat{S}^2(Y_1 - \hat{Y}_{i(i)})}}$$

la versión estandarizada del estadístico en consideración. Si la observación i-ésima sigue el modelo, entonces la esperanza de $Y_1 - \hat{Y}_{i(i)}$ debería ser cero. Si no lo sigue, será un valor no nulo. Luego, si llamamos δ a la esperanza poblacional de esa resta, $\delta = E(Y_1 - \hat{Y}_{i(i)})$, y asumimos normalidad de los errores, puede probarse que la distribución de t_i bajo la hipótesis H_0: $\delta = 0$ es una t de Student con $n-3$ grados de libertad, $t_i \sim t_{n-3}$ (recordar que hemos excluido una observación para el cálculo del error estándar que figura en el denominador, por eso tenemos un grado de libertad menos que con los anteriores test), y rechazar cuando este valor sea demasiado grande o demasiado pequeño.

Hay una fórmula sencilla para expresar a t_i sin necesidad de reajustar el modelo lineal con un dato menos, ya que es fácil escribir al desvío estándar estimado sin la observación i-ésima ($\hat{S}_{(i)}$) en términos del leverage de la observación i-ésima (h_i) y el desvío estándar estimado con toda la muestra ($\hat{S}$).

Es la siguiente

$$\boldsymbol{t_i = \frac{e_i}{\hat{S}_{(i)}\sqrt{1 - h_i}} = e_{zi}\sqrt{\frac{n-3}{n-2-e_{zi}}}}$$

siendo e_{zi} los residuos estandarizados. Esta cantidad se denomina el residuo estudentizado i-ésimo. La ecuación de t_i nos dice que los residuos estudentizados y los residuos estandarizados llevan la misma información, ya que pueden ser calculados uno en función de otro. Vemos entonces que para calcular los residuos estudentizados no es necesario descartar el caso i-ésimo y volver a ajustar la regresión.

Para completar el test, queda únicamente decidir contra qué valor comparar el t_i para decidir si la i-ésima observación es o no un outlier. Puede usarse un procedimiento conservativo conocido como método de Bonferroni para comparaciones múltiples.

Este procedimiento propone rechazar H_0: ninguna de las n observaciones es un outlier, versus H_1: hay al menos un outlier, cuando alguno de los $|t_i|$ es mayor que el percentil $1 - \alpha/2n$ de la t_{n-3}.

Este test, ubica un outlier, pero no nos dice qué hacer con él. Puede tratarse de un dato mal registrado, o que fue mal transcripto a la base de datos, en tal caso podremos eliminar el outlier (o corregirlo) y analizar los casos restantes. Pero si el dato es correcto y no hay razones para excluirlo del análisis entonces la estimación de los parámetros debería hacerse con un método robusto, que, a diferencia de mínimos cuadrados, no es tan sensible a observaciones alejadas de los demás datos.

Observaciones influyentes

Estudiar la influencia de las observaciones es, de alguna manera, estudiar los cambios en el análisis cuando se omiten uno o más datos. La idea es descubrir los efectos o la influencia que tiene cada caso en particular comparando el ajuste obtenido con toda la muestra con el ajuste obtenido sin ese caso particular (o sin esos pocos casos particulares). Una observación se denomina influyente si al excluirla de nuestro conjunto de datos la recta de regresión estimada cambia notablemente.

A continuación, se observan gráficos de cuatro conjuntos de 18 datos cada uno.

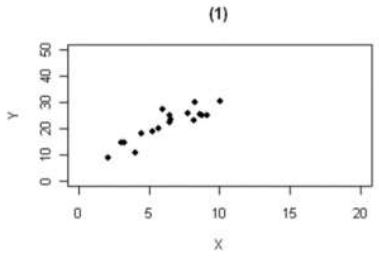


Figura 3.6. Ejemplo de diagramas de dispersión 1

En la figura 3.6, el conjunto de datos no presenta ni puntos influyentes ni outliers, ya que todas las observaciones siguen el mismo patrón.

En las figuras restantes se conservaron 17 de las observaciones de la figura 3.6 y se intercambió una de ellas por los puntos que aparecen indicados como A, B y C en los respectivos gráficos, y que son puntos atípicos en algún sentido, es decir, puntos que no siguen el patrón general de los datos. No todos los casos atípicos tendrán una fuerte influencia en el ajuste de la recta de regresión.

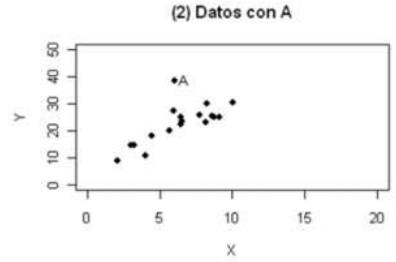


Figura 3.7. Ejemplo de diagramas de dispersión 2

En la figura 3.7, entre las observaciones figura una que rotulamos con la letra A. El caso A puede no ser muy influyente, ya que hay muchos otros datos en la muestra con valores similares de x que evitarán que la función de regresión se desplace demasiado lejos siguiendo al caso A.

Por otro lado, los casos B y C (figuras 3.8 y 3.9) ejercerán una influencia muy grande en el ajuste, ya que el leverage de ambas será bastante alto. Mientras mayor sea el leverage de la observación, menor será la variabilidad del residuo, esto quiere decir que para observaciones con gran leverage, el valor predicho tendrá que estar cerca del valor observado. Por eso se dice que tienen un alto grado de influencia, o que cada uno de ellos

es un punto de alta influencia. Luego la recta ajustada se verá matemáticamente obligada a acercarse a dichas observaciones, alejándose para ello, de los demás datos.

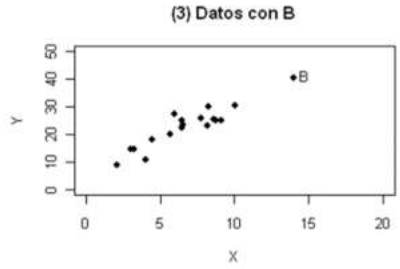


Figura 3.8. Ejemplo de diagramas de dispersión 3

En la Figura 3.8. aparece una observación indicada con la letra B. Esta observación será influyente en el ajuste, pero como sigue el patrón lineal de los datos (o sea, sigue la estructura de esperanza condicional de y cuando x es conocida que tienen el resto de los datos) no hará que la recta estimada cuando el punto está en la muestra varíe mucho respecto de la recta estimada en la situación en la que no está, pero reforzará la fuerza del ajuste observado y reforzará la significatividad de los test que se hagan sobre los parámetros.

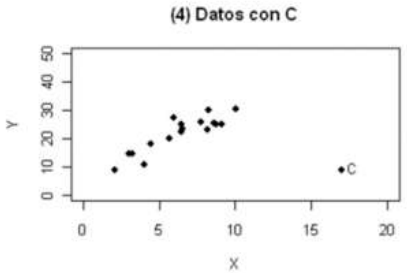


Figura 3.9. Ejemplo de diagramas de dispersión 4

La Figura 3.9. presenta la observación C. Esta observación será muy influyente en el ajuste, arrastrando a la recta estimada a acercarse a ella. Como no sigue la misma estructura de esperanza condicional que el resto de las observaciones, la recta ajustada en este caso diferirá mucho de la que se ajusta a los datos de la Figura 3.6. Sin embargo, si una vez realizado el ajuste intentamos identificar este punto mirando las observaciones de mayores residuos (o residuos estandarizados) es posible que no la detectemos (dependerá de cuán extrema sea) ya que, al arrastrar la recta hacia ella, tendrá un residuo mucho menor que el que tendría si usáramos la recta que ajusta a los datos de la figura 3.6.

A continuación, en la figura 3.10 se encuentran las mismas gráficas de las figuras 3.6 a 3.9 pero con las rectas ajustadas.

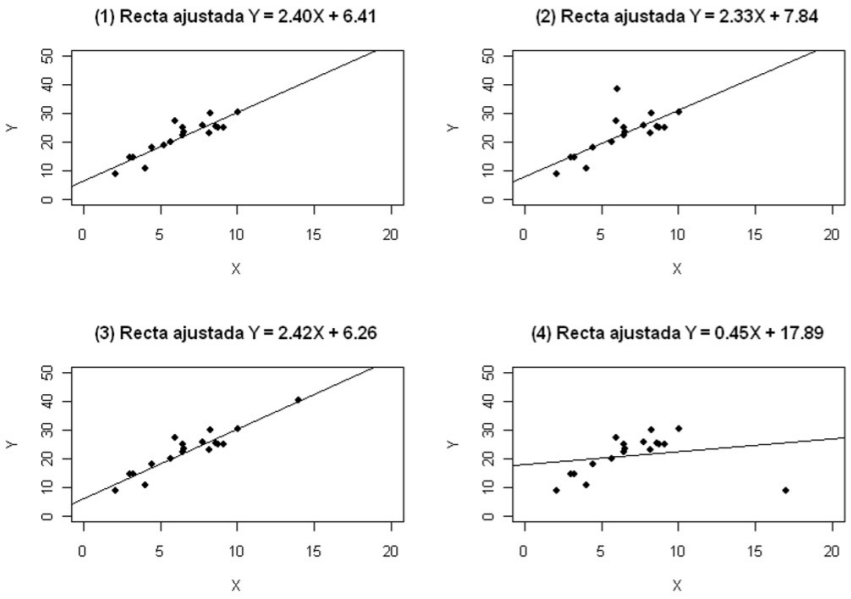


Figura 3.10. gráficos de las figuras 3.6 a 3.9 con las rectas ajustadas

Una vez realizado el ajuste vemos que se verifica lo anticipado. Las pendientes de las rectas estimadas en los 3 primeros gráficos no difieren demasiado entre sí, en el gráfico 3.10 (2) la ordenada al origen es mayor ya que la observación A está ubicada muy por encima de los datos. La recta estimada en 3.10 (3) pasa casi exactamente por el dato B y la significatividad del test para la pendiente aumenta en este caso, comparada con la del gráfico 3.10 (1). En el gráfico 3.10 (4) vemos que la recta ajustada difiere completamente de la recta estimada para el conjunto (1), de hecho, la pendiente que era significativa para los datos del gráfico 3.10 (1) deja de serlo en este caso. Vemos que la observación C arrastró la recta hacia ella. La observación C es la que más tergiversó las conclusiones del ajuste lineal.

En las situaciones prácticas, cuando hay más de un dato anómalo en un conjunto de datos, esta presencia simultánea puede enmascararse: la técnica de sacar las observaciones de a una muchas veces no logra detectar los problemas.

3.4. Uso de la ecuación de regresión para estimar y predecir

Una vez planteada la gráfica de residuos, el paper en estudio habla sobre utilizar la ecuación de regresión para realizar estimaciones y predecir ciertos valores. Esto puede realizarse solo si todo el análisis previo demuestra que existe una relación estadísticamente significativa entre las variables.

Luego muestra dos tipos de estimaciones por intervalos, el primer tipo de estimado es el de ***intervalo de confianza***, que es un estimado del valor medio de y para determinado valor de x. El segundo tipo es el estimado de ***intervalo de predicción***, que se usa cuando deseamos un estimado de intervalo de valor individual de y que corresponda a determinado valor de x. Con la estimación puntual se obtiene el mismo valor, sea que se esté estimando el valor medio de y o prediciendo un valor individual de y, pero con los estimados de intervalo se obtienen valores distintos.

Estimado del intervalo de confianza del valor medio de y

Para calcular este intervalo, el paper explica el procedimiento reemplazando un valor puntual de x (x_p) en la ecuación de la regresión y así obtener el $\hat{y}_p$ que define como estimado de un valor particular de y. Luego, dado que no se puede esperar que $\hat{y}_p$ sea exactamente igual a $E(\hat{y}_p)$ menciona que es necesario considerar la varianza de los estimados basados en la ecuación de regresión.

La fórmula para estimar la desviación estándar de $\hat{y}_p$ dado un valor particular de x, x_p, es:

$$\boldsymbol{S_{\hat{y}p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}}$$

Por consiguiente, se construye el intervalo de confianza de $100(1-\alpha)\%$ sobre la respuesta media $\bar{Y}$ a partir del estadístico

$$\boldsymbol{t = \frac{\hat{y}_p - \bar{Y}}{S_{\hat{y}p}}}$$

que tiene una distribución t con $n-2$ grados de libertad.

Entonces, se puede decir que:

Definición 3.1. *La ecuación general para un estimado del intervalo de confianza de $\hat{y}_p$ dado un valor particular de x es*

$$\boldsymbol{\hat{y}_p - t_{\alpha/2} * S_{\hat{y}p} < \hat{y}_p < \hat{y}_p + t_{\alpha/2} * S_{\hat{y}p}}$$

Donde el coeficiente de confianza es $1-\alpha$ y $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad

Se observa además que la desviación estándar estimada de x_p es mínima cuando $x_p = \bar{x}$. Esto implica que podemos hacer el mejor estimado, o el más preciso, del valor medio de y siempre que se esté usando el valor medio de x. Como resultado de ello, los intervalos de confianza para el valor medio de y se ensanchan a medida que x_p se aleja de $\bar{x}$.

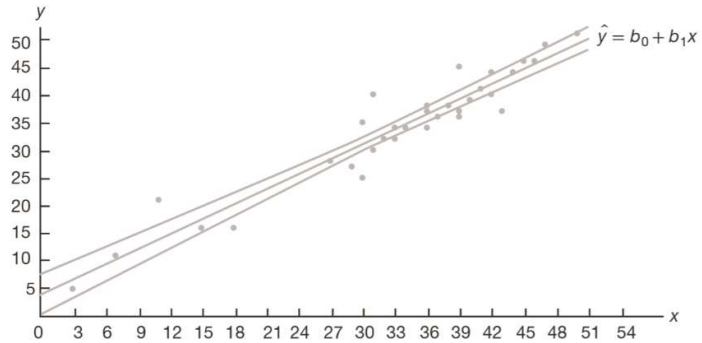


Figura 3.11. Límites de confianza para el valor medio de y

Estimado del intervalo de predicción para un valor particular de y

El otro tipo de estimado del intervalo que menciona el paper es el de predicción para un valor particular de y. Al igual que el anterior en una primera instancia se calcula el valor de $\hat{y}_p$, pero luego, es necesario estimar la varianza asociada al empleo de $\hat{y}_p$ la cual está formada por la suma de dos componentes: La varianza de los valores individuales de y respecto del promedio, cuyo estimado es s^2 y la varianza asociada al uso de $\hat{y}_p$ para estimar $E(y_p)$, cuyo estimado es $S_{\hat{y}p}$. Así, el estimado de la varianza de un valor individual es:

$$\boldsymbol{S^2_{ind} = S^2 + S_{\hat{y}_p}}$$

Por consiguiente, un estimado de la desviación estándar de un valor un individual de $\hat{y}_p$ es:

$$\boldsymbol{S_{ind} = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{XX}}}}$$

Así, un intervalo de predicción de $100(1-\alpha)\%$ para un solo valor pronosticado y_p se puede construir a partir del estadístico

$$\boldsymbol{t = \frac{\hat{y}_p - y_p}{S_{ind}}}$$

que tiene una distribución t con $n-2$ grados de libertad.

Entonces,

Definición 3.2. *La ecuación general para un estimado del intervalo de predicción para un valor individual de y dado un valor particular de x es:*

$$\boldsymbol{\hat{y}_p - t\alpha_{/2} * S_{ind} < \hat{y}_p < \hat{y}_p + t\alpha_{/2} * S_{ind}}$$

Donde el coeficiente de confianza es $1-\alpha$ y $t\alpha_{/2}$ es un valor de la distribución t con $n-2$ grados de libertad

La distinción básica entre los dos es que el intervalo de predicción predice en qué rango caerá una observación individual futura, mientras que un intervalo de confianza muestra el rango probable de valores asociados con algún parámetro estadístico de los datos, como la media de la población.

Esta es una distinción importante, porque el intervalo de confianza de los valores medios para las poblaciones muestreadas será más pequeño o más estricto que el intervalo de predicción para los mismos datos. El intervalo de predicción debe ser lo suficientemente amplio como para incluir casi todos los puntos de datos reales, mientras que el intervalo de confianza solo necesita incluir promedios de muestras de datos, que necesariamente caen dentro de un límite mucho más pequeño.

Estimación de los parámetros del modelo de regresión lineal

Para finalizar con todo el análisis de regresión el paper plantea identificar intervalos de confianza para los estimadores de β_0 y β_1 y escribe las ecuaciones para dicho intervalo. Para calcularlos se procede de la misma manera que los intervalos de confianza descritos anteriormente.

Para el intervalo de confianza del estimador β_1 se utiliza el mismo estadístico calculado en la prueba t:

$$\boldsymbol{t = \frac{b_1 - \beta_1}{S \Big/ \sqrt{S_{xx}}}}$$

El cual como se mencionó antes, tiene una distribución t con $n - 2$ grados libertad y por lo tanto se define que:

Definición 3.3. *Un intervalo de confianza de* $100(1-\alpha)\%$ *para el parámetro* $\boldsymbol{\beta_1}$ *en la recta de regresión* $Y = \beta_0 + \beta_1 x$, *es*

$$\boldsymbol{b_1 - t_{\alpha/2}\frac{S}{\sqrt{S_{XX}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{XX}}}}$$

donde $t_{\alpha/2}$ *es un valor de la distribución* t con $n - 2$ *grados de libertad.*

Por otro lado, para el cálculo del intervalo de confianza de β_0, el estadístico va a ser:

$$\boldsymbol{t = \frac{b_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{XX}}}}}$$

Con una distribución t con $n - 2$ grados de libertad.

Luego,

Definición 3.4. *Un intervalo de confianza de* $100(1-\alpha)\%$ *para el parámetro* $\boldsymbol{\beta_0}$ *en la recta de regresión* $Y = \beta_0 + \beta_1 x$, *es*

$$\boldsymbol{b_0 - t_{\alpha/2}S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{XX}}} < \beta_0 < b_0 + t_{\alpha/2}S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{XX}}}}$$

donde $t_{\alpha/2}$ *es un valor de la distribución* t con $n - 2$ *grados de libertad.*

Con todo esto se puede validar que los valores de los estimadores de b_0 y b_1 de los parámetros de β_0 y β_1 respectivamente, caigan dentro del intervalo de confianza calculado para cada uno.

Capítulo 4

Aplicación de regresión lineal a la pobreza en Argentina

A continuación, se plantea una aplicación de toda la teoría estudiada en los capítulos 2 y 3 en un análisis de regresión de la pobreza en Argentina.

Como punto de partida, la siguiente tabla fue realizada con los datos del cuadro "*4.3. Pobreza en hogares y personas. Regiones estadísticas y 31 aglomerados urbanos"* del informe técnico Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos publicado 1 de abril de 2020 por el INDEC.

Para los efectos explicativos pertinentes a este documento se considera únicamente la variable pobreza en personas, donde X será el porcentaje de pobreza en personas en el segundo semestre del 2018 y Y será el porcentaje de pobreza en personas en el segundo semestre del 2019. Por tanto, lo que nos interesa es evidenciar si el porcentaje de pobreza en personas en el segundo semestre del 2019 depende linealmente del porcentaje de pobreza en personas del segundo semestre del 2018 para estos aglomerados urbanos.

Tabla 4.1: Pobreza en hogares y personas en 31 aglomerados urbanos.

Área geográfica	2° semestre 2018		2° semestre 2019	
	Hogares	Personas	Hogares	Personas
Ciudad Autónoma de Buenos Aires	8,1	12,6	8,7	13,5
Partidos del GBA	28,2	35,9	31,8	40,5
Gran Mendoza	21,9	30,7	28,9	38,6
Gran San Juan	22,2	33,1	20,9	32,3
Gran San Luis	22,5	31,3	25,3	35
Corrientes	38,4	49,3	27,6	37,9
Formosa	23,8	32,5	31,2	41,6
Posadas	26,5	35,7	30,9	41,3
Gran Catamarca	26,9	35,5	30,4	39,6
Gran Tucumán - Tafí Viejo	25	32,2	27,9	37,3
Jujuy - Palpalá	24,7	31,7	28,2	37,8
La Rioja	22,4	30,5	27,2	38
Salta	29,3	37,7	34,5	45,5
Santiago del Estero - La Banda	28	38,9	34,5	45,2
Bahía Blanca - Cerri	17,5	25	20,8	28,1
Concordia	31,1	41,9	40,7	51,1
Gran Córdoba	24	36,5	25,5	37,4
Gran La Plata	21,5	30,9	22	30,6
Gran Rosario	23,5	31,8	25,5	35
Gran Paraná	20,8	29,4	20,6	30
Gran Santa Fe	23,1	34,4	23	34,4
Mar del Plata	18,6	24,8	18,4	25
Río Cuarto	19,6	29,3	21,2	29,4
Santa Rosa - Toay	22,6	32,1	23,9	33,9
San Nicolás - Villa Constitución	23,4	33,1	29,1	39,6
Comodoro Rivadavia - Rada Tilly	17,1	22	19,7	26,2
Neuquen - Plottier	21	27,1	21,9	28,6
Río Gallegos	15,3	18,6	19,3	25,7
Ushuaia - Río Grande	11,9	17,9	23,9	31,5
Rawson - Trelew	25,3	35,3	28,5	39,5
Viedma - Carmen de Patagones	20,1	29,7	22	32,9

Fuente: INDEC. Encuesta Permanente de Hogares.

Para comenzar con el análisis se realiza un diagrama de dispersión de los datos de la tabla 4.1. Los valores del porcentaje de pobreza en personas del segundo semestre del 2018 se representan en el eje horizontal y los valores del porcentaje de pobreza en personas en el segundo semestre del 2019 se representan en el eje vertical.

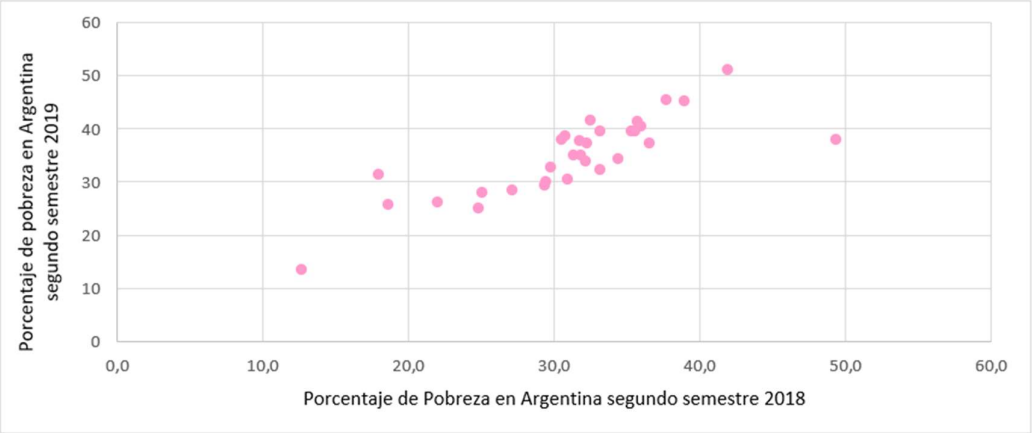


Figura 4.1. Gráfica de dispersión de pobreza en personas en Argentina

La figura 4.1. nos muestra que, conforme aumenta la pobreza en ciertas poblaciones en el segundo semestre del 2018, también aumenta la pobreza en el segundo semestre del año 2019, lo cual indica una relación directa entre las variables. Además, se observa que los puntos parecen aproximarse a una línea recta. En consecuencia, se elige el modelo de regresión lineal simple para representar la relación entre las variables

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

La ecuación estimada de regresión es

$$\hat{Y} = b_0 + b_1 x$$

4.1. Aplicación del método de mínimos cuadrados

Para calcular el valor de los regresores (b_0 y b_1) se emplea el método de los mínimos cuadrados en el cual se enuncian las siguientes ecuaciones normales:

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \qquad [\mathbf{1}]$$

$$\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2 \qquad [\mathbf{2}]$$

Donde n es el número de observaciones.

Para el caso de estudio, reemplazamos con los datos de la tabla 4.1 en las ecuaciones y resolviendo algebraicamente el sistema de ecuaciones obtenemos b_0 y b_1.

Observación: Para el siguiente desarrollo se realizaron cálculos adicionales, los cuales se pueden observar en la Tabla 7.2. "Cálculos auxiliares para la obtención de los regresores b_0 y b_1" del capítulo 7 "Cálculos y tablas auxiliares".

Reemplazando en [1] los datos se obtiene:

$$1083 = 31b_0 + b_1 967{,}4$$
$$1083 - b_1 967{,}4 = 31b_0$$
$$\frac{1083 - b_1 967{,}4}{31} = b_0$$
$$34{,}9355 - b_1 31{,}2064 = b_0 \qquad [\mathbf{3}]$$

Luego reemplazando la ecuación [3] en la [2]

$$\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2$$
$$\sum_{i=1}^{n} y_i x_i = (34{,}9355 - b_1 31{,}2064) \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2$$
$$35084{,}1 = (34{,}9355 - b_1 31{,}2064) 967{,}4 + b_1 31732{,}46$$
$$35084{,}1 = 33796{,}6027 - b_1 30189{,}0714 + b_1 31732{,}46$$
$$35084{,}1 - 33796{,}6027 = b_1 1543{,}3886$$
$$\frac{1287{,}4973}{1543{,}3886} = b_1$$
$$\mathbf{0{,}8342 = b_1}$$

Y el valor obtenido de b_1 lo reemplazamos en **[3]** para obtener b_0

$$34{,}9355 - 0{,}8342 * 31{,}2064 = b_0$$
$$34{,}9355 - 26{,}0324 = b_0$$
$$\mathbf{8{,}9031 = b_0}$$

Remplazando estos datos se obtiene que para las variables pobreza de personas en el segundo semestre 2018 y pobreza de personas en el segundo semestre en 2019 la ecuación estimada de la regresión es:

$$\boldsymbol{\widehat{Y} = 8{,}9031 + 0{,}8342x} \qquad [\mathbf{4}]$$

El coeficiente b_1 corresponde a la pendiente de la recta. En este caso, la pendiente de la recta es positiva, lo que implica, tal como se observa en el gráfico de dispersión, que en los aglomerados donde se observó mayor pobreza en el segundo semestre del 2018, también se observó mayor pobreza en el segundo semestre del 2019. Pero como la pendiente es un número entre cero y uno, significa que el incremento en el porcentaje de pobreza en el segundo semestre del 2019 entre un aglomerado y otro es menor que en el segundo semestre del 2018.

La figura 4.2 muestra la gráfica de esta ecuación sobre el diagrama de dispersión.

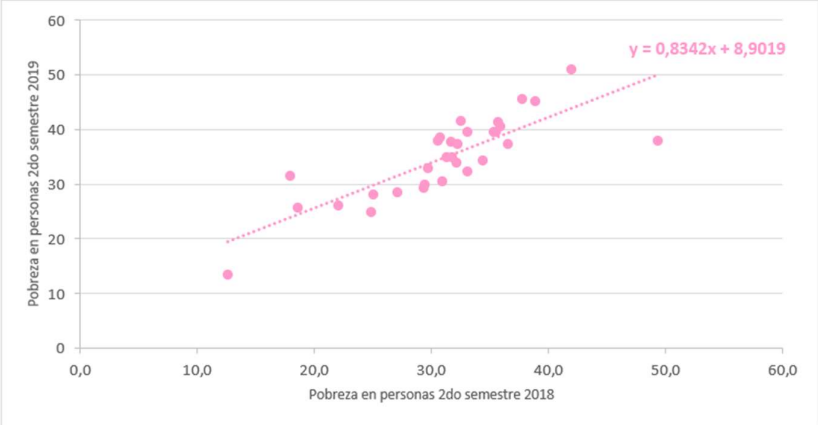


Figura 4.2. Grafica de la ecuación de regresión lineal.

Como se mencionó en los núcleos teóricos, con el fin de determinar la pertinencia de la ecuación de regresión hallada, es necesario hacer un análisis de la bondad de ajuste de la recta, demostrar si la relación es estadísticamente significativa y validar los supuestos acerca del término de error.

4.2. Análisis de los coeficientes de correlación y determinación

Para el próximo cálculo se necesita el valor de la media de Y ($\bar{Y}$):

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

$$\bar{Y} = \frac{1}{31}\sum_{i=1}^{n} y_i = \frac{1083{,}0}{31} = \mathbf{34{,}9355}$$

Además, para obtener el coeficiente de determinación se calcula previamente el valor de las siguientes sumas de cuadrados:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \, , SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \, , SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Al hacer los cálculos de las ecuaciones con los datos de la tabla 7.2 del capítulo 7 *"Cálculos y tablas auxiliares"*, se obtiene:

$$\boldsymbol{SSE = 540{,}5979}$$

$$\boldsymbol{SST = 1614{,}6910}$$

$$\boldsymbol{SSR = 1073{,}9934}$$

Se sabe además que existe una relación entre las tres sumas:

$$SST = SSR + SSE$$

$$1614{,}6910 = \ 1073{,}9934 + 540{,}5979$$

La cual podemos validar con los datos obtenidos, obsérvese que puede presentarse una mínima diferencia por tomar cuatro decimales para realizar los cálculos y no los números con todos sus decimales.

Luego, se puede calcular el valor del coeficiente de determinación

$$r^2 = \frac{SSR}{SST} = \frac{1073{,}9934}{1614{,}6910} = \mathbf{0{,}6651}$$

Esto revela que la ecuación de regresión explica en un 66,51% los valores observados de la pobreza en segundo semestre del 2019 según los valores de pobreza en el segundo semestre del 2018.

La segunda medida que se usa para describir qué tan bien explica una variable a la otra es el coeficiente de correlación:

$$r = (signo\ de\ b_1)\sqrt{r^2}$$

En la regresión que estamos analizando obtenemos

$$r = \sqrt{0{,}6651}\ = \mathbf{0{,}8155}$$

Este valor muestra un coeficiente de correlación alto, lo que implica una relación directa de dependencia lineal fuerte entre los valores de pobreza del segundo semestre del 2018 y el segundo semestre de 2019 en los 31 aglomerados urbanos de Argentina.

Como se mencionó en el núcleo teórico, el coeficiente de determinación da una medida entre 0 y 1, mientras que el coeficiente de correlación da una medida entre -1 y 1, ambos casos fueron validados en los cálculos realizados.

En la deducción de la ecuación de regresión por mínimos cuadrados, y en el cálculo de los coeficientes de determinación y correlación, no se realizaron pruebas estadísticas de significancia de la relación entre x e y. Es por ello que, a continuación, se realizan las pruebas t y F.

4.3 Aplicación de las pruebas de significancia

Para realizar las pruebas de significancia, primero se debe calcular el valor del estimado de σ^2 y desviación estándar de la estimación

El error cuadrado medio (s^2) es el estimado de σ^2. Se calcula mediante la ecuación:

$$s^2 = \frac{SSE}{n-2}$$

Reemplazando los resultados obtenidos anteriormente se obtiene que

$$s^2 = \frac{540{,}5979}{31-2} = \frac{540{,}5979}{29} = \mathbf{18,6413}$$

Por otro lado, la desviación estándar del estimado se calcula como la raíz cuadrada de la varianza del estimado.

$$S = \sqrt{\frac{SSE}{n-2}}$$

En el análisis en curso

$$S = \sqrt{s^2} = \sqrt{18{,}6413} = \mathbf{4,3176}$$

Antes de realizar las pruebas de significancia, se necesita calcular además el valor de S_{XX} que se utilizara en los próximos desarrollos.

$$S_{XX} = \sum {x_i}^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{XX} = 31732{,}46 - \frac{(\,967{,}4)^2}{31}$$

$$S_{XX} = 31732{,}46 - 30189{,}1213$$

$$\boldsymbol{S_{xx} = 1543,3387}$$

Prueba t

En el modelo de regresión lineal, si las variables tienen una relación lineal, debe suceder que $\beta_1 \neq 0$. Se usan los datos de la muestra para probar las siguientes hipótesis:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Como ya se vio en el capítulo 2 para realizar la prueba t, se necesita obtener previamente la desviación estándar estimada

$$S_b = \frac{S}{\sqrt{S_{XX}}}$$

En el análisis en curso S= 4,3176 y $S_{XX} = 1543{,}3387$, luego

$$S_b = \frac{4{,}3176}{\sqrt{1543{,}3387}} = \frac{4{,}3176}{39{,}2854} = \mathbf{0,1099}$$

Luego, el estadístico de prueba es:

$$t = \frac{b_1}{S_b} = \frac{0{,}8342}{0{,}1099} = \mathbf{7,5905}$$

De acuerdo con la tabla 7.5. *"Distribución t de student"*, del capítulo 7. "*Cálculos y tablas auxiliares*", se observa que el valor de t que corresponde a $\alpha = 0{,}01$ y $n - 2 = 31 - 2 = 29$ grados de libertad es

$$\boldsymbol{t_{0{,}005} = 2{,}756}.$$

Como $7{,}5905 > 2{,}756$ se rechaza H_0 y se concluye que, a un nivel de significancia de 0,01, β_1 no es cero.. La evidencia estadística es suficiente para concluir que hay una relación importante entre las variables.

Prueba F

Como solo hay una variable independiente, la prueba F debe indicar la misma conclusión que la prueba t, pero cuando hay más de una variable independiente solo se puede usar la prueba F.

Se plantea entonces,

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Para el análisis en curso, el estadístico de prueba F es:

$$F = \frac{SSR}{S^2} = \frac{1073{,}9934}{18{,}6413} = \boldsymbol{57{,}6136}$$

En la tabla 7.6 "*Distribución F de Fisher con probabilidad de 0,05*", del capítulo 7, observamos que el valor de F que corresponde a $\alpha = 0{,}01$, con un grado de libertad en el numerador y $n - 2 = 29$ grados de libertad en el denominador, es

$$\boldsymbol{F_{0{,}01} = 4{,}183}.$$

Como 57,6136>4,183, rechazamos H_0 y se concluye que, a un nivel de significancia del 0,01, β_1 no es cero.

4.4 Análisis de graficas de residuos

El primer paso para el análisis de residuales es calcular el valor de cada uno de los residuos, los cuales como vimos en el núcleo teórico son la diferencia entre el valor real de y y el valor estimado de y ($e_i = y_i - \hat{y}_i$). Nótese que los mismos fueron calculados previamente para la fórmula de SSE. El detalle de estos se encuentra en la tabla 7.2 del capítulo "*Cálculos y tablas auxiliares"*.

Luego se realiza la primera gráfica de residuos en función de la variable x

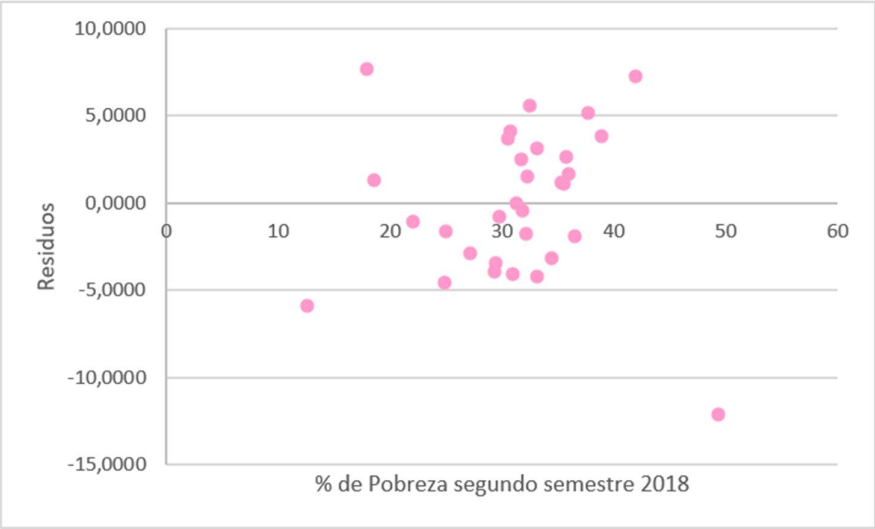


Figura 4.3. Gráfica de residuales de la pobreza en Argentina

Se estudio que la forma ideal de este gráfico es que los valores de los residuos ronden alrededor del valor 0 y podemos ver en la figura 4.3. que se cumple. Por otro lado, si realizamos la suma de los residuos y lo dividimos por las 31 observaciones vemos que la media es cero y por lo tanto se cumple la propiedad $E(e_i) = 0$.

Luego, para la siguiente gráfica, se debe calcular el valor de los leverage de cada dato (el cual se vio que resume cuán lejos cae el valor de x_i de la media muestral de las x) para luego poder calcular el desvío estándar de cada residuo. La fórmula del leverage es:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{S_{XX}}$$

Luego, se puede calcular el estimado de la desviación estándar del residual i, el cual depende del error estándar del estimado S (4,3176) y el valor h_i calculado en el paso anterior.

$$S_{ei} = s\sqrt{1 - h_i}$$

Una vez calculada la desviación estándar de cada residual, se procede a calcular el residual estandarizado.

$$e_{zi} = \frac{(y_i - \hat{y}_i)}{S_{ei}}$$

Todos los cálculos realizados para todas las observaciones y resultados necesarios para el análisis de residuales pueden observarse en la tabla 7.3 "*Cálculos auxiliares para el análisis de residuales*" del capítulo 7.

Posteriormente se grafican los residuales estandarizados calculados en función de x.

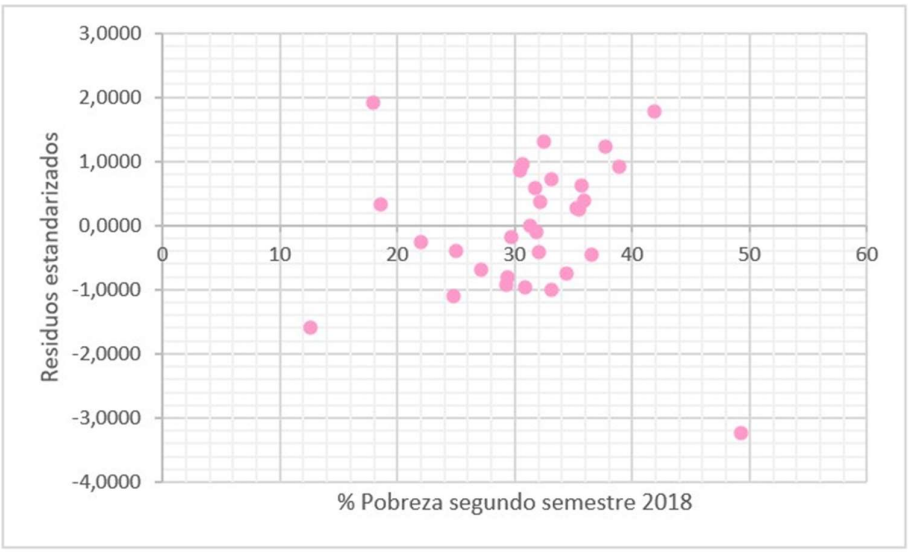


Figura 4.4. Gráfica de residuales estandarizados de la pobreza en Argentina

Como se mostró en la parte teórica, la gráfica de residuales estandarizados nos brinda información acerca de la hipótesis de que el término de error tiene distribución normal y se espera que aproximadamente el 95% de los residuales estandarizados estén entre –2 y 2. Observando la gráfica de residuales estandarizados (figura 4.4) notamos que solo un residual está fuera del intervalo mencionado (perteneciente al aglomerado de Corrientes). Puesto que se está trabajando con 31 observaciones, decir que uno de ellos está fuera del intervalo de desviaciones estándar implica que aproximadamente el 96,8% de los datos está dentro del intervalo y no habría razón suficiente para dudar de que el término de error tenga distribución normal.

También podemos validar que al sumar todos los residuos estandarizados y dividirlos por n obtenemos que la media es cero, y validamos además que el valor de todos los S_{ei} es constante, lo cual cumple con las propiedades que se habían planteado en el núcleo teórico. Observar que dado que se realizaron los cálculos con cuatro decimales puede variar el valor exacto de S_{ei} pero se puede ver en la tabla 7.3 "*Cálculos auxiliares para el análisis de residuales*" del capítulo 7, que en todos los casos se puede redondear a 4.

La tercera gráfica de residuo que se realizara es la gráfica de probabilidad normal la cual como vimos previamente confirma la suposición de que el término del error tiene una distribución normal.

Para el desarrollo de esta grafica se pueden observar los cálculos realizados según los pasos que desarrollaremos a continuación en la tabla 7.4. "*Cálculos auxiliares para la gráfica de probabilidad normal de los residuos*".

El primer paso para realizarla es ordenar los residuos de menor a mayor, luego se calcula el porcentaje empírico de residuos menor que el residuo específico que se está considerando de la siguiente manera:

$$P_e = \frac{i - 0{,}5}{n}$$

Donde i es el número de orden de cada dato y n es el total de datos.

Por último, se calcula el porcentaje teórico de residuos menor que el residuo específico usando la tabla de distribución normal (Tabla 7.7 capitulo 7. "*cálculos y tablas auxiliares*") de manera tal que

$$F(residuo) = P(Z < residuo\ estandar)$$

Una vez calculados todos estos datos se grafica la pareja $(F(residuo), P_e)$, el resultado es el siguiente

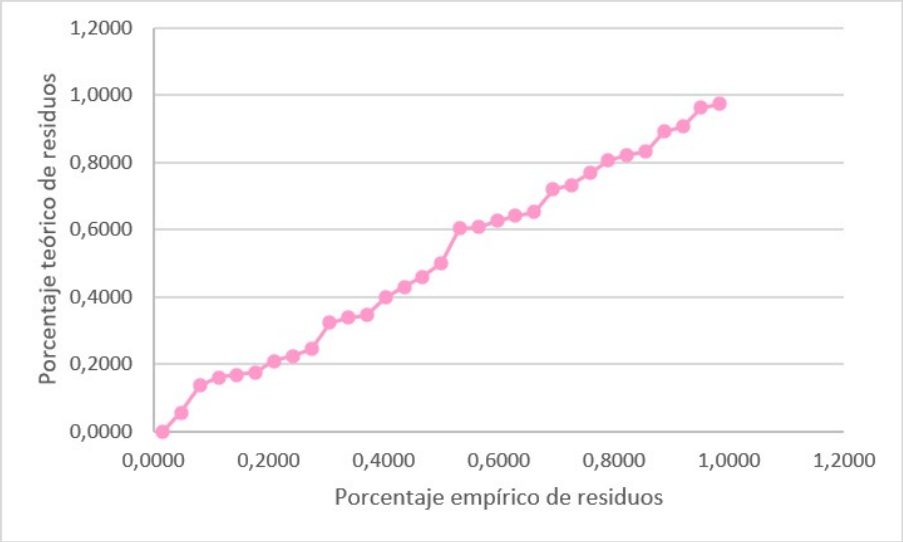


Figura 4.5. Gráfica de probabilidad normal de residuales estandarizados de la pobreza en Argentina

Como podemos observar en la figura 4.5. los puntos parecen ajustarse a una línea recta (de la forma $y = x$), lo cual confirma que los datos provienen de una distribución normal.

4.5. Búsqueda de outliers y observaciones influyentes

En el análisis en curso de la pobreza en argentina encontramos que un 77,4% de los e_{zi} está dentro del rango $(-1,1)$ y un 96,8% dentro de $(-2,2)$, quedando un solo residual estandarizado con valor -3,2318 que pertenece al aglomerado de Corrientes. Dado que este valor es superior al esperado podría ser un outlier.

Para no realizar todo el análisis nuevamente excluyendo al outlier detectado, se calcula el estadístico:

$$t_i = e_{zi}\sqrt{\frac{n-3}{n-2-e_{zi}}}$$

$$t_i = -3{,}2318\sqrt{\frac{28}{29 + 3{,}2318}}$$

$$t_i = -3{,}2318\sqrt{0{,}8687}$$

$$\boldsymbol{t_i = -3{,}0121}$$

Luego, se plantean las siguientes hipótesis

H_0: ninguna de las n observaciones es un outlier

H_1: hay al menos un outlier

Se debe comparar el $|t_i|$ obtenido con el percentil $1 - {}^{\alpha}/_{2n}$ de la t_{n-3}. En el análisis en curso

$$1 - {}^{\alpha}/_{2n} = 1 - {}^{0{,}05}/_{2 * 31} = 1 - 0{,}0008 = 0{,}9992$$

El valor del percentil obtenido en la tabla 7.8. "*Distribución t de student*". Del capítulo 7, para $n - 3 = 31 - 3 = 28$ es 3,408

Como $|t_i| = 3{,}0121 < 3{,}408$ se descarta la Hipótesis 1 y por lo tanto ninguna de las n observaciones es un outlier.

<u>Detección de observaciones influyentes</u>

Como se explicó en la teoría, si existen observaciones influyentes, al quitarlas del análisis la recta de regresión estimada debería cambiar su valor considerablemente.

En la siguiente gráfica se muestra el resultado de quitar de la base de porcentaje de pobreza de personas en Argentina el aglomerado de Corrientes $(49{,}3; 37{,}9)$. Se puede validar que no se observan cambios en la ecuación de la regresión ni en el coeficiente de correlación.

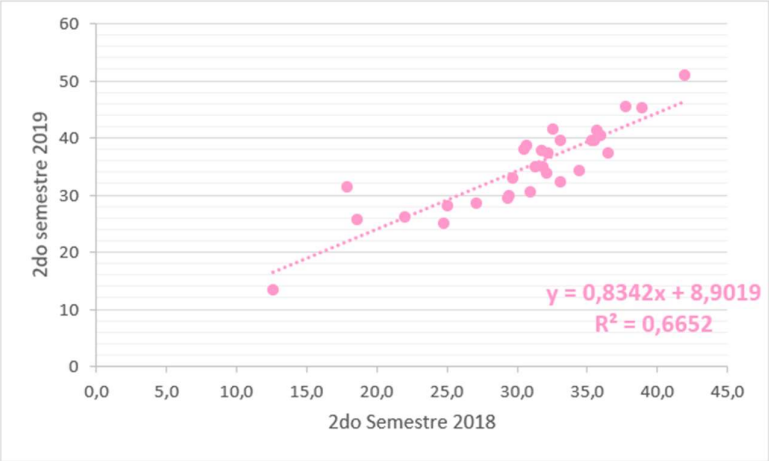


Figura 4.6. Grafica de la ecuación de regresión lineal sin el aglomerado Corrientes.

De esta misma manera, se realizó la prueba descartando ciertos valores en el análisis de regresión de la pobreza en argentina, y en ninguno de los casos cambio la ecuación de regresión, ni el coeficiente de regresión, por lo cual concluimos que no se detectan observaciones influyentes.

4.6. Aplicación de la ecuación de regresión para estimar y predecir

Como pudimos validar que la relación es estadísticamente significativa entre las variables, y que el ajuste que proporciona la ecuación es bueno, la ecuación puede usarse para estimaciones y predicciones. Cómo vimos en el capítulo 3 se realizan estimaciones por intervalo.

El primero que se va a calcular es el intervalo de confianza del valor medio de y.

A modo de ejemplo se tomará $x_p = 35{,}5$, es decir los aglomerados que en el segundo semestre del 2018 mostraron un índice de pobreza del 35,5%, que como se puede ver en la tabla 4.1 corresponde a Gran Catamarca.

Luego se calcula $\hat{y}_p$, es el estimado del valor particular de y

$$\hat{y}_p = b_0 + b_1(x_p)$$

Reemplazando con los datos obtenidos anteriormente, se obtiene:

$$\hat{y}_p = 8{,}9031 + 0{,}8342(35{,}5) = \mathbf{38,5172}$$

La fórmula para estimar la desviación estándar de $\hat{y}_p$ dado un valor particular de x, x_p, es:

$$S_{\hat{y}p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}}}$$

Entonces, reemplazando los datos obtenemos

$$S_{\hat{y}p} = 4{,}3176\sqrt{\frac{1}{31} + \frac{(35{,}5 - 31{,}2065)^2}{1543{,}3387}}$$

$$S_{\hat{y}p} = 4{,}3176\sqrt{0{,}0323 + \frac{18{,}4341}{1543{,}3387}}$$

$$S_{\hat{y}p} = 4{,}3176\sqrt{0{,}0442}$$

$$S_{\hat{y}p} = 4{,}3176 * 0{,}2103$$

$$\boldsymbol{S_{\hat{y}p} = 0{,}9082}$$

De acuerdo con la tabla 8.5 "*Distribución t de student*" del capítulo 7. "*Cálculos y tablas auxiliares*", vemos que $\boldsymbol{t_{0{,}025} = 2{,}045}$.

Así, con $y_p = 38{,}5172$ y $S_{\hat{y}p} = 0{,}9082$, se calcula

$$y_p \pm t_{\alpha/2} S_{\hat{y}p}$$

$$38{,}5172 \pm 2{,}045 * 0{,}9082$$

$$\mathbf{38{,}5172 \pm 1{,}8573}$$

Entonces, con una confianza del 95% se puede decir que el porcentaje promedio de pobreza en el segundo semestre del 2019 de todos los aglomerados que en el segundo semestre del 2018 mostraron un índice de pobreza del 35,5% está entre el 36,6599% y el 40,3745%.

Obsérvese que el valor real $y_{x_p} = 39{,}6$ y $\hat{y}_{x_p} = 38{,}5172$, en ambos casos el valor esta dentro del intervalo predicho.

El segundo es el intervalo de predicción para un valor particular de y, se utilizará el mismo ejemplo, aglomerado de Gran Catamarca con un índice de pobreza del 35,5%, ya sabemos que $y_p = 38{,}5172$ y además debemos calcular la desviación estándar de $\hat{y}_p$:

$$S_{ind} = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{XX}}}$$

Para el caso que estamos trabajando:

$$S_{ind} = 4{,}3176\sqrt{1 + \frac{1}{31} + \frac{(35{,}5 - 31{,}2065)^2}{1543{,}3387}}$$

$$S_{ind} = 4{,}3176\sqrt{1 + 0{,}0323 + 0{,}0119}$$

$$S_{ind} = 4{,}3176\sqrt{1{,}0442}$$

$$S_{ind} = 4{,}3176 * 1{,}0219$$

$$\boldsymbol{S_{ind} = 4{,}4122}$$

De acuerdo con la tabla 7.5 "*Distribución t de student*" del capítulo 7. "*Cálculos y tablas auxiliares",* vemos que $\boldsymbol{t_{0{,}025} = 2{,}045.}$ Así, con $y_p = 38{,}5172$ y $S_{ind} = 4{,}4122$, tenemos:

$$38{,}5172 \pm 2{,}045 * 4{,}4122$$

$$38{,}5172 \pm 2{,}045 * 4{,}4122$$

$$\mathbf{38{,}5172 \pm 9{,}0229}$$

Entonces, con una confianza del 95% se puede decir que el porcentaje de pobreza en el segundo semestre del 2019 del aglomerado de Gran Catamarca, que en el segundo semestre del 2018 mostró un índice de pobreza del 35,5%, está entre el 29,4943% y el 47,5401%.

De acuerdo con lo anterior, el intervalo de predicción es mayor que el intervalo de confianza del valor medio de y, y lo contiene. Por lo tanto y_{x_p} y $\hat{y}_{x_p}$ también pertenecen al intervalo de predicción.

Por último, se calculan los intervalos de confianza para los coeficientes de la ecuación de regresión

$$\beta_0 \ = b_0 \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{XX}}} \qquad [\mathbf{5}]$$

$$\beta_1 = b_1 \pm t_{\alpha/2} \frac{S}{\sqrt{S_{XX}}} \qquad [\mathbf{6}]$$

Siguiendo con el análisis de regresión de la pobreza en Argentina, al realizar los cálculos de los intervalos de confianza de los parámetros del modelo reemplazando los datos en **[5]** se tiene:

$$\beta_0 = 8{,}9031 \pm 2{,}045 * 4{,}3176 \sqrt{\frac{1}{31} * \frac{973{,}8426}{1543{,}3387}}$$

$$\beta_0 = 8{,}9031 \pm 2{,}045 * 4{,}3176 \sqrt{0{,}0323 * 0{,}6310}$$

$$\beta_0 = 8{,}9031 \pm 2{,}045 * 4{,}3176 * 0{,}1428$$

$$\boldsymbol{\beta_0 = 8{,}9031 \pm 1{,}2608}$$

Y, por otro lado, reemplazando en **[6]**

$$\beta_1 = 0{,}8342 \pm 2{,}045 * \frac{4{,}3176}{\sqrt{1543{,}3387}}$$

$$\beta_1 = 0{,}8342 \pm 2{,}045 * \frac{4{,}3176}{39{,}2853}$$

$$\beta_1 = 0{,}8342 \pm 2{,}045 * \frac{4{,}3176}{39{,}2853}$$

$$\boldsymbol{\beta_1 = 0{,}8342 \pm 0{,}1099}$$

De esta manera, podemos decir que con una confianza del 95%, el coeficiente de la ecuación de regresión β_0 esta entre 7,6423 y 10,1639 y el valor de β_1 entre 0,7243 y 0,9441.

Conclusión

Habiendo analizado y comprendido el contenido del paper "*Aplicación de la regresión lineal en un problema de pobreza*" de Colombia, y replicando el análisis para los datos de pobreza en las personas de 31 aglomerados urbanos principales de Argentina obtenidos de la encuesta permanente de hogares del INDEC, se demostró que existe una relación lineal significativa entre los datos del segundo semestre del año 2018 y los del segundo semestre del año 2019.

Comenzó el análisis con un diagrama de dispersión donde se detectó una relación directa entre las variables y, además, se observó que los puntos parecen aproximarse a una línea recta. En consecuencia, se eligió el modelo de regresión lineal simple para representar la relación entre las variables.

La ecuación estimada de la regresión obtenida fue $\hat{Y} = 8{,}9031 + 0{,}8342x$, la pendiente de la recta es positiva, lo que implica que en los aglomerados donde se observó mayor pobreza en el segundo semestre del 2018, también se observó mayor pobreza en el segundo semestre del 2019. Pero como la pendiente es un número entre cero y uno, significa que el incremento en el porcentaje de pobreza en el segundo semestre del 2019 entre un aglomerado y otro es menor que en el segundo semestre del 2018.

En el análisis se reveló que la ecuación de regresión explica en un 66,51% los valores observados de la pobreza en segundo semestre del 2019 según los valores de pobreza en el segundo semestre del 2018. Se demostró además con un coeficiente de correlación de 0,8155 que la relación directa de dependencia lineal es fuerte. Esta relación se pudo confirmar con las pruebas de significancia t y F.

En el análisis de residuales, se concluyó en primer lugar que la varianza de ε no es constante. En la gráfica de residuales estandarizados, se observó que aproximadamente el 96,8% de los datos está dentro del intervalo y no habría razón suficiente para dudar de que el término de error tenga distribución normal. Y, por último, en la gráfica de distribución normal los puntos parecen ajustarse a una línea recta, lo cual indica que los datos provienen de una distribución normal.

En el análisis de outliers se encontró que un 77,4% de los e_{zi} está dentro del rango (-1,1) y un 96,8% dentro de (-2,2), quedando un solo residual estandarizado con valor -3,2318, pero, al no haber error en la medición, el dato debe conservarse.

Continuando con el análisis, se realizó la prueba de observaciones influyentes descartando ciertos valores en el análisis de regresión de la pobreza en argentina, y en ninguno de los casos cambio la ecuación de regresión, ni el coeficiente de regresión, por lo cual se concluyó que no se detectan observaciones influyentes.

Como pudimos validar que la relación es estadísticamente significativa entre las variables, y que el ajuste que proporciona la ecuación es bueno, la ecuación puede usarse para estimaciones y predicciones. Por este motivo se calcularon los intervalos de confianza del valor medio de y, y el intervalo de predicción y se comprobó su validez con algunas observaciones. Así como también se calcularon los intervalos de confianza para los coeficientes de la ecuación de regresión.

Referencias

Paul L. Meyer. "*Probabilidad y aplicaciones estadísticas*". Ed. Addison Wesley Iberoamericana.

Murray R. Spiegel, John J. Schiller, R. Alu Srinivasan. "*Probabilidad y estadística. Segunda edición*". Traducido de la segunda edición en inglés de: *SCHAUM'S OUTLINE OF THEORY AND PROBLEMS OF PROBABILITY AND STATISTICS*. McGraw – Hill Interamericana Editores, S.A. México D.F. 2003

Devore, Jay. Probabilidad y estadística para ingeniería y ciencias. 2001 5ta edición. International Thomson Editores SA. México

Antonio Estepa Castro, María Magdalena Gea Serrano, Gustavo R. Cañadas de la Fuente, José Miguel Contreras García. "*Algunas notas históricas sobre la correlación y regresión y su uso en el aula*". Números: Revista de didáctica de las matemáticas Volumen 81, España 2012.

Gasparini, L., Tornarolli, L. y Gluzmann, P. (2019). "*El desafío de la pobreza en Argentina. Diagnóstico y perspectivas"*. Buenos Aires: CEDLAS, CIPPEC, PNUD.

Antonio Rustom J. "*ESTADÍSTICA DESCRIPTIVA, PROBABILIDAD E INFERENCIA*. *Una visión conceptual y aplicada"*. Departamento de Economía Agraria Facultad de Ciencias Agronómicas Universidad de Chile. Santiago de Chile 2012

RONALD E. WALPOLE, RAYMOND H. MYERS, SHARON L. MYERS Y KEYING YE. "*Probabilidad y estadística para ingeniería y ciencias*". Novena edición PEARSON EDUCACIÓN, México, 2012

María Eugenia Szretter Noste. "*Apunte de Regresión Lineal*". Carrera de Especialización en Estadística para Ciencias de la Salud, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires 2017

INDEC, "*Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos. Segundo semestre de 2019*". Buenos aires, abril 2020

Cálculos y tablas auxiliares

Tabla 7.1. Porcentaje de pobreza en personas de 31 aglomerados urbanos de Argentina

Dominio	Pobreza Personas	
	2018 x_i	2019 y_i
Ciudad Autónoma de Buenos Aires	12,6	13,5
Partidos del GBA	35,9	40,5
Gran Mendoza	30,7	38,6
Gran San Juan	33,1	32,3
Gran San Luis	31,3	35
Corrientes	49,3	37,9
Formosa	32,5	41,6
Posadas	35,7	41,3
Gran Catamarca	35,5	39,6
Gran Tucumán - Tafí Viejo	32,2	37,3
Jujuy - Palpalá	31,7	37,8
La Rioja	30,5	38
Salta	37,7	45,5
Santiago del Estero - La Banda	38,9	45,2
Bahía Blanca - Cerri	25	28,1
Concordia	41,9	51,1
Gran Córdoba	36,5	37,4
Gran La Plata	30,9	30,6
Gran Rosario	31,8	35
Gran Paraná	29,4	30
Gran Santa Fe	34,4	34,4
Mar del Plata	24,8	25
Río Cuarto	29,3	29,4
Santa Rosa - Toay	32,1	33,9
San Nicolás - Villa Constitución	33,1	39,6
Comodoro Rivadavia - Rada Tilly	22	26,2
Neuquen - Plottier	27,1	28,6
Río Gallegos	18,6	25,7
Ushuaia - Río Grande	17,9	31,5
Rawson - Trelew	35,3	39,5
Viedma - Carmen de Patagones	29,7	32,9
	967,4000	**1083,0000**

Tabla 7.2. Cálculos auxiliares para la obtención de los regresores b_0 y b_1

Pobreza Personas		Cálculos Auxiliares						
2018 x_i	2019 y_i	$x_i y_i$	${x_i}^2$	${y_i}^2$	$\widehat{y_i}$	$(y_i - \widehat{y_i})^2$	$(y_i - \bar{Y})^2$	$(\widehat{y_i} - \bar{Y})^2$
12,6	13,5	170,1000	158,7600	182,2500	19,4140	34,9756	459,4800	240,9158
35,9	40,5	1453,9500	1288,8100	1640,2500	38,8509	2,7196	30,9638	15,3303
30,7	38,6	1185,0200	942,4900	1489,9600	34,5130	16,7032	13,4287	0,1785
33,1	32,3	1069,1300	1095,6100	1043,2900	36,5151	17,7672	6,9458	2,4953
31,3	35	1095,5000	979,6900	1225,0000	35,0136	0,0002	0,0042	0,0061
49,3	37,9	1868,4700	2430,4900	1436,4100	50,0292	147,1165	8,7884	227,8191
32,5	41,6	1352,0000	1056,2500	1730,5600	36,0146	31,1967	44,4158	1,1645
35,7	41,3	1474,4100	1274,4900	1705,6900	38,6840	6,8432	40,5071	14,0517
35,5	39,6	1405,8000	1260,2500	1568,1600	38,5172	1,1725	21,7577	12,8287
32,2	37,3	1201,0600	1036,8400	1391,2900	35,7643	2,3583	5,5909	0,6870
31,7	37,8	1198,2600	1004,8900	1428,8400	35,3472	6,0160	8,2055	0,1695
30,5	38	1159,0000	930,2500	1444,0000	34,3462	13,3503	9,3913	0,3473
37,7	45,5	1715,3500	1421,2900	2070,2500	40,3524	26,4974	111,6090	29,3434
38,9	45,2	1758,2800	1513,2100	2043,0400	41,3535	14,7957	105,3603	41,1907
25	28,1	702,5000	625,0000	789,6100	29,7581	2,7493	46,7238	26,8053
41,9	51,1	2141,0900	1755,6100	2611,2100	43,8561	52,4744	261,2916	79,5770
36,5	37,4	1365,1000	1332,2500	1398,7600	39,3514	3,8080	6,0738	19,5003
30,9	30,6	945,5400	954,8100	936,3600	34,6799	16,6454	18,7964	0,0653
31,8	35	1113,0000	1011,2400	1225,0000	35,4307	0,1855	0,0042	0,2452
29,4	30	882,0000	864,3600	900,0000	33,4286	11,7552	24,3590	2,2708
34,4	34,4	1183,3600	1183,3600	1183,3600	37,5996	10,2373	0,2867	7,0974
24,8	25	620,0000	615,0400	625,0000	29,5913	21,0797	98,7138	28,5607
29,3	29,4	861,4200	858,4900	864,3600	33,3452	15,5643	30,6416	2,5291
32,1	33,9	1088,1900	1030,4100	1149,2100	35,6809	3,1717	1,0722	0,5557
33,1	39,6	1310,7600	1095,6100	1568,1600	36,5151	9,5165	21,7577	2,4953
22	26,2	576,4000	484,0000	686,4400	27,2555	1,1141	76,3087	58,9822
27,1	28,6	775,0600	734,4100	817,9600	31,5099	8,4676	40,1384	11,7345
18,6	25,7	478,0200	345,9600	660,4900	24,4192	1,6404	85,2942	110,5918
17,9	31,5	563,8500	320,4100	992,2500	23,8353	58,7479	11,8025	123,2145
35,3	39,5	1394,3500	1246,0900	1560,2500	38,3504	1,3217	20,8348	11,6614
29,7	32,9	977,1300	882,0900	1082,4100	33,6788	0,6066	4,1432	1,5792
967,4000	**1083,0000**	**35084,1000**	**31732,4600**	**39449,8200**	**1083,0012**	**540,5979**	**1614,6910**	**1073,9934**

Tabla 7.3. Cálculos auxiliares para el análisis de residuales.

Dominio	Pobreza Personas		Residuales				
	2018 (x)	2019 (y)	$(x_i - \bar{x})^2$	h_i	s_{ei}	$y_i - \widehat{y_i}$	e_{zi}
Ciudad Autónoma de Buenos Aires	12,6	13,5	346,2000	0,2566	3,7227	-5,9140	-1,5886
Partidos del GBA	35,9	40,5	22,0294	0,0465	4,2159	1,6491	0,3912
Gran Mendoza	30,7	38,6	0,2565	0,0324	4,2470	4,0870	0,9623
Gran San Juan	33,1	32,3	3,5855	0,0346	4,2422	-4,2151	-0,9936
Gran San Luis	31,3	35	0,0088	0,0323	4,2473	-0,0136	-0,0032
Corrientes	49,3	37,9	327,3765	0,2444	3,7531	-12,1292	-3,2318
Formosa	32,5	41,6	1,6733	0,0333	4,2450	5,5854	1,3158
Posadas	35,7	41,3	20,1920	0,0453	4,2185	2,6160	0,6201
Gran Catamarca	35,5	39,6	18,4346	0,0442	4,2211	1,0828	0,2565
Gran Tucumán - Tafí Viejo	32,2	37,3	0,9871	0,0329	4,2459	1,5357	0,3617
Jujuy - Palpalá	31,7	37,8	0,2436	0,0324	4,2470	2,4528	0,5775
La Rioja	30,5	38	0,4991	0,0326	4,2466	3,6538	0,8604
Salta	37,7	45,5	42,1662	0,0596	4,1870	5,1476	1,2294
Santiago del Estero - La Banda	38,9	45,2	59,1907	0,0706	4,1623	3,8465	0,9241
Bahía Blanca - Cerri	25	28,1	38,5200	0,0572	4,1922	-1,6581	-0,3955
Concordia	41,9	51,1	114,3520	0,1064	4,0815	7,2439	1,7748
Gran Córdoba	36,5	37,4	28,0217	0,0504	4,2073	-1,9514	-0,4638
Gran La Plata	30,9	30,6	0,0939	0,0323	4,2472	-4,0799	-0,9606
Gran Rosario	31,8	35	0,3523	0,0325	4,2468	-0,4307	-0,1014
Gran Paraná	29,4	30	3,2633	0,0344	4,2427	-3,4286	-0,8081
Gran Santa Fe	34,4	34,4	10,1988	0,0389	4,2328	-3,1996	-0,7559
Mar del Plata	24,8	25	41,0426	0,0589	4,1886	-4,5913	-1,0961
Río Cuarto	29,3	29,4	3,6346	0,0346	4,2422	-3,9452	-0,9300
Santa Rosa - Toay	32,1	33,9	0,7984	0,0328	4,2462	-1,7809	-0,4194
San Nicolás - Villa Constitución	33,1	39,6	3,5855	0,0346	4,2422	3,0849	0,7272
Comodoro Rivadavia - Rada Tilly	22	26,2	84,7588	0,0872	4,1251	-1,0555	-0,2559
Neuquen - Plottier	27,1	28,6	16,8629	0,0432	4,2233	-2,9099	-0,6890
Río Gallegos	18,6	25,7	158,9226	0,1352	4,0150	1,2808	0,3190
Ushuaia - Río Grande	17,9	31,5	177,0617	0,1470	3,9876	7,6647	1,9221
Rawson - Trelew	35,3	39,5	16,7571	0,0431	4,2235	1,1496	0,2722
Viedma - Carmen de Patagones	29,7	32,9	2,2694	0,0337	4,2441	-0,7788	-0,1835
Sumatorias	**967,4**	**1083,0**	**1543,3387**	**2,0000**	**129,3902**	**-0,0012**	**-0,3621**

Tabla 7.4. Cálculos auxiliares para la gráfica de probabilidad normal de los residuos.

Dominio	e_{zi}	i	P_e	F(residuo)
Corrientes	-3,2318	1	0,0161	0,0006
Ciudad Autónoma de Buenos Aires	-1,5886	2	0,0484	0,0559
Mar del Plata	-1,0961	3	0,0806	0,1357
Gran San Juan	-0,9936	4	0,1129	0,1611
Gran La Plata	-0,9606	5	0,1452	0,1685
Río Cuarto	-0,9300	6	0,1774	0,1762
Gran Paraná	-0,8081	7	0,2097	0,2090
Gran Santa Fe	-0,7559	8	0,2419	0,2236
Neuquen - Plottier	-0,6890	9	0,2742	0,2451
Gran Córdoba	-0,4638	10	0,3065	0,3228
Santa Rosa - Toay	-0,4194	11	0,3387	0,3372
Bahía Blanca - Cerri	-0,3955	12	0,3710	0,3446
Comodoro Rivadavia - Rada Tilly	-0,2559	13	0,4032	0,3974
Viedma - Carmen de Patagones	-0,1835	14	0,4355	0,4286
Gran Rosario	-0,1014	15	0,4677	0,4602
Gran San Luis	-0,0032	16	0,5000	0,5000
Gran Catamarca	0,2565	17	0,5323	0,6026
Rawson - Trelew	0,2722	18	0,5645	0,6064
Río Gallegos	0,3190	19	0,5968	0,6255
Gran Tucumán - Tafí Viejo	0,3617	20	0,6290	0,6406
Partidos del GBA	0,3912	21	0,6613	0,6517
Jujuy - Palpalá	0,5775	22	0,6935	0,7190
Posadas	0,6201	23	0,7258	0,7324
San Nicolás - Villa Constitución	0,7272	24	0,7581	0,7673
La Rioja	0,8604	25	0,7903	0,8051
Santiago del Estero - La Banda	0,9241	26	0,8226	0,8212
Gran Mendoza	0,9623	27	0,8548	0,8315
Salta	1,2294	28	0,8871	0,8907
Formosa	1,3158	29	0,9194	0,9066
Concordia	1,7748	30	0,9516	0,9616
Ushuaia - Río Grande	1,9221	31	0,9839	0,9726

Tabla 7.5. Distribución t de student

α / r	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,689
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,660
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,290

Tabla 7.6. Distribución F de Fisher con probabilidad de 0,05

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	70	80	100	120
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95	248.02	249.05	250.10	251.14	251.77	252.20	252.50	252.72	253.04	253.25
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.412	19.429	19.446	19.454	19.463	19.471	19.476	19.48	19.48	19.48	19.49	19.49
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.745	8.703	8.660	8.638	8.617	8.594	8.581	8.572	8.566	8.561	8.554	8.549
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803	5.774	5.746	5.717	5.699	5.688	5.679	5.673	5.664	5.658
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558	4.527	4.496	4.464	4.444	4.431	4.422	4.415	4.405	4.398
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874	3.841	3.808	3.774	3.754	3.74	3.73	3.722	3.712	3.705
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445	3.410	3.376	3.340	3.319	3.304	3.294	3.286	3.275	3.267
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150	3.115	3.079	3.043	3.020	3.005	2.994	2.986	2.975	2.967
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936	2.900	2.864	2.826	2.803	2.787	2.776	2.768	2.756	2.748
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774	2.737	2.700	2.661	2.637	2.621	2.609	2.601	2.588	2.580
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646	2.609	2.570	2.531	2.507	2.490	2.478	2.469	2.457	2.448
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544	2.505	2.466	2.426	2.401	2.384	2.372	2.363	2.350	2.341
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459	2.420	2.380	2.339	2.314	2.297	2.284	2.275	2.261	2.252
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388	2.349	2.308	2.266	2.241	2.223	2.210	2.201	2.187	2.178
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328	2.288	2.247	2.204	2.178	2.160	2.147	2.137	2.123	2.114
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276	2.235	2.194	2.151	2.124	2.106	2.093	2.083	2.068	2.059
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230	2.190	2.148	2.104	2.077	2.058	2.045	2.035	2.020	2.011
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191	2.150	2.107	2.063	2.035	2.017	2.003	1.993	1.978	1.968
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155	2.114	2.071	2.026	1.999	1.980	1.966	1.955	1.940	1.930

Continúa

20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124	2.082	2.039	1.994	1.966	1.946	1.932	1.922	1.907	1.896
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.176	2.096	2.054	2.010	1.965	1.936	1.916	1.902	1.891	1.876	1.866
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.151	2.071	2.028	1.984	1.938	1.909	1.889	1.875	1.864	1.849	1.838
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.128	2.048	2.005	1.961	1.914	1.885	1.865	1.850	1.839	1.823	1.813
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.108	2.027	1.984	1.939	1.892	1.863	1.842	1.828	1.816	1.800	1.790
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.089	2.007	1.964	1.919	1.872	1.842	1.822	1.807	1.796	1.779	1.768
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	2.072	1.990	1.946	1.901	1.853	1.823	1.803	1.788	1.776	1.76	1.749
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	2.056	1.974	1.930	1.884	1.836	1.806	1.785	1.770	1.758	1.742	1.731
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	2.041	1.959	1.915	1.869	1.820	1.790	1.769	1.754	1.742	1.725	1.714
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	2.027	1.945	1.901	1.854	1.806	1.775	1.754	1.738	1.726	1.71	1.698
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932	1.887	1.841	1.792	1.761	1.740	1.724	1.712	1.695	1.683
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114	2.041	1.963	1.878	1.833	1.786	1.735	1.703	1.681	1.665	1.652	1.635	1.623
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.924	1.839	1.793	1.744	1.693	1.660	1.637	1.621	1.608	1.589	1.577
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049	1.974	1.895	1.808	1.762	1.713	1.660	1.626	1.603	1.586	1.573	1.554	1.541
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.952	1.871	1.784	1.737	1.687	1.634	1.599	1.576	1.558	1.544	1.525	1.511
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.836	1.748	1.700	1.649	1.594	1.559	1.534	1.516	1.502	1.481	1.467
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969	1.893	1.812	1.722	1.674	1.622	1.566	1.530	1.505	1.486	1.471	1.45	1.435
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951	1.875	1.793	1.703	1.654	1.602	1.545	1.508	1.482	1.463	1.448	1.426	1.411
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938	1.861	1.779	1.688	1.639	1.586	1.528	1.491	1.465	1.445	1.429	1.407	1.391
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.850	1.768	1.676	1.627	1.573	1.515	1.477	1.450	1.430	1.415	1.392	1.376
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.750	1.659	1.608	1.554	1.495	1.457	1.429	1.408	1.392	1.369	1.352

Fuente: López, 2010.

Continúa

Tabla 7.7. Distribución normal estandarizada

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0160	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,5160	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0855	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1631	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,7210	0,4681	0,4641

(Continua)

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9989	0,9990	0,9990

Fuente: ebookbrowse, 2012.

Tabla 7.8. Distribución t de student

v	0,6	0,75	0,9	0,95	0,975	0,99	0,995	0,9975	0,999	0,9995
1	0,325	1,000	3,078	6,314	12,706	31,821	63,656	127,321	318,289	636,578
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,328	31,600
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,214	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	5,894	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	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
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	3,660
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
∞	0,253	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,290