



MODELO DE REGRESIÓN

Abstract

Se presenta un apunte de cátedra de **Regresión**, un modelo estadístico utilizado en la Minería de Datos, y su aplicación con R, una importante herramienta de software libre. El desarrollo del apunte es clásico. Comenzamos con la presentación del modelo, analizamos los supuestos que lo sustentan y finalmente desarrollamos métodos de predicción.

Palabras clave

Regresión. Minería de Datos, Data Mining. Coeficiente de determinación. Intervalo de predicción. Intervalo de confianza, The R Project for Statistical Computing.

Introducción

La Minería de Datos (Data Mining) se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Una tarea es un tipo de problema de Minería de Datos. Existen dos tipos de tareas: predictivas y descriptivas. En las primeras debemos predecir uno o más valores para uno o más ejemplos. En las segundas el objetivo es describir los datos existentes. La regresión es una tarea predictiva en la que, dados dos conjuntos de datos E y S (entrada y salida, respectivamente), debemos encontrar una función que represente la correspondencia existente entre los ejemplos, es decir, para cada valor de E tenemos un único valor para S . En este documento trataremos en profundidad el concepto y las aplicaciones de la regresión lineal simple como modelo particular de regresión.

Usaremos como herramienta para cálculo y gráficos el **R**, un entorno para computación estadística y gráficos, con algunas aplicaciones de Minería de Datos, de licencia GNU GPL. Empleamos asimismo dos paquetes adicionales de R para nuestro propósito: **R Commander**, una interfaz gráfica para R, y **SimpleR**, un conjunto de funciones que incluyen las necesarias para intervalos, este último desarrollado por el Departamento de Matemática del College of Staten Island.



Modelización estadística

El objetivo de la modelización estadística es explicar el comportamiento de una variable a partir del conocimiento de otras. Como su nombre lo indica una *variable* tiene *variabilidad*, y esta variabilidad puede relacionarse con el comportamiento de otras variables, por ejemplo, el saldo total bancario de las personas de una cierta edad, con un mismo nivel profesional, residentes en la misma localidad no es igual para todos sino que sigue una cierta distribución. Si conocemos la edad de una persona, su nivel profesional y su lugar de residencia podremos aproximar su saldo bancario.

A la variable bajo estudio, en este ejemplo el saldo bancario, se le denomina variable de salida (*output*), explicada, de respuesta o endógena y se denota por la letra Y , mientras que las variables edad, nivel profesional, localidad, se denominan variables de entrada (*input*), predictoras, explicativas, regresores o exógenas y se denotan por x_j . En el presente documento, utilizaremos la terminología de variable de respuesta y variables explicativas para referirnos a ambos conjuntos de variables.

La modelización estadística consiste en descomponer los valores que toma la variable de respuesta Y en dos componentes, uno función de las variables explicativas y otro que es específico del valor en cuestión:

$$Y = r(x_1, \dots, x_n) + \varepsilon \quad [1]$$

donde r es la función que relaciona los valores de la variable de respuesta con las explicativas, mientras que ε es la variabilidad aleatoria de Y que no se explica por las variables x_j .

La función r representa pues la parte determinista, estructural del modelo, que explica parte del comportamiento de la variable de respuesta, mientras que la segunda componente representa la parte impredecible, aleatoria y se denomina término de error. Ambas permiten elaborar predicciones. La distribución de probabilidad del término de error puede suponerse, sin pérdida de generalidad, centrada en 0, es decir:

$$E[\varepsilon_i] = E[Y_i - r(x_{i1}, \dots, x_{in})] = 0 \quad [2]$$

Los modelos estadísticos propuestos en este documento se caracterizan según el tipo de variables explicativas, que pueden ser numéricas o categóricas, respecto de la función r y de la distribución de probabilidad del error aleatorio.



Regresión Lineal Simple

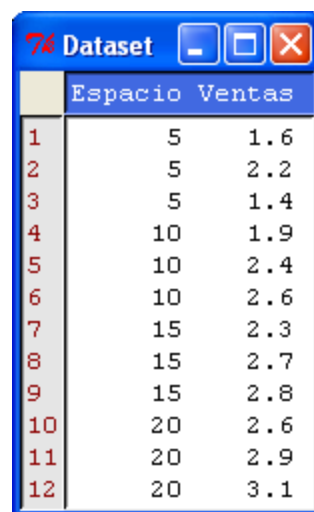
Introducción

Hablamos de *modelo de regresión* cuando la variable de respuesta y las variables explicativas son cuantitativas. Si solo disponemos de una variable explicativa es *regresión simple*, mientras que si disponemos de varias variables explicativas se trata de una *regresión múltiple*.

Comenzaremos tratando la regresión simple. Dado que la variable aleatoria X toma un valor específico, esperamos una respuesta en la variable aleatoria Y . Es decir, el valor que toma X influye en el valor de Y . O alternativamente, Y depende de X .

Es importante destacar que el modelo de regresión requiere que la variable x sea determinística. Sin embargo, es posible extender la validez del modelo para el caso de una variable X aleatoria, tal como explicaremos más adelante.

Para ilustrar la idea mencionada consideremos el siguiente ejemplo, que muestra los valores de las ventas semanales de comidas para mascotas (en cientos de pesos), y el espacio (medido en decímetros del frente de la góndola) de los estantes de 12 supermercados de igual tamaño.



	Espacio	Ventas
1	5	1.6
2	5	2.2
3	5	1.4
4	10	1.9
5	10	2.4
6	10	2.6
7	15	2.3
8	15	2.7
9	15	2.8
10	20	2.6
11	20	2.9
12	20	3.1

Figura 1. Venta de alimentos para mascotas

El objetivo del análisis de regresión es encontrar un modelo para esta relación. Para muchos problemas, es razonable asumir inicialmente un modelo lineal, al menos en el rango estudiado, verificándolo visualmente con los datos muestrales.

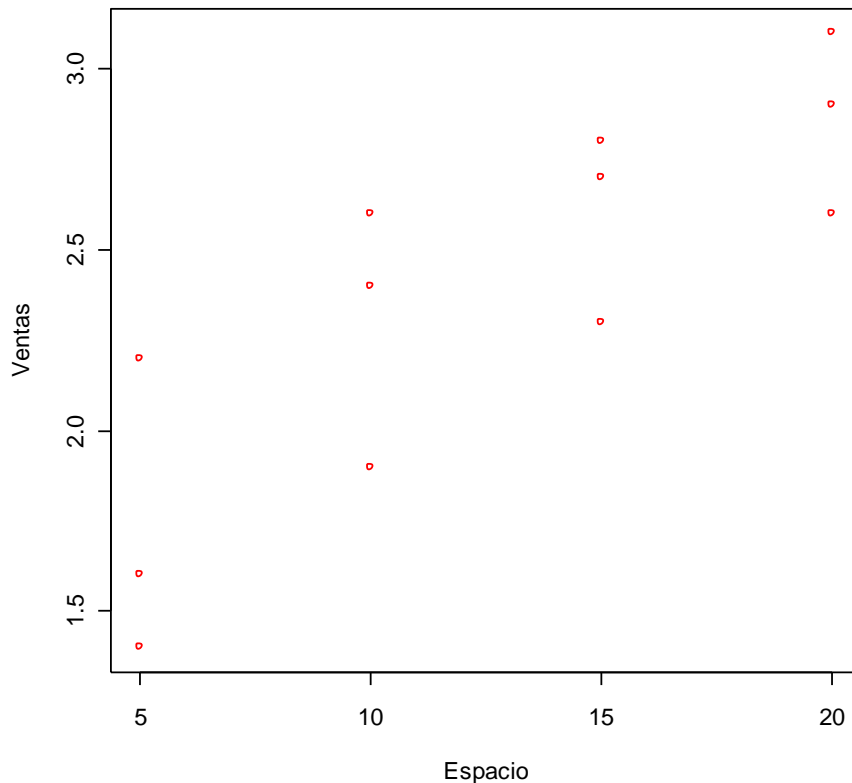


Figura 2. Nube de puntos de los datos
sobre espacio en estantes – ventas de alimentos para mascotas

Reemplazando en [1] el modelo de la regresión lineal simple es:

$$Y = \alpha + \beta x + \varepsilon$$

Ahora quizás estemos interesados en pronosticar el valor aproximado que tomará la variable aleatoria Y , cuando la variable x toma un valor específico. Por ejemplo, conocer el valor de las ventas en un supermercado en que el espacio dedicado al alimento es de 12 dm. En la realidad, la relación entre las dos variables no es exacta, luego no es razonable pensar en un único valor para las ventas, dado un valor particular para el espacio en estantes. Por el contrario, es más realista concebir, para cada posible longitud del estante, una *distribución* de los valores de ventas resultantes, una *distribución condicional* de las ventas de alimentos para mascotas cuando el espacio en estantes toma un valor específico, por ejemplo, 12 dm.

Parámetros del modelo

Nos preguntamos cuál es el **valor esperado** de las ventas de alimentos para mascotas en supermercados que dedican 12 dm de espacio en estantes. Denotaremos por $E(Y/X = x_i)$ el valor esperado de la variable aleatoria Y cuando la variable x toma el valor específico x_i . Nuestro supuesto de linealidad se traduce en una esperanza condicional lineal en x :



$$E(Y/X = x_i) = \alpha + \beta x_i$$

[3]

donde las constantes α y β son los parámetros del modelo. Las dos tienen una interpretación sencilla, una extremadamente importante, la otra no tanto.

Como ejemplo supongamos que las ventas promedio de alimentos para mascotas están relacionadas con el espacio en estantes mediante la ecuación

$$E(Y/X = x_i) = 1,5 + 0,07x_i; \quad x_i = 5, \dots, 20$$

[4]

luego, en la ecuación **[4]**, $\alpha = 1,5$ y $\beta = 0,07$. Sustituyendo $x = 0$ en **[3]** se obtiene

$$E(Y/X = 0) = \alpha$$

[5]

El parámetro α es el valor esperado de la **variable dependiente** Y cuando la **variable independiente** x toma el valor 0. Este número no siempre es valioso. En nuestro ejemplo si el espacio dedicado a la exhibición de comida para mascotas en estantes fuese de 0 dm (no exhibimos comida para mascotas), se esperaría que las ventas fuesen de 150 pesos. En realidad, si no exhibimos el producto es de esperar que las ventas sean nulas. **No conviene extender el supuesto de linealidad fuera del rango estudiado.** Contamos con observaciones del espacio en estantes en supermercados en un rango de 5 a 20 dm, y aunque la linealidad en este tramo parece razonable, sería peligroso extrapolar nuestras conclusiones fuera de este intervalo. El modelo de regresión es útil para interpolar, peligroso para extrapolar.

Volvamos a la ecuación

[3]. Supongamos que x se incrementa en una unidad, de x_i a $(x_i + 1)$. Entonces, tenemos

$$E(Y/X = x_i + 1) = \alpha + \beta(x_i + 1)$$

[6]

luego

$$E(Y/X = x_i + 1) - E(Y/X = x_i) = \alpha + \beta(x_i + 1) - (\alpha + \beta x_i) = \beta$$

[7]

El parámetro β , la pendiente de la recta, es el incremento esperado de Y para un incremento unitario de x . En nuestro ejemplo, para un incremento de un decímetro en el espacio en estan-



tes, se espera un aumento promedio de 0,07 cientos de pesos = 7 pesos en las ventas de alimentos para mascotas.

Ya hemos visto que el objetivo de la regresión es describir la dependencia de una variable aleatoria respecto de otra. Una manera de entender esta dependencia es en términos del cambio de la variable dependiente, Y , producido por un cambio en la variable independiente, x . En cada caso, la magnitud del cambio esperado en la variable dependiente es un múltiplo β del cambio en la variable independiente.

Los puntos de la Figura 3 no están alineados, ni se ajustan exactamente a ninguna otra curva que pudiéramos dibujar. Sin embargo, observamos una tendencia lineal, que el modelo reflejará en la esperanza. Supongamos que la variable independiente toma el valor x_i . Entonces

$$E(Y_i / X = x_i) = \alpha + \beta x_i$$

[8]

En la práctica, el valor observado de Y_i se desviará, casi inevitablemente, de su valor esperado. Si la diferencia se representa mediante la variable aleatoria ε_i , podemos escribir

$$\varepsilon_i = Y_i - E(Y_i / X = x_i) = Y_i - (\alpha + \beta x_i)$$

[9]

En virtud de la ecuación

[8], esta diferencia entre el valor esperado y el observado tendrá media 0. Podemos reescribir la [9] como sigue:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

[10]

La ecuación [10] se denomina **recta de regresión poblacional**. En virtud de ella, la diferencia entre el valor esperado y el observado tendrá media 0. Por tanto, el volumen de ventas de alimentos para mascotas para un valor x_i , del espacio en estantes, será la suma de dos partes: una esperanza $(\alpha + \beta x_i)$, reflejando su relación promedio, y una discrepancia ε_i de la esperanza. Se puede pensar en la discrepancia, o **término de error** ε_i , como en la componente que engloba la multitud de factores, *distintos del espacio en estantes*, que influyen en las ventas de alimentos para mascotas. Este término de error es muy importante y forma parte del modelo. No podemos pensar en un modelo de regresión sin considerar el error aleatorio.

El modelo de regresión que acabamos de describir se ilustra en la Figura 4, la recta representa la relación lineal entre el valor esperado de la variable dependiente y el valor que toma la



variable independiente. Para cada posible valor de la variable independiente, el valor de la variable dependiente se representa mediante una variable aleatoria cuya media está sobre la recta de regresión. En la figura dibujamos una serie de funciones de densidad para la variable dependiente, dados algunos valores de la variable independiente. Para un valor dado x_i , la desviación de la variable dependiente Y respecto de la recta de regresión es el término de error ε_i . Las funciones de densidad dibujadas en la Figura 4 corresponden a las variables aleatorias ε_i .

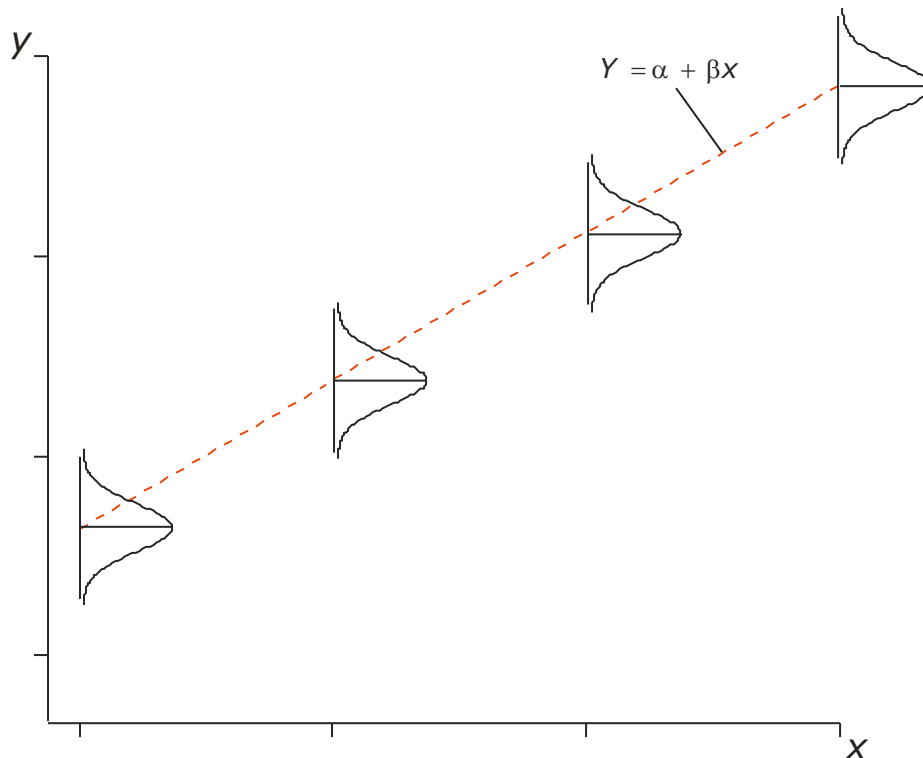


Figura 4. Modelo de regresión poblacional. Se representan funciones de densidad de la variable dependiente para x valores dados de la variable independiente

La figura muestra las medidas de posición y de dispersión del modelo propuesto. Para describir un conjunto de datos univariados (una sola variable, X) utilizamos medidas de resumen, al menos una de posición, por ejemplo el promedio $E(X)=\mu$, y otra de dispersión como la varianza σ^2 . Por analogía, en nuestro modelo multivariado (dos variables, X e Y) presentamos también medidas de resumen, de posición $E(Y/X)=(\alpha + \beta x_i)$ y de dispersión σ_e^2 , (error aleatorio). Nos extenderemos sobre esta última medida más adelante.

Estimación por mínimos cuadrados

La recta de regresión poblacional introducida en la sección anterior es una valiosa construcción teórica. Sin embargo, en aplicaciones prácticas, nunca seremos capaces de determinar los parámetros, sólo podremos estimarlos a partir de los datos disponibles.



Supongamos que disponemos de n pares de observaciones $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Nos gustaría encontrar la recta que se ajusta mejor a estos datos. En otras palabras, deseamos estimar los coeficientes desconocidos α y β de la recta de regresión poblacional. Un procedimiento obvio sería dibujar, a mano alzada una recta que pase razonablemente cerca de todos los puntos. Existen procedimientos formales, más precisos, para hallar la recta de regresión. Usaremos el *Método de los Mínimos Cuadrados*.

Consideremos, como posibles estimaciones de α y β los números a y b . La recta estimada es, entonces

$$y = a + bx$$

[11]

Para determinar la recta, necesitamos definir una medida de la distancia de los puntos (x_i, y_i) a la misma. La **Figura 5** muestra, para un solo punto, cómo se mide esta distancia.

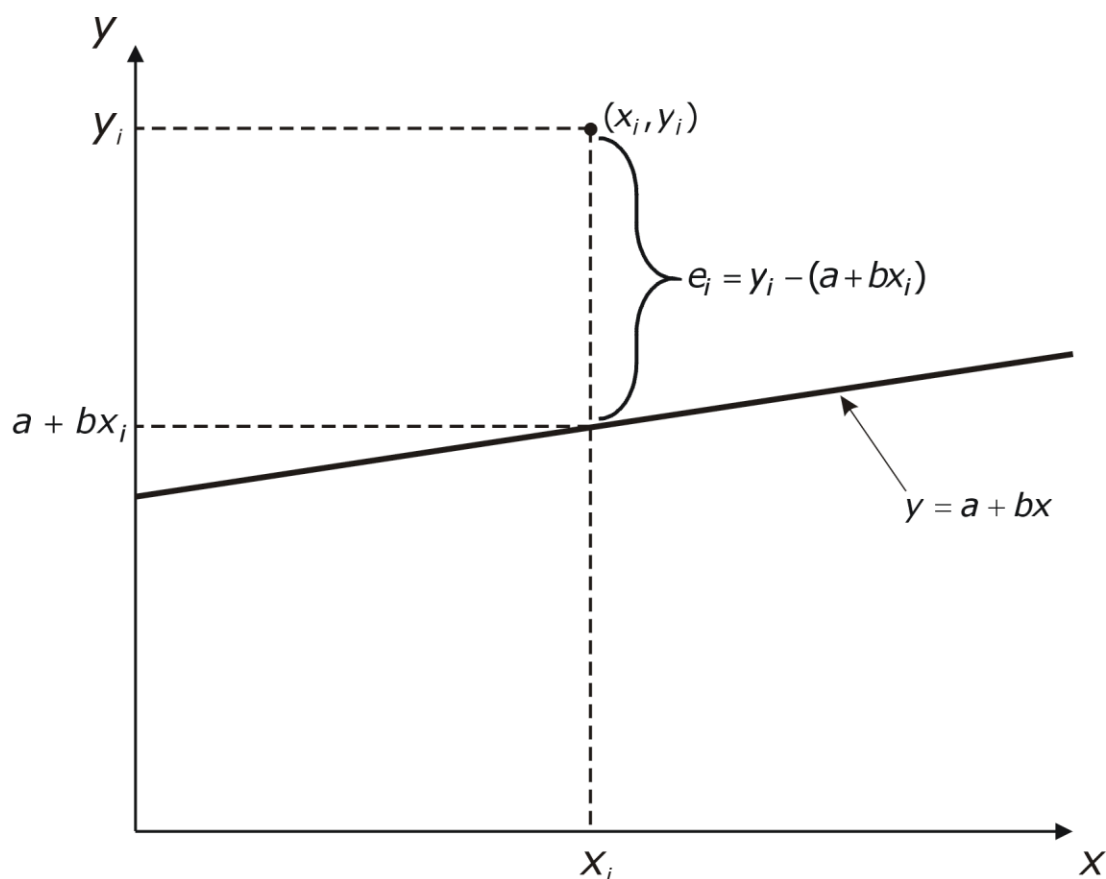


Figura 5. Distancia $e_i = y_i - (a + bx_i)$ del punto (x_i, y_i) a la recta $y = a + bx$

Para el valor x_i , el correspondiente valor y_i en nuestra recta es $\alpha + \beta x_i$, mientras que el valor realmente observado para la variable dependiente es y_i . La diferencia entre ambos es



$$\varepsilon_i = y_i - (\alpha + \beta x_i)$$

[12]

A primera vista, puede parecer sorprendente que no tomemos la distancia más corta entre el punto y la recta. La discrepancia o error ε_i refleja la desviación de la variable dependiente del valor $\alpha + \beta x_i$ predicho por la recta. Recordemos que la variable explicativa x no es aleatoria, sino que toma algunos valores predeterminados sobre los cuales se calculará el error. Para ilustrar estas ideas, consideremos las primeras observaciones del espacio disponible en estantes y las ventas de alimentos para mascotas de la Figura 1. Estas son

$$x_i = 5 \qquad y_i = 1,6$$

Consideremos también la posible recta

$$y = 1,5 + 0,07x$$

es decir, la recta con valores

$$a = 1,5 \qquad b = 0,07$$

para la constante y la pendiente. El valor promedio predicho por esta recta para las ventas de alimentos para mascotas, cuando el espacio es de 5 decímetros, es

$$a + bx_1 = 1,5 + (0,07)(5) = 1,85$$

La diferencia e_{ij} entre el verdadero valor y_{ij} y el valor predicho es el *residual*. Para el ejemplo, es

$$e_{11} = y_{11} - (a + bx_1) = 1,6 - 1,85 = -0,25$$

Cualquier estimación razonable de la recta de regresión verdadera dejará algunos de los datos observados por debajo y otros por encima de ella. Por tanto, algunos de los ε_{ij} de la ecuación [12] serán positivos y otros negativos. Podríamos penalizar por igual a todos los errores, o si las características del problema son particulares, utilizar ponderaciones distintas para cada error. El primer caso corresponde al *método de mínimos cuadrados*, tal como lo explicamos. Para el segundo caso existe el *método de mínimos cuadrados ponderados*. En este documento se utilizará el primer método.



Método de Mínimos Cuadrados

Si queremos penalizar por igual los todos los errores observados, una posibilidad es trabajar con los *cuadrados de ε_{ij}* . La suma de los errores al cuadrado entre los puntos y la recta es

$$SC = \sum_i \sum_j \varepsilon_{ij}^2 = \sum_i \sum_j (y_{ij} - \alpha - \beta x_i)^2$$

El **método de mínimos cuadrados** selecciona, como estimación de la recta de regresión poblacional, aquella para la cual esta suma de cuadrados es mínima.

Estimación por mínimos cuadrados y recta de regresión muestral¹

Sea $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra de n pares de observaciones de un proceso cuya recta de regresión poblacional es

$$Y_i = \alpha + \beta x_i$$

Las **estimaciones de mínimos cuadrados** de los coeficientes α y β son los valores a y b para los cuales se minimiza la suma de los errores al cuadrado

$$\min \left\{ SC = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \right\}$$

Los estimadores resultantes, a y b , se obtienen de la siguiente manera: se debe hallar el mínimo de SC , es decir, el par de valores (a, b) para los cuales las primeras derivadas parciales de SC sean iguales a 0. Estas son las componentes del vector gradiente, que al ser nulas satisfacen la condición necesaria para la existencia de extremo relativo. En el caso de funciones cóncavas o convexas como la que nos ocupa, la condición necesaria es también suficiente.

$$\begin{cases} \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - a - bx_i)^2 = 0 \\ \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - a - bx_i)^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^n (y_i - a - bx_i)(-2) = 0 \\ \sum_{i=1}^n (y_i - a - bx_i)(-2)(x_i) = 0 \end{cases}$$

Dividimos miembro a miembro por (-2)

$$\begin{cases} \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \sum_{i=1}^n (y_i - a - bx_i)(x_i) = 0 \end{cases}$$

¹ El desarrollo a continuación corresponde al caso de un único valor de y por cada valor de x .



Para la primera ecuación

$$\sum_{i=1}^n (y_i - a - bx_i) = 0 \Leftrightarrow \sum_{i=1}^n y_i = na + b \sum_{i=1}^n x_i$$

Dividimos miembro a miembro por n

$$\frac{\sum_{i=1}^n y_i}{n} = \frac{na}{n} + \frac{b \sum_{i=1}^n x_i}{n} \Leftrightarrow \bar{y} = a + b\bar{x}$$

donde \bar{x} e \bar{y} son las respectivas medias muestrales. Por lo tanto

$$a = \bar{y} - b\bar{x}$$

[13]

Para la segunda ecuación

$$\sum_{i=1}^n (y_i - a - bx_i)(x_i) = 0 \Leftrightarrow \sum_{i=1}^n y_i x_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2$$

Teniendo en cuenta la expresión de a para la primera ecuación, reescribimos lo anterior

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

Distribuimos

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

Esto equivale a

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

De donde deducimos b



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

[14]

O bien, multiplicamos miembro a miembro por $\frac{n}{\frac{n}{n}}$ y nos queda la expresión más simplifi-

cada

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

Esto es equivalente a (no demostrado aquí)

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

La recta $y = a + bx$ se denomina *recta de regresión muestral de Y sobre X*.

La recta de regresión muestral de los datos sobre las ventas de alimentos para mascotas se obtiene usando los cálculos de la Figura 6, donde tenemos que

$$\begin{aligned} \sum_{i=1}^n x_i &= 150 & \sum_{i=1}^n y_i &= 28,5 \\ \sum_{i=1}^n x_i y_i &= 384 & \sum_{i=1}^n x_i^2 &= 2250 \end{aligned}$$

[15]



	x_i	y_i	$x_i y_i$	x_i^2
	5	1,6	8	25
	5	2,2	11	25
	5	1,4	7	25
	10	1,9	19	100
	10	2,4	24	100
	10	2,6	26	100
	15	2,3	34,5	225
	15	2,7	40,5	225
	15	2,8	42	225
	20	2,6	52	400
	20	2,9	58	400
	20	3,1	62	400
Sumas	150	28,5	384	2250

Figura 6. Cálculos para obtener la recta de regresión muestral de las ventas de alimentos para mascotas sobre el espacio disponible en estantes de supermercados

Por consiguiente, las medias muestrales son

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{150}{12} = 12,5$$

[16]

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{28,5}{12} = 2,375$$

Los estimadores de mínimos cuadrados de los coeficientes de la recta de regresión poblacional son

$$b = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{384 - 12 \times 12,5 \times 2,375}{2250 - 12 \times 12,5^2} = 0,074$$

[17]

$$a = \bar{y} - b \bar{x} = 2,375 - 0,074 \times 12,5 = 1,45$$

La recta de regresión muestral, o estimada, es, por tanto

$$y = 1,45 + 0,074x$$

La salida de R para el ejemplo es

```
Call:
lm(formula = Ventas ~ Espacio, data = Dataset)
```



```

Residuals:
Min        1Q        Median        3Q        Max
-0.4200   -0.2675    0.0550    0.2175    0.4100

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.45000     0.21783   6.657 5.66e-05 ***
Espacio       0.07400     0.01591   4.652 0.000906 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3081 on 10 degrees of freedom
Multiple R-Squared:  0.6839,    Adjusted R-squared:  0.6523
F-statistic: 21.64 on 1 and 10 DF,  p-value: 0.0009057
    
```

Figura 7. Salida de R para el modelo de regresión
del espacio disponible-venta de alimentos

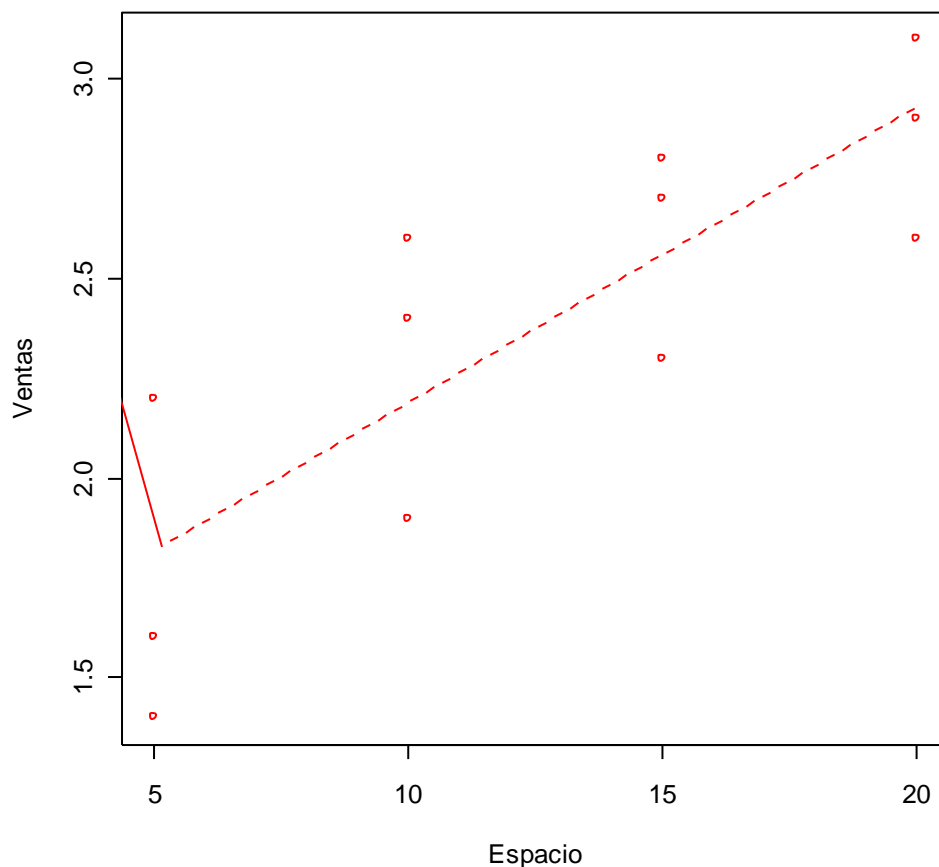


Figura 8. Datos del espacio disponible-venta de alimentos y recta de regresión estimada por
mínimos cuadrados, $y = 1,45 + 0,074x$



Recordando la interpretación de la pendiente de la recta de regresión, hemos estimado que un incremento de un decímetro en el espacio disponible en estantes produce, *en promedio*, un incremento de 7,40 pesos en las ventas de alimentos para mascotas. La Figura 8 muestra cómo la recta de regresión muestral ajusta a los doce puntos observados.

Estimación del desvío estándar del error

Una vez definidos estimadores para las medidas de resumen del modelo, debemos hacer una consideración importante. A menos que todos los datos observados estén sobre la recta de regresión, la ecuación de regresión no es un pronosticador perfecto. De la misma forma que no se espera que todos los valores sean idénticos a su media aritmética, tampoco puede pensarse que todos los datos estén justo sobre la recta de regresión. Por lo tanto, es necesario desarrollar un estadístico que mida la variabilidad de los valores de Y reales a partir de los valores de Y pronosticados, de la misma manera que se usa el desvío estándar como medida de variabilidad de cada observación univariada alrededor de la media.

Siguiendo con la analogía expuesta, la varianza de error, σ_e^2 , es una medida de dispersión del modelo de regresión. Ahora bien, ¿cómo estimar la varianza del error, del mismo modo que lo hicimos con la medida de posición? Una respuesta es generalizar a partir del modelo univariado. En él, estimamos σ^2 , la varianza poblacional, mediante s^2

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

[18]

que es sencillamente la suma de los cuadrados de los desvíos de las observaciones con respecto al promedio muestral, dividido por los grados de libertad del modelo univariado ($n-1$). En su análogo bivariado, el cálculo del estimador de la varianza del error, s_e^2 , se hace de un modo similar

$$s_e^2 = \frac{1}{n-2} \sum_i \sum_j [y_{ij} - (a + bx_i)]^2$$

[19]

donde de acuerdo a $\varepsilon_i = y_i - (\alpha + \beta x_i)$

[12] esta operación es en esencia la misma, consistiendo en la suma de los cuadrados de las discrepancias de las observaciones con respecto al promedio muestral $(a + bx_i)$, dividido por los grados de libertad del modelo, en este caso ($n-2$).



La raíz cuadrada de la varianza estimada, que usaremos para nuestro modelo de regresión y que mide la variabilidad alrededor de la recta de regresión se llama *desvío estándar del error*. Se representa con S_e y se define a continuación.

Varianza y Desvío estándar del error

Dada una recta de regresión poblacional

$$Y_i = \alpha + \beta X_i$$

y suponiendo que se verifican los supuestos mencionados más adelante (en

[33],

[34],

[35] y

[29]), un estimador insesgado de la varianza del error, σ_e^2 , se define como S_e^2 y se obtiene mediante

$$s_e^2 = \frac{1}{n-2} \sum_i \sum_j e_{ij}^2 = \frac{SCE}{n-2}$$

[20]

La raíz de este estimador se define como desvío estándar del error, se denota como S_e y es igual a

$$\sqrt{S_e^2} = S_e = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{\sum_i \sum_j e_{ij}^2}{n-2}}$$

[21]

donde SCE es la suma de cuadrados residual (o del error) definida en

[26]

Volviendo al ejemplo del espacio disponible-venta de alimentos para mascotas, tenemos según [27] que $SCE = 0,949$ y además $n = 12$. Si aplicamos la ecuación [20], el resultado es

$$s_e^2 = \frac{SCE}{n-2} = \frac{0,949}{12-2} = 0,0949$$

[22]

y

$$s_e = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{0,949}{12-2}} = \sqrt{0,0949} = 0,3081$$



Este valor representa una medida de la variación alrededor de la recta de regresión ajustada. Se mide en unidades de la variable de respuesta, es decir, para nuestro caso es de 30,81 pesos. La interpretación del desvío estándar del error es similar a la del desvío estándar de un modelo univariado. Ésta mide la variabilidad alrededor de la media aritmética, mientras que el desvío estándar del error mide la variabilidad alrededor de la recta de regresión ajustada.

La capacidad explicativa de una ecuación de Regresión Lineal: evaluación de la idoneidad del modelo²

Una ecuación de regresión puede considerarse como un intento de emplear la información proporcionada por una variable independiente, X , para *explicar* el comportamiento de una variable dependiente, Y . Aquí presentaremos una medida del grado de éxito de ese intento con los datos de la muestra. Las observaciones de la variable dependiente tienen *variabilidad*. Esencialmente, nos preguntaremos qué *proporción* de esa variabilidad puede explicarse por la dependencia lineal y estocástica de Y sobre X .

El modelo de regresión estimado puede escribirse como

$$y_i = a + bx_i + e_i$$

o bien

$$y_i = \hat{y}_i + e_i \text{ [23]}$$

donde

$$\hat{y}_i = a + bx_i$$

y_i	$\hat{y}_i = a + bx_i$ $= 1,45 + 0,074x$	$e_i = y_i - \hat{y}_i$	$y_i - \bar{y}$ $\bar{y} = 2,375$	$\hat{y}_i - \bar{y}$
1,6	1,82	-0,22	-0,775	-0,555
2,2	1,82	0,38	-0,175	-0,555
1,4	1,82	-0,42	-0,975	-0,555
1,9	2,19	-0,29	-0,475	-0,185
2,4	2,19	0,21	0,025	-0,185
2,6	2,19	0,41	0,225	-0,185
2,3	2,56	-0,26	-0,075	0,185
2,7	2,56	0,14	0,325	0,185
2,8	2,56	0,24	0,425	0,185
2,6	2,93	-0,33	0,225	0,555
2,9	2,93	-0,03	0,525	0,555

² Aquí trabajamos nuevamente para el caso de un único valor de y por cada valor de x



y_i	$\hat{y}_i = a + bx_i$ $= 1,45 + 0,074x$	$e_i = y_i - \hat{y}_i$	$y_i - \bar{y}$ $\bar{y} = 2,375$	$\hat{y}_i - \bar{y}$
3,1	2,93	0,17	0,725	0,555

Figura 9. Valores real y predicho por la recta de regresión del espacio disponible-venta de alimentos

La cantidad \hat{y}_i es el valor promedio predicho por la recta de regresión para la variable dependiente, y el residuo e_i es la diferencia entre los valores observado y predicho. Por tanto, el residuo representa la parte del comportamiento de la variable dependiente que no puede ser explicada por su relación lineal con la variable independiente. En las tres primeras columnas de la Figura 9 aparecen los valores de los tres términos de la ecuación [23] para los datos relativos al espacio disponible-venta de alimentos.

Para nuestros propósitos, es útil modificar ligeramente la ecuación [23]. Podemos pensar en la variabilidad muestral de la variable dependiente en términos de las desviaciones respecto a la media muestral. Restando \bar{y} a cada lado de la ecuación [23], podemos escribir

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

[24]

o

$$\begin{aligned} &\text{Desvío observado respecto de la media muestral} \\ &= \text{desvío predicho respecto de la media muestral} + \text{residuo} \end{aligned}$$

En las dos últimas columnas de la Figura 9, aparecen los valores de los dos primeros términos de la ecuación

[24] para nuestros datos.

Elevando al cuadrado ambos miembros de la ecuación

[24], sumando respecto al índice muestral i y operando algebraicamente, el resultado es

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

[25]

La ecuación

[25] posee una valiosa interpretación. El término del lado izquierdo representa la variabilidad total en la muestra de la variable dependiente en torno a su media. Esta variabilidad puede descomponerse en dos partes. El primer término del lado derecho de



[25] representa la variabilidad explicada por la regresión, mientras que el segundo término representa la variabilidad no explicada, atribuida al error aleatorio. La ecuación resulta

$$\begin{aligned} & \text{Variabilidad total en la muestra} \\ &= \text{variabilidad explicada} + \text{variabilidad no explicada} \end{aligned}$$

Hasta cierto punto, a mayor proporción de variabilidad explicada mayor capacidad explicativa tiene la regresión.

A partir de lo explicado anteriormente, surgen las siguientes definiciones:

Descomposición de la suma de cuadrados y el coeficiente de determinación

Supongamos que se ajusta una ecuación de regresión lineal por mínimos cuadrados a n pares de observaciones, obteniendo

$$y_i = a + bx_i + e_i = \hat{y}_i + e_i \quad (i = 1, 2, \dots, n)$$

donde a y b son las estimaciones de mínimos cuadrados de la constante y la pendiente de la regresión poblacional y e_i son los residuos de la recta de regresión ajustada.

Definamos las siguientes expresiones (donde \bar{y} es la media muestral de la variable dependiente):

$$\text{Suma de cuadrados total: } SCT = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{Suma de cuadrados de la regresión: } SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\text{Suma de cuadrados residual (o del error): } SCE = \sum_{i=1}^n e_i^2$$

[26]

De acuerdo a [25] resulta

$$SCT = SCR + SCE$$



A partir de estos conceptos, se define el coeficiente de determinación, R^2

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Ésta es la proporción de variabilidad de la variable dependiente explicada por su relación lineal con la variable independiente.

De esta definición se deduce que la proporción de variabilidad explicada verifica

$$0 \leq R^2 \leq 1$$

y que a mayor R^2 , mayor capacidad explicativa de la regresión, siempre que los grados de libertad no sean demasiado escasos.

Para nuestro ejemplo, las sumas de cuadrados de las columnas 3 y 4 de la Figura 9 son, respectivamente

$$SCE = \sum_{i=1}^n e_i^2 = 0,949$$

[27]

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2 = 3,0025$$

[28]

El coeficiente de determinación es, por tanto,

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{0,949}{3,0025} = 0,68393006$$

que es parte de la salida de R mostrada en la Figura 7. Este resultado indica que alrededor del 68% de la variabilidad muestral de las ventas de alimentos para mascotas está explicada por la regresión. Por consiguiente, podemos concluir que usando el espacio en estantes como variable independiente, tendremos bastante éxito al explicar la variabilidad de las ventas de alimentos para mascotas. En el ámbito de las ciencias sociales, en general, los R^2 mayores a 0,5 son “buenos”, es decir, justifican un análisis de la regresión. Esta consideración varía dependiendo del ámbito de aplicación del problema.

Supuestos para el Modelo de Regresión Lineal

El método de mínimos cuadrados presentado anteriormente es un procedimiento para estimar la recta de regresión poblacional, aunque no siempre es el más apropiado. No obstante, si se formu-



lan ciertos supuestos acerca del término de error, se demuestra que los estimadores de mínimos cuadrados poseen propiedades deseables.

El primer supuesto fue esbozado al principio de este documento:

Hipótesis de **no aleatoriedad de la variable explicativa**:

Cada $x_i \forall i$ es un valor concreto definido por el investigador

[29]

En el caso de que

[29] no se cumpla, el error debe ser independiente de la variable explicativa:

Hipótesis de **independencia del error con respecto**

a la variable explicativa: $cor(\varepsilon_i, x_i) = 0$

[30]

De aquí en más asumiremos que, cuando exigimos

[29], queda implícito que ante su incumplimiento exigiremos

[30].

El término de error ε se supone que depende de innumerables factores, cada uno de ellos con una influencia sobre la variable de respuesta muy pequeña, tales que están nada o poco relacionados entre sí. Representa pues la parte impredecible de la variable de respuesta. En estas condiciones se aplica el Teorema Central del Límite, por el que el término de error sigue aproximadamente una distribución normal centrada:

Hipótesis de **normalidad**: $\varepsilon_i \sim N$

[31]

Hipótesis de **centrado de los errores**: $E(\varepsilon_i) = 0$

[32]

Esta hipótesis requiere que el error alrededor de la recta de regresión siga una distribución normal en cada valor de X . Hagamos la aclaración de que

[31] no es estrictamente necesaria hasta la realización de inferencias, que detallaremos más adelante.

Además suponemos que los términos de error están generados por distribuciones de probabilidad con la misma varianza para todos los subgrupos:



Hipótesis de **homocedasticidad**: $\text{var}(\varepsilon_i) = \sigma^2$

[33]

Esta hipótesis requiere que la variación alrededor de la recta de regresión sea constante para todos los valores de x . Esto quiere decir que los errores tienen la misma variabilidad cuando el valor de x es bajo o es alto.

Por otro lado el valor del error en un individuo es independiente del valor de error en otro individuo (observaciones independientes) que, bajo normalidad, equivale a falta de correlación:

Hipótesis de **independencia de los errores**: $\text{cor}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

[34]

Esta hipótesis requiere que los errores sean independientes para cada valor de x . Esta suposición es en especial importante cuando los datos no se recopilan simultáneamente. En esas situaciones, los errores para un período dado con frecuencia se correlacionan con los del período anterior (efecto denominado *autocorrelación*).

Podemos resumir

[31],

[32],

[33] y

[34] en la siguiente expresión:

$$\varepsilon_i \overset{IND}{\sim} N(0, \sigma^2)$$

[35]

En caso de tener varianzas distintas para cada individuo y/o correlaciones no nulas entre distintas observaciones, también se podría efectuar el ajuste si es que poseemos alguna estimación de estas varianzas y correlaciones, usando el método de mínimos cuadrados ponderados mencionado anteriormente. Este último caso queda fuera del alcance de este documento.

El Teorema de Gauss-Markov

Presentaremos aquí una justificación del uso del método de mínimos cuadrados en la estimación de la recta de regresión poblacional. Supongamos que disponemos de n pares de observaciones $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Pueden construirse varios estimadores de los parámetros α y β . Una po-



sibilidad es restringir nuestra atención a aquellos que sean combinaciones lineales de los y_i , es decir, estimadores de la forma general

$$c_1Y_1 + c_2Y_2 + \dots + c_nY_n$$

donde los c_i son números que no dependen de y_i . Dadas ciertas condiciones, para este tipo de estimadores, los estimadores de mínimos cuadrados son óptimos, como lo señala el siguiente teorema.

El teorema de Gauss-Markov

Denotemos la recta de regresión poblacional por

$$Y_i = \alpha + \beta x_i$$

y asumamos que se dispone de n pares de observaciones. Supongamos, además que se verifican los supuestos mencionados en

[33],

[34],

[35] y

[29].

Entonces, de todos los posibles estimadores insesgados de α y β que son combinación lineal de los Y_i , **los estimadores de mínimos cuadrados** (es decir, las variables aleatorias correspondientes a las estimaciones de mínimos cuadrados a y b de [13] y **Error! Reference source not found.** respectivamente) **tienen la menor varianza**.

Además, si d_0 y d_1 son dos números fijos, y queremos estimar

$$d_0\alpha + d_1\beta$$

entonces, el estimador

$$d_0a + d_1b$$

tiene **la menor varianza en la clase de los estimadores insesgados que son combinación lineal de las Y_i** (este resultado es útil cuando se usa la recta de regresión para obtener predicciones acerca de la variable dependiente).

En virtud de este teorema se dice que, **dados los supuestos mencionados, los estimadores de mínimos cuadrados son los mejores estimadores lineales insesgados**.



Comprobación de los supuestos

La comprobación de los tres supuestos para ε_i mencionados en

[33],

[34] y

[35] puede hacerse evaluando e_i . A los fines de este documento, esta evaluación será gráfica o con un estadístico particular (llamado de *Durbin-Watson*) para

[34], cuando los datos cumplen una determinada condición. Existen otras técnicas para verificar cada supuesto, pero aquí haremos sólo un análisis elemental.

Con respecto a

[30], hipótesis de independencia del error y la variable explicativa, se analiza en ciertas ocasiones cuando los x_i son aleatorios y no pueden medirse con precisión. Dejaremos estos casos excepcionales fuera del alcance de nuestro estudio y asumiremos de ahora en más que se cumple

[29].

Hipótesis de normalidad

Para evaluar la distribución de los errores construiremos un *Diagrama de Probabilidad Normal* para los errores de las observaciones muestrales. Si estos errores se distribuyen alrededor de una recta sin mostrar algún patrón de comportamiento, concluiremos que siguen una distribución aproximadamente normal. A continuación haremos la evaluación para el problema del espacio disponible-venta de alimentos para mascotas, repasando el procedimiento explicado en la sección 4.8 de la referencia [i].

Repaso: construcción de un Diagrama de Probabilidad Normal

1. Colocar los valores del conjunto de datos en un arreglo ordenado, de menor a mayor:

$e_i = y_i - \hat{y}_i$ (Desordenado)		$e_i = y_i - \hat{y}_i$ (Ordenado)
-0,22		-0,42
0,38		-0,33
-0,42		-0,29
-0,29		-0,26
0,21	→	-0,22
0,41		-0,03
-0,26		0,14
0,14		0,17
0,24		0,21
-0,33		0,24
-0,03		0,38



0,17

0,41

2. Obtener los valores de los cuantiles estándar correspondientes, teniendo en cuenta que el i -ésimo cuantil normal estándar O_i es el valor de Z de una distribución normal estándar debajo del cual se encuentra la proporción $\frac{i}{n+1}$ del área bajo la curva:

$e_i = y_i - \hat{y}_i$ (Ordenado)	i	$\frac{i}{n+1} = \frac{i}{13}$ (Área bajo la curva normal)	O_i (Valor de Z correspondiente)
-0,42	1	0,076923077	-1,42607719
-0,33	2	0,153846154	-1,02007629
-0,29	3	0,230769231	-0,73631568
-0,26	4	0,307692308	-0,5024022
-0,22	5	0,384615385	-0,2933814
-0,03	6	0,461538462	-0,09655846
0,14	7	0,538461538	0,09655846
0,17	8	0,615384615	0,2933814
0,21	9	0,692307692	0,5024022
0,24	10	0,769230769	0,73631568
0,38	11	0,846153846	1,02007629
0,41	12	0,923076923	1,42607719

3. Graficar los pares de puntos correspondientes utilizando los valores de los datos observados en el eje vertical y los valores de los cuantiles normales estándar en el eje horizontal:

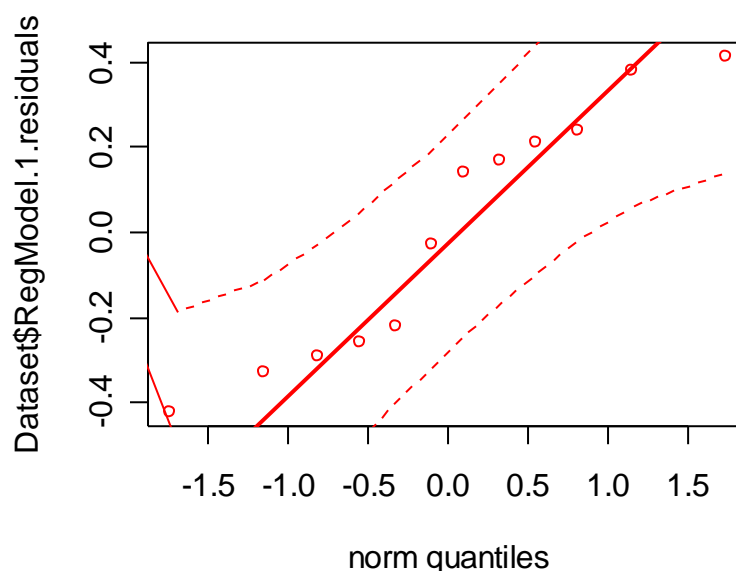


Figura 10. Diagrama de Probabilidad Normal para los errores del espacio disponible-venta de alimentos.

4. Evaluar la probabilidad de que la variable aleatoria de interés tenga una distribución normal (o por lo menos se aproxime) con un examen del diagrama en busca



de evidencias de linealidad (o de una línea recta)

En la Figura 10 se dibujan los puntos del Diagrama de Probabilidad Normal. Podemos decir que la normalidad se cumple cuando la distribución de los puntos es aleatoria alrededor de una línea recta ascendente imaginaria, o bien cuando no presenta ningún patrón no lineal (Un ejemplo de patrón no lineal sería una curva ascendente o descendente).

Para nuestro caso, podemos concluir que la distribución de los puntos es aproximadamente aleatoria, por lo que no tenemos una justificación suficiente para negar la existencia de una distribución normal para los errores del problema.

Hipótesis de homocedasticidad

Para evaluar la homocedasticidad debemos comprobar que la varianza del error sea constante en cada subgrupo, es decir, que no existan diferencias importantes en la variabilidad de los errores para distintos valores de x_i . Como a y b son constantes, esto equivale a decir que la variabilidad de los errores no debe presentar diferencias importantes con $\hat{y}_i = a + bx_i$. Esto se logra observando el *Gráfico de los Residuales* de la **Error! Reference source not found.**, que muestra los errores o residuos en el eje Y y los valores de las x_i en el eje X.

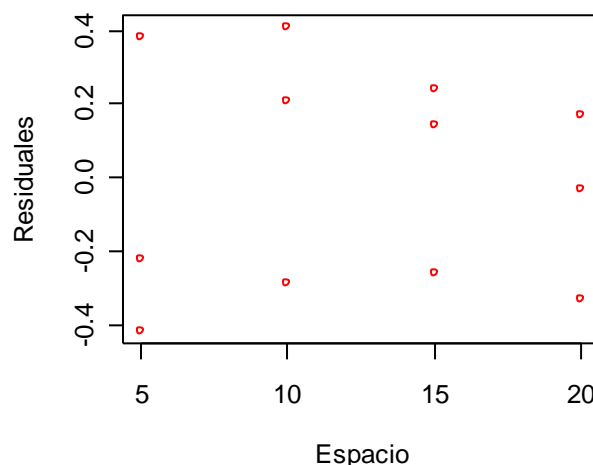


Figura 11. Gráfico de los Residuales
para los errores del espacio disponible-venta de alimentos

Hipótesis de centrado de los errores

Para evaluar el centrado de los errores, generaremos un gráfico de los residuales con respecto a cada valor de la variable independiente, lo que nos permitirá visualizar su dispersión alrededor del cero. Vemos esto en la Figura 12.

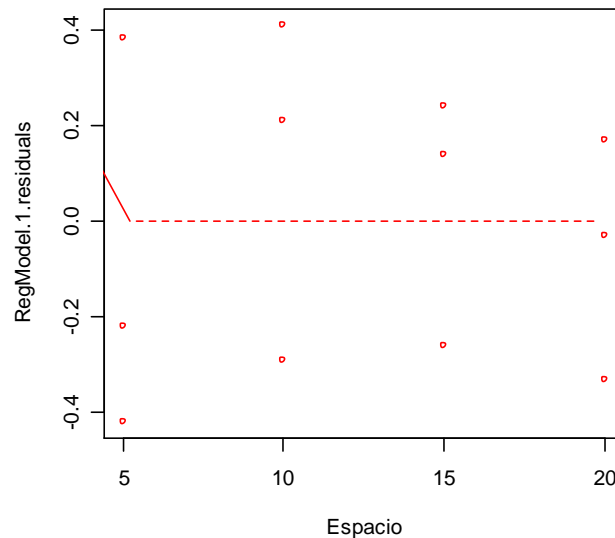


Figura 12. Gráfico de los Residuales vs. Espacio para los errores del espacio disponible-venta de alimentos. Incluye línea punteada donde los residuales son cero para visualizar mejor la dispersión.

Hipótesis de independencia de los errores

La hipótesis de independencia de los errores puede evaluarse con un *Gráfico de los Residuales vs. Tiempo* que muestre los puntos en el orden o secuencia en que fueron obtenidos, para examinar si hay alguna relación aparente entre un dato cualquiera y el dato recolectado inmediatamente antes o después. Suponemos, para el problema del espacio disponible-venta de alimentos, que los datos se tomaron simultáneamente, con lo que se cumple este supuesto.

En otras situaciones, debemos prestar especial atención a los datos recopilados en períodos sucesivos en puntos adyacentes a lo largo del tiempo (años, meses o semanas por ejemplo). En estas circunstancias, será más probable que residuales positivos sigan a residuales positivos y que residuales negativos sigan a residuales negativos. Esto será evidente en el *Gráfico de los Residuales vs. Tiempo*. Un patrón de este tipo en los residuales se conoce como *autocorrelación positiva*. Cuando está presente una autocorrelación importante, se viola esta suposición y en consecuencia se invalida el modelo de regresión ajustado. Para medir adecuadamente este efecto, como anticipábamos al inicio de esta sección, utilizaremos el *Estadístico de Durbin-Watson*, explicado a continuación.

Estadístico de Durbin-Watson

El Estadístico de Durbin-Watson, D , se define como

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

[36]



donde e_i = residual del período i

Como ejemplo consideremos el siguiente caso. El gerente de una tienda desea predecir las ventas semanales con los datos del número de clientes que hacen compras en un período de 15 semanas. Como los datos se recopilaron durante 15 semanas consecutivas en la misma tienda, deberá estudiarse el efecto de autocorrelación de los residuales. Los datos de esta tienda se resumen en la **Error! Reference source not found.** La **Error! Reference source not found.** representa un *Gráfico de los Residuales vs. Tiempo* para el ejemplo.

Dataset. 2				
	Semana	Clientes	Ventas	Residuos
1	1	794	9.33	0.93857269
2	2	799	8.26	-0.28522845
3	3	837	7.48	-2.23411711
4	4	855	9.08	-1.18780121
5	5	845	9.83	-0.13019893
6	6	844	10.09	0.16056130
7	7	863	11.01	0.49611697
8	8	875	11.49	0.60699423
9	9	880	12.07	1.03319310
10	10	905	12.55	0.74418740
11	11	886	11.92	0.69863173
12	12	843	10.27	0.37132153
13	13	904	11.80	0.02494763
14	14	950	12.15	-1.04002285
15	15	841	9.64	-0.19715802

Figura 13. Clientes y ventas para 15 semanas consecutivas. Se incluyen los residuos

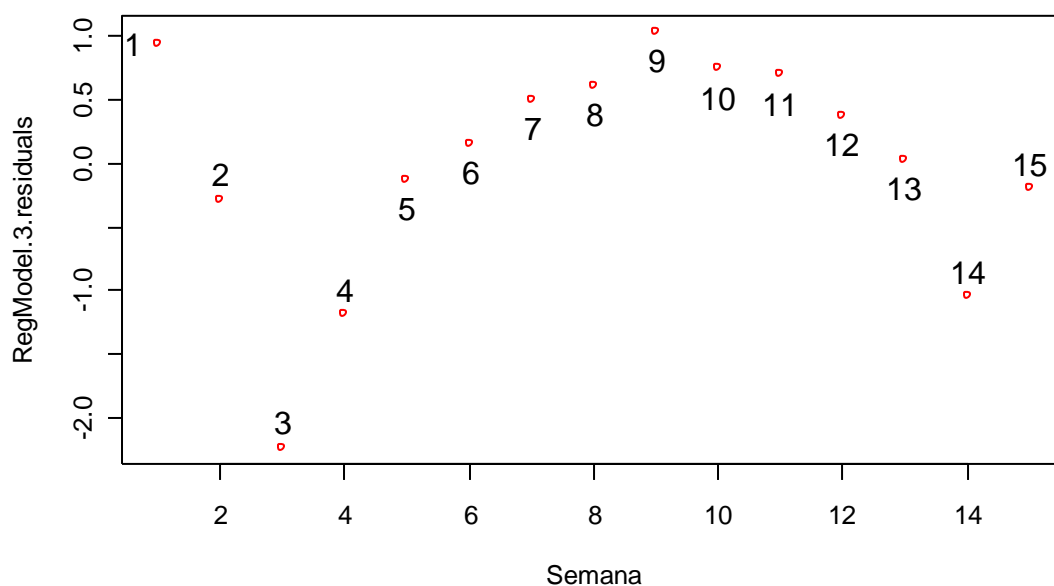




Figura 14. Gráfico de los Residuales vs. Tiempo
para los errores de la cantidad de clientes-ventas semanales

Para entender mejor qué mide el estadístico de Durbin-Watson, es necesario examinar la composición del estadístico D presentado en la ecuación

[36]. El numerador representa los cuadrados de la diferencia entre dos residuales sucesivos, sumados desde la segunda observación hasta la n -ésima. El denominador constituye la suma de los cuadrados de los residuales. Cuando los residuales sucesivos tienen una autocorrelación positiva, el valor de D se acerca a 0. Si los residuales no se correlacionan, el valor de D será cercano a 2. $D > 2$ es indicio de autocorrelación negativa.

Para el caso citado, usamos la ecuación

[36] con los datos de la Figura 13 y se obtiene

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{10,058}{11,39} = 0,883$$

La salida correspondiente de R es

DW = 0.883, p-value = 0.003491
alternative hypothesis: true autocorrelation is greater than 0

Un punto importante al usar el estadístico de Durbin-Watson es determinar cuándo la autocorrelación positiva es lo suficientemente grande para invalidar el modelo. La respuesta a esta pregunta se obtiene llevando a cabo un test de hipótesis para $\rho = \text{Cor}(\varepsilon_i; \varepsilon_{i-1})$, que dependerá de n , el número de observaciones que se analizan y P , el número de variables independientes en el modelo (en la regresión lineal simple, $P = 1$). Los valores críticos para este test se encuentran registrados en tablas de Durbin-Watson. Adjuntamos una en la

Tabla **4**, que se encuentra en el Anexo I: Tablas útiles. Para esta tabla, d_L es el valor crítico inferior. Si D es menor que d_L se concluye que existe evidencia de una autocorrelación positiva entre los residuales. En ese caso, los métodos de mínimos cuadrados no son adecuados y se requieren métodos alternativos. Por su parte, d_U es el valor crítico superior de D , arriba del cual se concluye que no hay evidencia de correlación entre los residuales. Si D estuviera entre d_L y d_U , no es posible llegar a una conclusión.



Para nuestro caso, $P=1$, $n=15$, $dL = 1,08$ y $dU = 1,36$. Dado que $D = 0,883 < 1,08$ se concluye que existe autocorrelación entre los residuales. Por tanto, el análisis de regresión con el método de mínimos cuadrados para los datos de la figura 14 no es adecuado debido a la presencia de una autocorrelación importante entre los residuales. Deberán considerarse enfoques alternativos no incluidos en este documento.

Inferencias acerca de la pendiente

Recordando a partir de

[5] la importancia escasa de α , la ordenada al origen de la recta de regresión, nos concentraremos en β , su pendiente. En primer lugar, estimaremos su varianza, σ_b^2 . Tengamos presente, como indicábamos en la página 21, que para que los cálculos de inferencia sean válidos, debe comprobarse, además de los supuestos mencionados en

[33],

[34],

[35] y

[29], el supuesto de normalidad señalado en

[31].

Estimación de la varianza del estimador de la pendiente de la recta de regresión poblacional

Denotemos por b la estimación de mínimos cuadrados de la pendiente de la recta de regresión poblacional. Si se verifican los supuestos mencionados en

[31],

[33],

[34],

[35] y

[29], el estimador correspondiente a β es insesgado y se demuestra que tiene varianza

$$\sigma_b^2 = \frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_e^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Un estimador insesgado de σ_b^2 se obtiene mediante

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_e^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$



En el ejemplo sobre espacio disponible-venta de alimentos para mascotas, tenemos de **[22]** que $s_e^2 = 0,0949$, de **[15]** que $\sum_{i=1}^n x_i^2 = 2250$, y de **[16]** que $\bar{x} = 12,5$. Entonces

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{0,0949}{2250 - 12 \times 12,5^2} = 0,00025306$$

luego, el error estándar estimado del estimador de mínimos cuadrados de la pendiente de la recta de regresión poblacional es

$$s_b = \sqrt{s_b^2} = \sqrt{0,00025306} = 0,01591$$

[37]

Un caso especial de interés práctico para β es estudiar si su valor es 0. En tal caso el modelo de regresión poblacional

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

se convierte en

$$Y_i = \alpha + \varepsilon_i.$$

con lo cual cualquiera sea el valor que tome la variable explicativa x_i , la variable de respuesta será una variable aleatoria con media α y varianza igual a la varianza de los errores, es decir σ_e^2 . Por lo tanto, el valor esperado de la variable de respuesta no se verá afectado linealmente por el valor de la variable explicativa. En otras palabras, la variable de respuesta no puede explicarse por una relación lineal con la variable explicativa.

Para verificarlo, debemos calcular b , estimador de β , la pendiente de la recta de regresión poblacional, y constatar que difiere significativamente de 0. Esto puede hacerse de la siguiente manera

Pruebas para determinar la existencia de una relación significativa entre la variable explicativa y la variable de respuesta

1. Prueba t

Las hipótesis nula y alternativa se establecen como sigue

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

H_0 establece que no hay una relación lineal entre la variable explicativa y la variable de respuesta. H_1 establece que hay una relación lineal.



El estadístico t es igual a la diferencia entre la pendiente de la muestra y la pendiente hipotética dividida por el error estándar de la pendiente

$$t = \frac{b - \beta}{s_b}$$

donde, por hipótesis, $\beta = 0$.

El estadístico de prueba sigue una distribución t con $n - 2$ grados de libertad, para un nivel de significancia α . La regla de decisión es rechazar H_0 si

$$\frac{b - \beta}{s_b} = \frac{b}{s_b} > t_{n-2, \alpha/2}$$

o si

$$\frac{b - \beta}{s_b} = \frac{b}{s_b} < -t_{n-2, \alpha/2}$$

para el ejemplo, tenemos de **[17]** que $b = 0,074$ y de **[37]** que $s_b = 0,01591$. Por lo tanto, para un nivel de significancia de 0,05 se obtiene

$$\frac{b}{s_b} = 4,652$$

y

$$t_{n-2, \alpha/2} = t_{10, 0,025} = 2,2281$$

[38]

por lo tanto se rechaza H_0 y se concluye con un 95% de confianza que β es distinto de 0, o lo que es lo mismo, que existe una relación significativa entre la variable explicativa y la variable de respuesta.

2. Prueba F

Un enfoque alternativo para probar si la pendiente en una regresión lineal simple es estadísticamente significativa es usar una prueba F . Esta prueba se usa para probar la razón de dos varianzas. Al probar la significancia de la pendiente, la prueba F es la razón de la varianza que se debe a la regresión dividida por la varianza del error

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n - p - 1}}$$

El numerador se llama *cuadrado medio de la regresión* y representa la varianza debida a la regresión. El denominador es el *cuadrado medio de los residuos* y representa la varianza del error. Por su parte, p es el número de variables explicativas en el modelo de regresión, para nuestro caso $p = 1$. La prueba F establece la hipótesis nula de que ambas varianzas, en la población, son iguales. El estadístico de prueba F sigue una distribución F con p y $n - p - 1$ grados de libertad. La regla de decisión es rechazar H_0 si $F > F_U$, el valor crítico superior de F .

Para el ejemplo, tenemos de **[27]** que $SCE = 0,949$ y a partir de **[27]** y **[28]** deducimos $SCR = 3,0025 - 0,949 = 2,0535$. Por lo tanto, para un nivel de significancia de 0,05 se obtiene



$$F = \frac{\frac{2,0535}{1}}{\frac{0,949}{12-1-1}} = 21,6386$$

y

$$F_{U(0,05, 1, 10)} = 4,96$$

por lo tanto se rechaza H_0 y se concluye con un 95% de confianza que existe una relación significativa entre la variable explicativa y la variable de respuesta.

3. Prueba de Intervalo de Confianza para β

Otra alternativa para probar la existencia de una relación lineal entre las variables es establecer una estimación del intervalo de confianza de β y determinar si el valor hipotético $\beta = 0$ está incluido en ese intervalo.

La estimación del intervalo de confianza para la pendiente se obtiene tomando la pendiente de la muestra b y sumando y restando el valor crítico del estadístico t multiplicado por el error estándar de la pendiente

$$b \pm t_{n-2, \alpha/2} \times s_b$$

Para nuestro ejemplo tenemos **[17]** que $b = 0,074$ y de **[37]** que $s_b = 0,01591$. Además $t_{n-2, \alpha/2} = t_{10, 0,025} = 2,2281$. Por lo tanto

$$0,074 \pm 2,2281 \times 0,01591 = 0,074 \pm 0,03545 = [0,03855; 0,10945]$$

Este intervalo no cubre al 0, por lo tanto se concluye con un 95% de confianza que existe una relación significativa entre la variable explicativa y la variable de respuesta.

Si se verifica la hipótesis de $\beta \neq 0$, según adelantábamos en la página **Error! Bookmark not defined.**, existirá una relación lineal entre la variable explicativa y la variable de respuesta, que será positiva para una estimación de $\beta > 0$ y negativa para una estimación de $\beta < 0$.

Pronósticos

Una aplicación importante de la regresión es realizar el pronóstico de la variable de respuesta para un valor determinado de la variable explicativa. Supongamos que la variable explicativa es igual a cierto valor específico x_k , y que la relación entre las variables dependiente e independiente es lineal. El correspondiente valor de la variable de respuesta será

$$Y_k = \alpha + \beta x_k + \varepsilon_k \quad [39]$$

cuya esperanza es

$$E(Y/x_k) = \alpha + \beta x_k \quad [40]$$

Estamos interesados en dos tipos de pronóstico distintos:

1. Estimar el valor de Y_k en la ecuación **[39]**.
2. Estimar la esperanza condicional $E(Y/x_k)$ de la ecuación **[40]**, es decir, el valor promedio de la variable de respuesta cuando se fija en x_k la variable explicativa.



Si los supuestos mencionados en

[31],

[33],

[34],

[35] y

[29] se verifican, el estimador puntual es el mismo para los dos casos. Es lógico sustituir los α y β desconocidos por sus estimaciones de mínimos cuadrados, a y b . Por lo tanto, $\alpha + \beta x_k$ se estima mediante $a + bx_k$. Por el teorema de Gauss-Markov, sabemos que el estimador correspondiente es el mejor entre los lineales e insesgados. En consecuencia, para los dos casos, un estimador puntual adecuado bajo nuestras hipótesis es

$$\hat{Y}_k = a + bx_k$$

Esto se deduce del hecho de que, en el presente contexto, el error aleatorio ε_k de la ecuación [39] tiene media 0.

Una estimación completa es aquella que incluye el error de estimación. Para este fin, se recomiendan las estimaciones por intervalos que se construyen con la estimación puntual más/menos el error de estimación multiplicado por una constante. Los dos casos planteados tienen distintas soluciones. El motivo es que existe incertidumbre sobre el valor que tomará la variable aleatoria ε_k que aparece en la ecuación [39] pero no en la ecuación [40]. Los procedimientos apropiados se resumen a continuación.

Intervalo de confianza para la esperanza

Supongamos que se verifica el modelo de regresión poblacional

$$Y = \alpha + \beta X + \varepsilon$$

y los supuestos mencionados en

[31],

[33],

[34],

[35] y

[29]. Sean a y b las estimaciones de mínimos cuadrados de α y β , basadas en $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Entonces, puede demostrarse que el intervalo de confianza del $100(1 - \alpha)\%$ para la estimación de la esperanza condicional $E(Y_k/x_k)$ es

$$\hat{Y}_k \pm t_{n-2, \alpha/2} \times \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}} \times S_e^2 \quad [41]$$

Intervalo de predicción para valores individuales

El intervalo de predicción de Y_k con probabilidad $(1 - \alpha)$ es



$$\hat{Y}_k \pm t_{n-2, \alpha/2} \times \sqrt{\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right] \times s_e^2} \quad [42]$$

donde

$$\hat{Y}_k = a + bx_k$$

Para nuestro ejemplo, supongamos que queremos predecir exactamente cuántas ventas habrá para un espacio en estantes de 13 decímetros. Entonces $x_k = 13$ y tenemos de **[15]** que $\sum_{i=1}^n x_i^2 = 2250$, de **[16]** que $\bar{x} = 12,5$, de **[22]** que $s_e^2 = 0,0949$ y para un nivel de significancia de $\alpha = 0,05$, tenemos de **[38]** que $t_{10, 0,025} = 2,2281$. Calculamos además $\hat{Y}_k = a + bx_k = 1,45 + 0,074 \times 13 = 2,412$. Por lo tanto, con estos datos se obtiene

$$\begin{aligned} \hat{Y}_k \pm t_{n-2, \alpha/2} \times \sqrt{\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right] \times s_e^2} &= 2,412 \pm 2,2281 \times \sqrt{\left[1 + \frac{1}{12} + \frac{(13 - 12,5)^2}{2250 - 12 \times 12,5^2} \right] \times 0,0949} \\ &= 2,412 \pm 0,7146 \\ &= [1,6974 ; 3,1266] \end{aligned}$$

La salida de R es

```
> predict(RegModel.1, data.frame(Espacio=13), level=.95,
         interval="prediction")
         fit          lwr          upr
2.412      1.697356    3.126644
```

Si queremos construir un intervalo de confianza para la esperanza de las ventas cuando el espacio en estantes sea de 13 decímetros, con los mismos datos anteriores obtendremos

$$\begin{aligned} \hat{Y}_k \pm t_{n-2, \alpha/2} \times \sqrt{\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right] \times s_e^2} &= 2,412 \pm 2,2281 \times \sqrt{\left[\frac{1}{12} + \frac{(13 - 12,5)^2}{2250 - 12 \times 12,5^2} \right] \times 0,0949} \\ &= 2,412 \pm 0,1989 \\ &= [2,2130 ; 2,6109] \end{aligned}$$

La salida de R es



```
> predict(RegModel.1, data.frame(Espacio=13), level=.95, inter-  
val="confidence")
```

<i>fit</i>	<i>lwr</i>	<i>upr</i>
2.412	2.213063	2.610937

En la **Error! Reference source not found.** ilustramos estos intervalos. La amplitud del intervalo de confianza para la media representa la confianza que nos merece la estimación realizada de la recta de regresión poblacional con el tamaño de muestra elegido. Por otra parte, la amplitud del intervalo de predicción para un valor puntual x_k es una medida de nuestra incertidumbre sobre el valor que tomarán las ventas para un espacio en estantes determinado, y como indicamos en $Y_k = \alpha + \beta x_k + \varepsilon_k$ [39] incluye dos fuentes de variación: el error de estimación de la recta y la variabilidad del error aleatorio. Es por esta última razón que la amplitud del intervalo de predicción es mayor que la del intervalo de confianza.

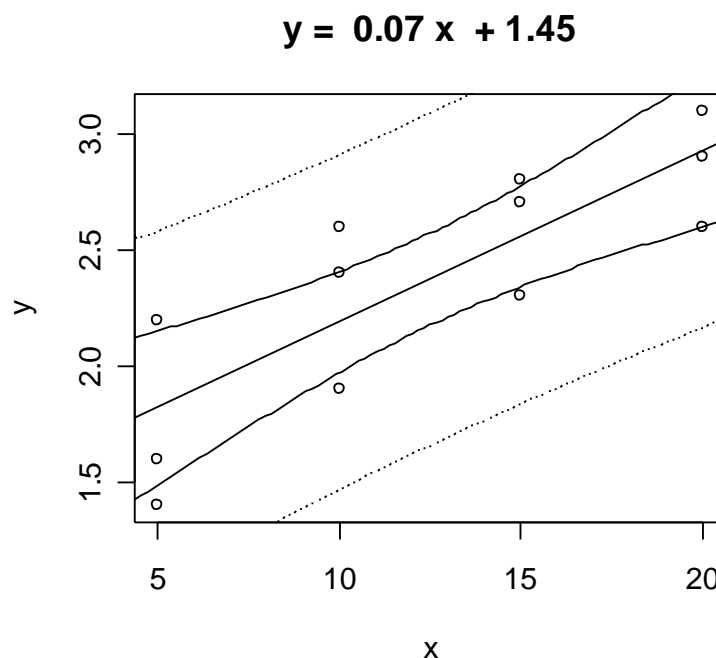


Figura 15. Datos del espacio disponible-venta de alimentos y recta de regresión estimada por mínimos cuadrados. Se incluyen los intervalos de confianza del 95% para la media (línea continua) y de predicción para un valor puntual (línea punteada)

Pasos del análisis de regresión

A continuación se explican, a modo de resumen, los pasos necesarios para hacer un análisis de regresión. Hacemos hincapié en el orden de ejecución de los pasos, que permitirá justificar adecuadamente el procedimiento.



Supongamos que disponemos de dos variables, X e Y , y supongamos que deseamos hacer un análisis tomando a X como variable explicativa y a Y como variable de respuesta

Paso 1: ¿Existirá una relación entre las variables explicativa y de respuesta?

Comenzamos con un diagrama de dispersión para observar la relación posible entre X e Y . Construir la ecuación para la recta estimada de regresión $y = a + bx$ calculando a y b , y continuar con el paso siguiente para obtener conclusiones más precisas acerca del modelo.

Paso 2: ¿Puedo estimar una recta de regresión poblacional en forma óptima con el método de mínimos cuadrados?

Para responder esta pregunta, debo justificar el uso del método comprobando los supuestos de centrado, homocedasticidad e independencia de los errores. No olvidar aplicar el cálculo del estadístico de Durbin-Watson para observaciones no simultáneas (observaciones tomadas en puntos adyacentes a lo largo del tiempo). Si los supuestos se comprueban, seguir con el Paso 3. Si no, no se aplica el Teorema de Gauss-Markov, y en consecuencia el método de mínimos cuadrados no proporcionará estimadores óptimos, por lo que los métodos a utilizar serán distintos a los incluidos en este documento.

Paso 3: ¿Existirá una real influencia lineal de la variable explicativa sobre la variabilidad de la variable de respuesta?

Debemos comprobar esto haciendo los tests t , F o de intervalo de confianza y verificando, con un nivel α de significación, que β sea distinto de 0. Recordar que para que el test t sea válido debe comprobarse el supuesto de normalidad de los errores. Si el resultado de cualquiera de los tests rechaza la hipótesis nula, seguir con el Paso 4. Si no, se concluye que la variabilidad de la variable de respuesta no puede explicarse por una relación lineal con la variable explicativa y en consecuencia se recomienda buscar otra u otras variables explicativas.

Paso 4: ¿Qué tan bueno es el modelo para estimar?

Para saber esto, debo evaluar la capacidad explicativa del modelo, calculando R^2 , que indicará qué porcentaje de la variabilidad de la variable de respuesta queda explicada por el modelo de regresión. Si este valor es bajo, no se justifica el análisis de regresión y debemos descartar el modelo (el umbral de R^2 depende del ámbito de aplicación del problema) Continuar con el paso 5.

Paso 5: Análisis de los resultados del modelo

Si llegamos hasta aquí, es porque tenemos una justificación suficiente para la validez del modelo. Por lo tanto, podemos analizar los resultados del modelo, evaluando intervalos de confianza para la esperanza o de predicción para un valor individual, dado un valor especificado de la variable explicativa.



Bibliografía

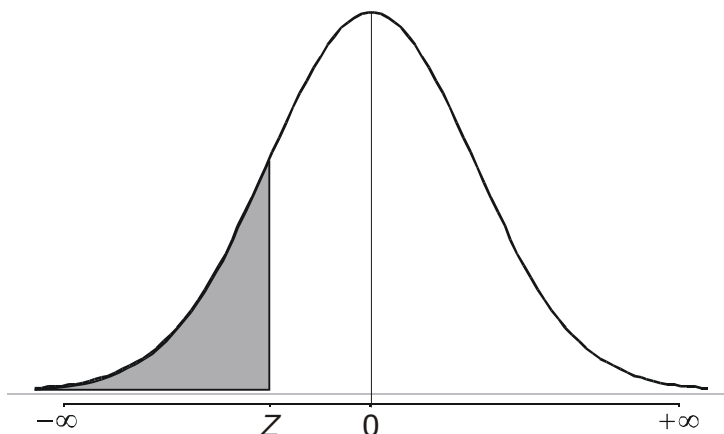
- [i] J. Hernández Orallo, M. J. Ramírez Quintana, C. Ferri Ramírez
Introducción a la Minería de Datos
Pearson Educación, Madrid, 2004
- [ii] M. Berenson, D. Levine, T. Krehbiel et al
Estadística para Administración, Segunda Edición
Pearson Educación, México, 2001
- [iii] P. Newbold
Estadística para los Negocios y la Economía
Prentice Hall, Madrid, 1997
- [iv] R for Windows
<http://www.r-project.org/>
- [v] R Commander
<http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>
- [vi] SimpleR
<http://www.math.csi.cuny.edu/Statistics/R/simpleR/>
- [vii] Cátedra de Sistemas de Gestión II – UTN-FRRO
<http://www.frro.utn.edu.ar/isi/gestion2/index.htm>



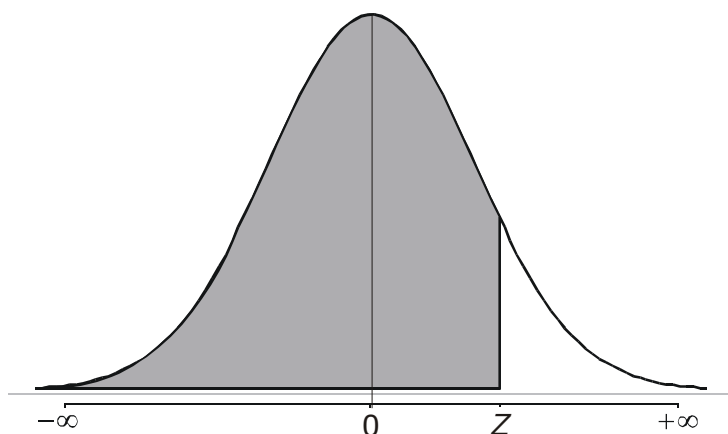
Anexo I: Tablas útiles

Tabla 1: Distribución Normal Estándar acumulada

Cada elemento representa el área bajo la distribución normal estándar acumulada, de $-\infty$ a Z



Z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00
-4	0,000022	0,000023	0,000024	0,000025	0,000026	0,000027	0,000028	0,000029	0,000030	0,000032
-3,9	0,000033	0,000034	0,000036	0,000037	0,000039	0,000041	0,000042	0,000044	0,000046	0,000048
-3,8	0,000050	0,000052	0,000054	0,000057	0,000059	0,000062	0,000064	0,000067	0,000070	0,000072
-3,7	0,000075	0,000078	0,000082	0,000085	0,000088	0,000092	0,000096	0,000100	0,000104	0,000108
-3,6	0,000112	0,000117	0,000121	0,000126	0,000131	0,000136	0,000142	0,000147	0,000153	0,000159
-3,5	0,000165	0,000172	0,000179	0,000185	0,000193	0,000200	0,000208	0,000216	0,000224	0,000233
-3,4	0,000242	0,000251	0,000260	0,000270	0,000280	0,000291	0,000302	0,000313	0,000325	0,000337
-3,3	0,000350	0,000362	0,000376	0,000390	0,000404	0,000419	0,000434	0,000450	0,000467	0,000483
-3,2	0,000501	0,000519	0,000538	0,000557	0,000577	0,000598	0,000619	0,000641	0,000664	0,000687
-3,1	0,000711	0,000736	0,000762	0,000789	0,000816	0,000845	0,000874	0,000904	0,000936	0,000968
-3	0,001001	0,001035	0,001070	0,001107	0,001144	0,001183	0,001223	0,001264	0,001306	0,001350
-2,9	0,001395	0,001441	0,001489	0,001538	0,001589	0,001641	0,001695	0,001750	0,001807	0,001866
-2,8	0,001926	0,001988	0,002052	0,002118	0,002186	0,002256	0,002327	0,002401	0,002477	0,002555
-2,7	0,002635	0,002718	0,002803	0,002890	0,002980	0,003072	0,003167	0,003264	0,003364	0,003467
-2,6	0,003573	0,003681	0,003793	0,003907	0,004025	0,004145	0,004269	0,004397	0,004527	0,004661
-2,5	0,004799	0,004940	0,005085	0,005234	0,005386	0,005543	0,005703	0,005868	0,006037	0,006210
-2,4	0,006387	0,006569	0,006756	0,006947	0,007143	0,007344	0,007549	0,007760	0,007976	0,008198
-2,3	0,008424	0,008656	0,008894	0,009137	0,009387	0,009642	0,009903	0,010170	0,010444	0,010724
-2,2	0,011011	0,011304	0,011604	0,011911	0,012224	0,012545	0,012874	0,013209	0,013553	0,013903
-2,1	0,014262	0,014629	0,015003	0,015386	0,015778	0,016177	0,016586	0,017003	0,017429	0,017864
-2	0,018309	0,018763	0,019226	0,019699	0,020182	0,020675	0,021178	0,021692	0,022216	0,022750
-1,9	0,023295	0,023852	0,024419	0,024998	0,025588	0,026190	0,026803	0,027429	0,028067	0,028716
-1,8	0,029379	0,030054	0,030742	0,031443	0,032157	0,032884	0,033625	0,034379	0,035148	0,035930
-1,7	0,036727	0,037538	0,038364	0,039204	0,040059	0,040929	0,041815	0,042716	0,043633	0,044565
-1,6	0,045514	0,046479	0,047460	0,048457	0,049471	0,050503	0,051551	0,052616	0,053699	0,054799
-1,5	0,055917	0,057053	0,058208	0,059380	0,060571	0,061780	0,063008	0,064256	0,065522	0,066807
-1,4	0,068112	0,069437	0,070781	0,072145	0,073529	0,074934	0,076359	0,077804	0,079270	0,080757
-1,3	0,082264	0,083793	0,085344	0,086915	0,088508	0,090123	0,091759	0,093418	0,095098	0,096801
-1,2	0,098525	0,100273	0,102042	0,103835	0,105650	0,107488	0,109349	0,111233	0,113140	0,115070
-1,1	0,117023	0,119000	0,121001	0,123024	0,125072	0,127143	0,129238	0,131357	0,133500	0,135666
-1	0,137857	0,140071	0,142310	0,144572	0,146859	0,149170	0,151505	0,153864	0,156248	0,158655
-0,9	0,161087	0,163543	0,166023	0,168528	0,171056	0,173609	0,176186	0,178786	0,181411	0,184060
-0,8	0,186733	0,189430	0,192150	0,194894	0,197662	0,200454	0,203269	0,206108	0,208970	0,211855
-0,7	0,214764	0,217695	0,220650	0,223627	0,226627	0,229650	0,232695	0,235762	0,238852	0,241964
-0,6	0,245097	0,248252	0,251429	0,254627	0,257846	0,261086	0,264347	0,267629	0,270931	0,274253
-0,5	0,277595	0,280957	0,284339	0,287740	0,291160	0,294598	0,298056	0,301532	0,305026	0,308538
-0,4	0,312067	0,315614	0,319178	0,322758	0,326355	0,329969	0,333598	0,337243	0,340903	0,344578
-0,3	0,348268	0,351973	0,355691	0,359424	0,363169	0,366928	0,370700	0,374484	0,378281	0,382089
-0,2	0,385908	0,389739	0,393580	0,397432	0,401294	0,405165	0,409046	0,412936	0,416834	0,420740
-0,1	0,424655	0,428576	0,432505	0,436441	0,440382	0,444330	0,448283	0,452242	0,456205	0,460172
0	0,464144	0,468119	0,472097	0,476078	0,480061	0,484047	0,488033	0,492022	0,496011	0,500000

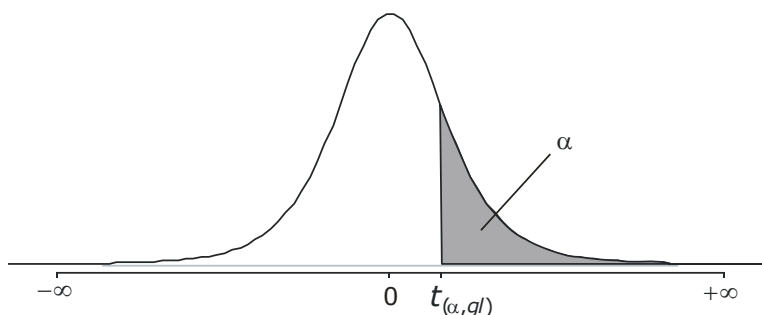


Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500000	0,503989	0,507978	0,511967	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,1	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,2	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,3	0,617911	0,621719	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,5	0,691462	0,694974	0,698468	0,701944	0,705402	0,708840	0,712260	0,715661	0,719043	0,722405
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,7	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,8	0,788145	0,791030	0,793892	0,796731	0,799546	0,802338	0,805106	0,807850	0,810570	0,813267
0,9	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,1	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,878999	0,881000	0,882977
1,2	0,884930	0,886860	0,888767	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,3	0,903199	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914656	0,916207	0,917736
1,4	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,5	0,933193	0,934478	0,935744	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,7	0,955435	0,956367	0,957284	0,958185	0,959071	0,959941	0,960796	0,961636	0,962462	0,963273
1,8	0,964070	0,964852	0,965621	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,9	0,971284	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691
2,1	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,2	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,3	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,6	0,995339	0,995473	0,995603	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,7	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,8	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,9	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605
3	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,1	0,999032	0,999064	0,999096	0,999126	0,999155	0,999184	0,999211	0,999238	0,999264	0,999289
3,2	0,999313	0,999336	0,999359	0,999381	0,999402	0,999423	0,999443	0,999462	0,999481	0,999499
3,3	0,999517	0,999533	0,999550	0,999566	0,999581	0,999596	0,999610	0,999624	0,999638	0,999650
3,4	0,999663	0,999675	0,999687	0,999698	0,999709	0,999720	0,999730	0,999740	0,999749	0,999758
3,5	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999821	0,999828	0,999835
3,6	0,999841	0,999847	0,999853	0,999858	0,999864	0,999869	0,999874	0,999879	0,999883	0,999888
3,7	0,999892	0,999896	0,999900	0,999904	0,999908	0,999912	0,999915	0,999918	0,999922	0,999925
3,8	0,999928	0,999930	0,999933	0,999936	0,999938	0,999941	0,999943	0,999946	0,999948	0,999950
3,9	0,999952	0,999954	0,999956	0,999958	0,999959	0,999961	0,999963	0,999964	0,999966	0,999967
4	0,999968	0,999970	0,999971	0,999972	0,999973	0,999974	0,999975	0,999976	0,999977	0,999978



Tabla 2: Valores críticos de t

Para un número dado de grados de libertad, el elemento representa el valor crítico de t que corresponde a un área de la cola superior especificada (α)



Grados de Libertad	Áreas de la cola superior									
	0,4	0,25	0,1	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	0,3249	1,0000	3,0777	6,3137	12,7062	31,8210	63,6559	127,3211	318,2888	636,5776
2	0,2887	0,8165	1,8856	2,9200	4,3027	6,9645	9,9250	14,0892	22,3285	31,5998
3	0,2767	0,7649	1,6377	2,3534	3,1824	4,5407	5,8408	7,4532	10,2143	12,9244
4	0,2707	0,7407	1,5332	2,1318	2,7765	3,7469	4,6041	5,5975	7,1729	8,6101
5	0,2672	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321	4,7733	5,8935	6,8685
6	0,2648	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074	4,3168	5,2075	5,9587
7	0,2632	0,7111	1,4149	1,8946	2,3646	2,9979	3,4995	4,0294	4,7853	5,4081
8	0,2619	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554	3,8325	4,5008	5,0414
9	0,2610	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498	3,6896	4,2969	4,7809
10	0,2602	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693	3,5814	4,1437	4,5868
11	0,2596	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058	3,4966	4,0248	4,4369
12	0,2590	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545	3,4284	3,9296	4,3178
13	0,2586	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123	3,3725	3,8520	4,2209
14	0,2582	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768	3,3257	3,7874	4,1403
15	0,2579	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467	3,2860	3,7329	4,0728
16	0,2576	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208	3,2520	3,6861	4,0149
17	0,2573	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982	3,2224	3,6458	3,9651
18	0,2571	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784	3,1966	3,6105	3,9217
19	0,2569	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609	3,1737	3,5793	3,8833
20	0,2567	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453	3,1534	3,5518	3,8496
21	0,2566	0,6864	1,3232	1,7207	2,0796	2,5176	2,8314	3,1352	3,5271	3,8193
22	0,2564	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188	3,1188	3,5050	3,7922
23	0,2563	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073	3,1040	3,4850	3,7676
24	0,2562	0,6848	1,3178	1,7109	2,0639	2,4922	2,7970	3,0905	3,4668	3,7454
25	0,2561	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874	3,0782	3,4502	3,7251
26	0,2560	0,6840	1,3150	1,7056	2,0555	2,4786	2,7787	3,0669	3,4350	3,7067
27	0,2559	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707	3,0565	3,4210	3,6895
28	0,2558	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633	3,0470	3,4082	3,6739
29	0,2557	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564	3,0380	3,3963	3,6595
30	0,2556	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500	3,0298	3,3852	3,6460
31	0,2555	0,6825	1,3095	1,6955	2,0395	2,4528	2,7440	3,0221	3,3749	3,6335
32	0,2555	0,6822	1,3086	1,6939	2,0369	2,4487	2,7385	3,0149	3,3653	3,6218
33	0,2554	0,6820	1,3077	1,6924	2,0345	2,4448	2,7333	3,0082	3,3563	3,6109
34	0,2553	0,6818	1,3070	1,6909	2,0322	2,4411	2,7284	3,0020	3,3480	3,6007
35	0,2553	0,6816	1,3062	1,6896	2,0301	2,4377	2,7238	2,9961	3,3400	3,5911
36	0,2552	0,6814	1,3055	1,6883	2,0281	2,4345	2,7195	2,9905	3,3326	3,5821
37	0,2552	0,6812	1,3049	1,6871	2,0262	2,4314	2,7154	2,9853	3,3256	3,5737
38	0,2551	0,6810	1,3042	1,6860	2,0244	2,4286	2,7116	2,9803	3,3190	3,5657
39	0,2551	0,6808	1,3036	1,6849	2,0227	2,4258	2,7079	2,9756	3,3127	3,5581
40	0,2550	0,6807	1,3031	1,6839	2,0211	2,4233	2,7045	2,9712	3,3069	3,5510
41	0,2550	0,6805	1,3025	1,6829	2,0195	2,4208	2,7012	2,9670	3,3012	3,5443
42	0,2550	0,6804	1,3020	1,6820	2,0181	2,4185	2,6981	2,9630	3,2959	3,5377
43	0,2549	0,6802	1,3016	1,6811	2,0167	2,4163	2,6951	2,9592	3,2909	3,5316
44	0,2549	0,6801	1,3011	1,6802	2,0154	2,4141	2,6923	2,9555	3,2861	3,5258
45	0,2549	0,6800	1,3007	1,6794	2,0141	2,4121	2,6896	2,9521	3,2815	3,5203
46	0,2548	0,6799	1,3002	1,6787	2,0129	2,4102	2,6870	2,9488	3,2771	3,5149
47	0,2548	0,6797	1,2998	1,6779	2,0117	2,4083	2,6846	2,9456	3,2729	3,5099
48	0,2548	0,6796	1,2994	1,6772	2,0106	2,4066	2,6822	2,9426	3,2689	3,5050
49	0,2547	0,6795	1,2991	1,6766	2,0096	2,4049	2,6800	2,9397	3,2651	3,5005
50	0,2547	0,6794	1,2987	1,6759	2,0086	2,4033	2,6778	2,9370	3,2614	3,4960
51	0,2547	0,6793	1,2984	1,6753	2,0076	2,4017	2,6757	2,9343	3,2579	3,4917
52	0,2546	0,6792	1,2980	1,6747	2,0066	2,4002	2,6737	2,9318	3,2545	3,4877
53	0,2546	0,6791	1,2977	1,6741	2,0057	2,3988	2,6718	2,9293	3,2513	3,4837
54	0,2546	0,6791	1,2974	1,6736	2,0049	2,3974	2,6700	2,9270	3,2481	3,4799



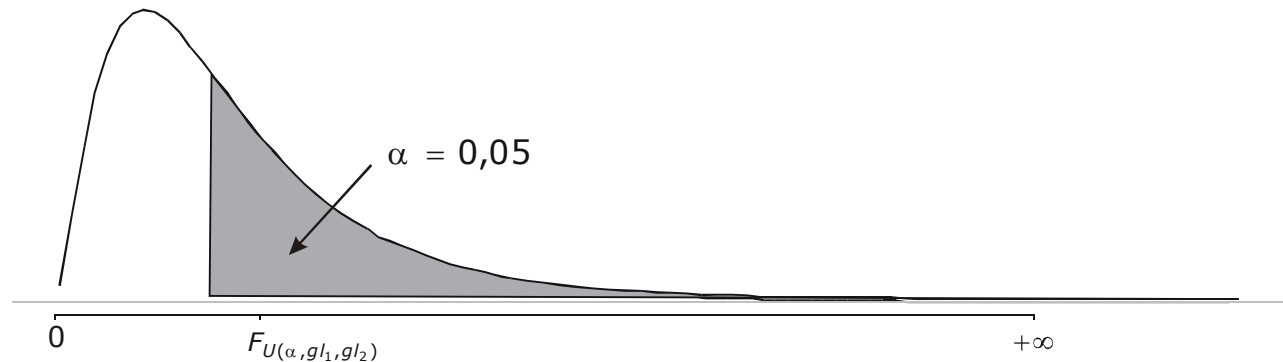
Grados de Libertad	Áreas de la cola superior									
	0,4	0,25	0,1	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
55	0,2546	0,6790	1,2971	1,6730	2,0040	2,3961	2,6682	2,9247	3,2451	3,4765
56	0,2546	0,6789	1,2969	1,6725	2,0032	2,3948	2,6665	2,9225	3,2422	3,4730
57	0,2545	0,6788	1,2966	1,6720	2,0025	2,3936	2,6649	2,9204	3,2395	3,4695
58	0,2545	0,6787	1,2963	1,6716	2,0017	2,3924	2,6633	2,9184	3,2368	3,4663
59	0,2545	0,6787	1,2961	1,6711	2,0010	2,3912	2,6618	2,9164	3,2342	3,4632
60	0,2545	0,6786	1,2958	1,6706	2,0003	2,3901	2,6603	2,9146	3,2317	3,4602
61	0,2545	0,6785	1,2956	1,6702	1,9996	2,3890	2,6589	2,9127	3,2293	3,4572
62	0,2544	0,6785	1,2954	1,6698	1,9990	2,3880	2,6575	2,9110	3,2270	3,4545
63	0,2544	0,6784	1,2951	1,6694	1,9983	2,3870	2,6561	2,9093	3,2247	3,4517
64	0,2544	0,6783	1,2949	1,6690	1,9977	2,3860	2,6549	2,9076	3,2225	3,4491
65	0,2544	0,6783	1,2947	1,6686	1,9971	2,3851	2,6536	2,9060	3,2204	3,4466
66	0,2544	0,6782	1,2945	1,6683	1,9966	2,3842	2,6524	2,9045	3,2184	3,4441
67	0,2544	0,6782	1,2943	1,6679	1,9960	2,3833	2,6512	2,9030	3,2164	3,4418
68	0,2543	0,6781	1,2941	1,6676	1,9955	2,3824	2,6501	2,9015	3,2144	3,4395
69	0,2543	0,6781	1,2939	1,6672	1,9949	2,3816	2,6490	2,9001	3,2126	3,4372
70	0,2543	0,6780	1,2938	1,6669	1,9944	2,3808	2,6479	2,8987	3,2108	3,4350
71	0,2543	0,6780	1,2936	1,6666	1,9939	2,3800	2,6469	2,8974	3,2090	3,4329
72	0,2543	0,6779	1,2934	1,6663	1,9935	2,3793	2,6458	2,8961	3,2073	3,4308
73	0,2543	0,6779	1,2933	1,6660	1,9930	2,3785	2,6449	2,8948	3,2056	3,4289
74	0,2543	0,6778	1,2931	1,6657	1,9925	2,3778	2,6439	2,8936	3,2040	3,4270
75	0,2542	0,6778	1,2929	1,6654	1,9921	2,3771	2,6430	2,8924	3,2024	3,4249
76	0,2542	0,6777	1,2928	1,6652	1,9917	2,3764	2,6421	2,8913	3,2010	3,4232
77	0,2542	0,6777	1,2926	1,6649	1,9913	2,3758	2,6412	2,8902	3,1995	3,4214
78	0,2542	0,6776	1,2925	1,6646	1,9908	2,3751	2,6403	2,8891	3,1981	3,4197
79	0,2542	0,6776	1,2924	1,6644	1,9905	2,3745	2,6395	2,8880	3,1966	3,4180
80	0,2542	0,6776	1,2922	1,6641	1,9901	2,3739	2,6387	2,8870	3,1952	3,4164
81	0,2542	0,6775	1,2921	1,6639	1,9897	2,3733	2,6379	2,8860	3,1939	3,4148
82	0,2542	0,6775	1,2920	1,6636	1,9893	2,3727	2,6371	2,8850	3,1926	3,4132
83	0,2542	0,6775	1,2918	1,6634	1,9890	2,3721	2,6364	2,8840	3,1914	3,4116
84	0,2542	0,6774	1,2917	1,6632	1,9886	2,3716	2,6356	2,8831	3,1901	3,4101
85	0,2541	0,6774	1,2916	1,6630	1,9883	2,3710	2,6349	2,8822	3,1889	3,4086
86	0,2541	0,6774	1,2915	1,6628	1,9879	2,3705	2,6342	2,8813	3,1877	3,4073
87	0,2541	0,6773	1,2914	1,6626	1,9876	2,3700	2,6335	2,8804	3,1866	3,4059
88	0,2541	0,6773	1,2912	1,6624	1,9873	2,3695	2,6329	2,8795	3,1854	3,4046
89	0,2541	0,6773	1,2911	1,6622	1,9870	2,3690	2,6322	2,8787	3,1843	3,4033
90	0,2541	0,6772	1,2910	1,6620	1,9867	2,3685	2,6316	2,8779	3,1832	3,4019
91	0,2541	0,6772	1,2909	1,6618	1,9864	2,3680	2,6309	2,8771	3,1822	3,4006
92	0,2541	0,6772	1,2908	1,6616	1,9861	2,3676	2,6303	2,8763	3,1812	3,3995
93	0,2541	0,6771	1,2907	1,6614	1,9858	2,3671	2,6297	2,8755	3,1802	3,3982
94	0,2541	0,6771	1,2906	1,6612	1,9855	2,3667	2,6291	2,8748	3,1792	3,3970
95	0,2541	0,6771	1,2905	1,6611	1,9852	2,3662	2,6286	2,8741	3,1783	3,3958
96	0,2541	0,6771	1,2904	1,6609	1,9850	2,3658	2,6280	2,8733	3,1773	3,3948
97	0,2540	0,6770	1,2903	1,6607	1,9847	2,3654	2,6275	2,8727	3,1764	3,3937
98	0,2540	0,6770	1,2903	1,6606	1,9845	2,3650	2,6269	2,8720	3,1755	3,3926
99	0,2540	0,6770	1,2902	1,6604	1,9842	2,3646	2,6264	2,8713	3,1746	3,3915
100	0,2540	0,6770	1,2901	1,6602	1,9840	2,3642	2,6259	2,8707	3,1738	3,3905
101	0,2540	0,6769	1,2900	1,6601	1,9837	2,3638	2,6254	2,8700	3,1729	3,3894
102	0,2540	0,6769	1,2899	1,6599	1,9835	2,3635	2,6249	2,8694	3,1720	3,3886
103	0,2540	0,6769	1,2898	1,6598	1,9833	2,3631	2,6244	2,8688	3,1712	3,3875
104	0,2540	0,6769	1,2897	1,6596	1,9830	2,3627	2,6239	2,8681	3,1704	3,3865
105	0,2540	0,6768	1,2897	1,6595	1,9828	2,3624	2,6235	2,8676	3,1697	3,3856
106	0,2540	0,6768	1,2896	1,6594	1,9826	2,3620	2,6230	2,8670	3,1689	3,3848
107	0,2540	0,6768	1,2895	1,6592	1,9824	2,3617	2,6226	2,8664	3,1682	3,3838
108	0,2540	0,6768	1,2894	1,6591	1,9822	2,3614	2,6221	2,8659	3,1674	3,3829
109	0,2540	0,6767	1,2894	1,6590	1,9820	2,3610	2,6217	2,8653	3,1666	3,3820
110	0,2540	0,6767	1,2893	1,6588	1,9818	2,3607	2,6213	2,8648	3,1660	3,3811
111	0,2540	0,6767	1,2892	1,6587	1,9816	2,3604	2,6209	2,8642	3,1653	3,3804
112	0,2539	0,6767	1,2892	1,6586	1,9814	2,3601	2,6204	2,8637	3,1646	3,3795
113	0,2539	0,6767	1,2891	1,6584	1,9812	2,3598	2,6200	2,8632	3,1639	3,3787
114	0,2539	0,6766	1,2890	1,6583	1,9810	2,3595	2,6196	2,8627	3,1633	3,3779
115	0,2539	0,6766	1,2890	1,6582	1,9808	2,3592	2,6193	2,8622	3,1626	3,3772
116	0,2539	0,6766	1,2889	1,6581	1,9806	2,3589	2,6189	2,8617	3,1620	3,3763
117	0,2539	0,6766	1,2888	1,6580	1,9804	2,3586	2,6185	2,8612	3,1613	3,3756
118	0,2539	0,6766	1,2888	1,6579	1,9803	2,3584	2,6181	2,8608	3,1607	3,3749
119	0,2539	0,6766	1,2887	1,6578	1,9801	2,3581	2,6178	2,8603	3,1601	3,3742
120	0,2539	0,6765	1,2886	1,6576	1,9799	2,3578	2,6174	2,8599	3,1595	3,3734
∞	0,6745	1,2816	1,6449	1,9600	2,3263	2,5758	2,8070	3,0902	3,2905	



Tabla 3: Valores críticos de F

Para una combinación dada de grados de libertad en numerador y denominador, el elemento representa el valor crítico de F que corresponde a un área de la cola superior especificada (α). Presentamos aquí las tablas correspondientes a $\alpha = 0,05$ y $\alpha = 0,01$

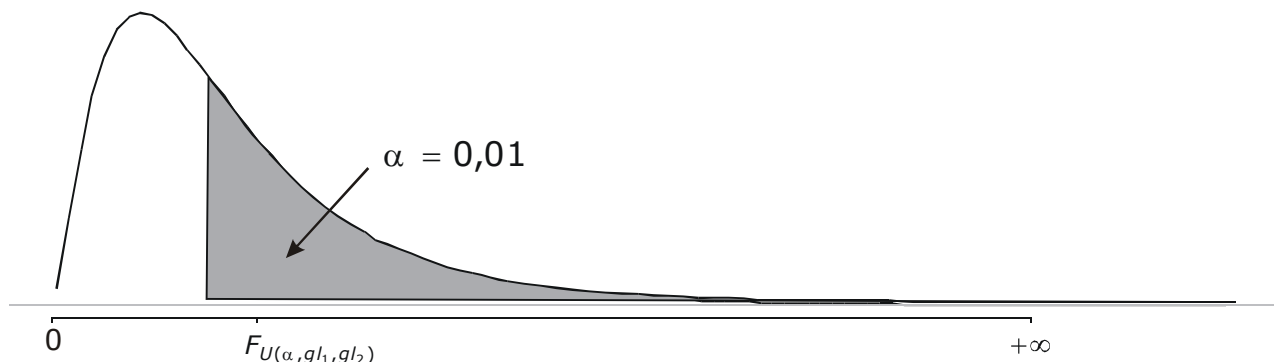
Tabla F con $\alpha = 0,05$



← Denominador, gl_2	Numerador, gl_1																			
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	261,1	252,2	253,3	254,3	
2	18,5	119,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50	
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53	
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63	
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36	
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67	
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23	
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93	
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71	
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54	
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40	
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30	
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21	
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13	
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07	
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01	
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96	
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92	
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88	
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84	
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81	
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78	
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76	
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73	
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71	
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69	
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67	
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65	
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64	
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62	
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51	
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39	
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25	
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,62	1,46	1,39	1,32	1,22	1,00	



Tabla F con $\alpha = 0,01$



← Denominador, gl_2	Numerador, gl_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,76	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00



Tabla 4: Valores críticos del Estadístico de Durbin-Watson

Los valores críticos son de un lado.

n es el número de observaciones.

P es el número de variables independientes.

n	$\alpha = 0,05$										$\alpha = 0,01$									
	$P = 1$		$P = 2$		$P = 3$		$P = 4$		$P = 5$		$P = 1$		$P = 2$		$P = 3$		$P = 4$		$P = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65