MODELO LINEAL

DESARROLLOS Y ANÁLISIS CON EXCEL

Autor: María Alejandra Metelli
Colaboradores:
Marina Russo
Luciano Pizarro

Metelli, María Alejandra
Modelo lineal: desarrollo y análisis con Excel / María Alejandra Metelli ; Contribuciones de Marina Russo ; Luciano Pizarro. - 1a ed - Ciudad Autónoma de Buenos Aires: Universidad de Buenos Aires. Facultad de Ciencias Económicas, 2021.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-29-1916-4

1. Estadísticas. 2. Economía. 3. Finanzas. I. Russo, Marina, colab. II. Pizzarro, Luciano, colab. III. Título.
CDD 005

Índice

Capítulo 1 4
Conceptos Preliminares 4
Introducción 4
Definiciones 4
Diversos tipos de modelos lineales 2
Capítulo 2 4
Correlación y regresión 4
Conceptos básicos 4
Clasificación 5
Análisis para una variable explicativa – independiente - 6
Modelo de regresión lineal 7
Modelo lineal general 8
Capítulo 3 23
Análisis estadístico 23
Supuestos de Gauss Markov 23
Propiedades de los estimadores mínimo cuadráticos 28
Distribución F de Snedecor 39
Aplicaciones a la inferencia estadística 40
Test de hipótesis para variables conjuntas 45
Compatibilidad con los test de hipótesis para parámetros individuales 47
Coeficiente de determinación corregido o ajustado 47
Capítulo 4 50
Un caso particular 50
Determinación de las varianzas y covarianza de los estimadores 52
Cálculo de la varianza de $\boldsymbol{Y}$ 55
Coeficiente de correlación 57
Relación entre el coeficiente de correlación y la pendiente 58
Capítulo 5 60
Modelos no lineales 60
Parábola 60

Modelo exponencial 60
Modelo potencial 61
Capítulo 6 63
Distintos métodos de resolución aplicando Excel 63
Primer método: Método de determinantes para resolución de sistemas de ecuaciones 64
Resolución el sistema a través de los determinantes 67
Resolución en forma matricial 70
El cuadro siguiente muestra los valores de las varianzas y de los desvíos estimados de los estimadores mínimo cuadráticos: 77
La función análisis de datos se encuentra en la solapa de datos en la planilla de Excel. 77
Lectura de los resultados obtenidos 80
Capítulo 7 82
Ejercicios 82
Apéndice 97
1. Aplicación directa de la herramienta regresión 97
2. Resolución en forma matricial 99

Capítulo 1

Conceptos Preliminares

Introducción

Uno de los propósitos de la ciencia es describir y predecir eventos del mundo en el que se vive. Una de las maneras de hacerlo es encontrando una fórmula o ecuación que relacione cantidades del mundo real. Se puede estar interesado, por ejemplo, en la relación entre consumo e ingreso o, entre las ventas de determinado producto y el precio del mismo. Es conclusión, lo primero que se debe realizar es plantear cuáles son las variables que se pretenden analizar a fin de determinar el tipo de relación que las vincula.

Cuando se dice que existe una relación funcional entre una serie de variables significa que en la realidad la función describe la relación entre las variables en forma aproximada.

Por ejemplo, si se clasifica a los individuos por peso y se evalúa la estatura promedio de éstos en cada una de esas clases, se encuentra que existe una relación funcional entre peso y estatura promedio de todos los individuos que tienen un peso determinado.

Definiciones

Se define como modelo una ecuación matemática que incluye variables aleatorias, variables matemáticas y parámetros.

Si está dada la distribución de las variables aleatorias, ésta puede ser considerado como parte del modelo y puede haber parámetros desconocidos en la distribución que, en consecuencia, deban ser estimados.

En forma análoga, se define como 'modelo lineal' a una ecuación que incluye variables aleatorias, variables matemáticas y parámetros y que, además, es lineal tanto en los parámetros como en las variables.

Por ejemplo, si a_0, a_1 y a_2 son parámetros desconocidos, entonces

$$a_0 + a_1X + a_2Y = 0$$

es un modelo lineal.

Las siguientes expresiones

$$a_0 + a_1X + a_2Y\cos X = 0$$

$$a_0Y + a_1 \operatorname{sen} X + a_2 = 0$$

no responden en forma completa a la definición de modelo lineal dado que no son lineales en las variables.

Se considera entonces el siguiente modelo lineal:

$$a_0 + a_1X + a_2Y = 0 \qquad (1)$$

Donde X e Y pueden ser variables matemáticas o aleatorias, observables o no observables. Este modelo puede escribirse como:

$$Y = -a_0/a_2 - a_1/a_2X \qquad Si \qquad a_2 \neq 0 \qquad (2)$$

O como,

$$X = -a_0/a_1 - a_2/a_1Y \qquad Si \qquad a_1 \neq 0 \qquad (3)$$

En (1), X e Y actúan de la misma forma en la ecuación. En (2), Y se denomina variable dependiente, X variable independiente. En la ecuación (3) la variable dependiente es X, mientras que la variable Y es independiente.[1]

Si X e Y pudieran ser medidas sin error y un experimentador deseara predecir Y a partir del conocimiento de X, entonces tendría que usar la ecuación (2). Como los valores de a_i son desconocidos, deben ser observados al menos tres pares de valores (X, Y) y resolver las ecuaciones resultantes donde las a_i serán las incógnitas.

Supóngase ahora que se quiere efectuar una predicción en el modelo dado por la siguiente expresión:

$$Y^* = -a_0/a_2 - a_1/a_2X$$

[1] NOTA: los términos *independiente* y *dependiente* son usados aquí en un sentido matemático y no estadístico.

Donde $-a_0/a_2$ y - a_1/a_2 son desconocidos e Y^* es no observable.

Entonces, si se pudiera observar Y, donde $Y = Y^* + u$, el modelo anterior puede escribirse como:

$$Y = -a_0/a_2 - a_1/a_2 X + u \qquad (4)$$

Como no hay una relación funcional exacta entre X e Y no se pueden obtener los valores exactos de los coeficientes a_0/a_2 y a_1/a_2 sino estimarlos mediante la aplicación de métodos estadísticos.

Diversos tipos de modelos lineales

Existen cinco tipos de modelos lineales que se usan comúnmente, a saber:

1. Modelos relacionados funcionalmente
2. Modelos relacionados en promedio
3. Modelos de regresión
4. Modelos experimentales
5. Modelos de componente varianza

Una vez seleccionado el tipo de modelo y habiendo obtenido los datos necesarios, se pueden usar las técnicas estadísticas para llevar a cabo el correspondiente análisis.

Se dan a continuación las características básicas de los tres primeros modelos señalados:

1) *Modelos relacionados funcionalmente*

Se caracterizan por presentar una relación funcional entre las variables matemáticas, las cuales no pueden ser observadas debido a la existencia de errores de medición. Se suponen dos cantidades Y^* y X que satisfacen la ecuación:

$$Y^* = a_0 + a_1 X \qquad (5)$$

Esto implica que se puede observar X, pero no Y^*. Se supone que hay un error de medición y, en lugar de observar Y^*, se observa $Y = Y^* + u$. Se supone que u es un

error aleatorio con media cero, de donde surge entonces que $E(Y) = Y^*$. De este modo, se puede escribir el modelo como:

$$Y - u = a_0 + a_1X \qquad o \qquad Y = a_0 + a_1X + u \tag{6}$$

Es importante señalar que Y^* y u son variables aleatorias mientras que Y y X son variables observables.

2) *Modelos relacionados en promedio*

Estos modelos se caracterizan por contener errores de ecuación. Supóngase que una relación funcional está dada por $f(Y, X_1, X_2, \ldots, X_k) = 0$.

Y que la misma puede ser escrita del siguiente modo:

$$Y = a_0 + a_1X_1 + a_2X_2 + g(X_3 \ldots X_k) \tag{7}$$

Si el término $g(X_3,\ldots X_k)$ se desprecia, se ve claramente que Y no está exactamente determinada por los pares de valores X_1, X_2 dado que las otras variables también inciden.

Supóngase que, sin embargo, se asume que las variables X_1, X_2 pueden predecir en forma confiable a Y pero que, además, se necesita de los valores X_3, $X_4 \ldots X_k$ para predecir en forma exacta. Es decir, puede que no se conozca de qué variables se trata pero puede considerarse que $g(X_3,\ldots X_k)$ actúa como una variable aleatoria en razón de que X_3, $X_4 \ldots X_k$ varían. Si esto es cierto, el modelo puede escribirse como:

$$Y = a_0 + a_1X_1 + a_2X_2 + u \ ^{2} \tag{8}$$

En este modelo, Y es una variable aleatoria mientras que X_1 y X_2 no lo son, son variables matemáticas predeterminadas.

De este modo el valor esperado de la variable dependiente Y no da por resultado la relación funcional con X_1 y X_2, es decir $E(Y) = Y$. Es lo mismo que decir que la relación funcional es:

$$Y = a_0 + a_1X_1 + a_2X_2 + g(X_3 \ldots X_k)$$

---

[2] NOTA: Debe puntualizarse que X_i pueden a su vez ser funciones de otras variables, por ejemplo: $X_1 = t \quad \wedge \quad X_2 = 4 \cdot t$

Mientras que

$$E(Y) = a_0 + a_1X_1 + a_2X_2$$

Es decir, es igual al valor promedio de Y.

El modelo expuesto se utiliza en problemas en los cuales muchos factores influyen en Y y, donde no todos los factores pueden ser conocidos.

Supóngase que se quieren predecir las ventas de determinado producto. Los factores que pueden influir son:

- X_1: precio del producto
- X_2: precio del artículo sustituto
- X_3: ingreso del consumidor
- X_4: ahorros de períodos anteriores

De acuerdo con lo expuesto, se puede considerar el siguiente modelo:

$$Y = a_0 + a_1X_1 + a_2X_2 + u$$

donde u incluye al resto de las variables enunciadas (X_3 y X_4).

Una de las formas para determinar si el modelo expuesto puede ser aceptado es calcular la varianza. De este modo, puede concluirse que: si la varianza de u es pequeña quiere decir que este modelo funciona; en cambio, si es grande quiere decir que habrá que considerar en el modelo, alguna de las variables incluidas en el residuo.

3) *Modelos de regresión*

Estos modelos se caracterizan por el hecho de que todas las variables consideradas en el modelo son aleatorias.

Sea $f(Y, X)$, la función de distribución conjunta de Y y X de modo tal que

$$E(Y/X_1 = x_1) = a_0 + a_1X_1.$$

Se puede escribir entonces:

$$Y = a_0 + a_1X_1 + u$$

Donde:

$$Y = Y^* + u \qquad \wedge \qquad X = X^* + d$$

Ya que Y^* y X^* son variables no observables.

Capítulo 2

Correlación y regresión

Conceptos básicos

El propósito del análisis de correlación es estudiar conjuntamente un grupo de variables con el principal objetivo de ver cómo se relacionan entre sí.

Cuando dos variables están ligadas en su comportamiento se dice que están correlacionadas. El monto de ventas e inversión en publicidad representan un ejemplo de la situación mencionada.

Sin embargo, se podría estar interesado en resolver un problema de predicción o estimación como, por ejemplo, "*¿a qué nivel debe ascender el monto de ventas para un determinado nivel de inversión en publicidad?*". Este problema recibe el nombre de regresión.

El análisis de regresión permite entonces expresar matemáticamente la relación existente entre las variables, mientras que el análisis de correlación permite medir la intensidad de la relación que existe entre las variables.

En caso de que las variables estén altamente correlacionadas la ecuación de regresión puede utilizarse para estimar los valores de la variable dependiente para determinado valor que tome la variable independiente.

Consideraciones acerca de la correlación

Corresponde resaltar los siguientes conceptos:

- Un alto valor del coeficiente de correlación no necesariamente determina una **causalidad** entre las variables; dos variables pueden aparecer correlacionadas por casualidad y no porque exista una relación de dependencia entre ellas.
- Es necesario que las variables aparezcan depuradas de las influencias de otras variables; dos variables pueden mostrar estrecha relación porque hay una tercera que exagera el grado de correlación.

- Puede ocurrir también que el coeficiente de correlación sea cercano a |1| porque el tamaño de la muestra resulta insuficiente.
- La proyección por regresión y correlación es válida en tanto se mantengan los supuestos y circunstancias implícitas en los datos y antecedentes disponibles.

Clasificación

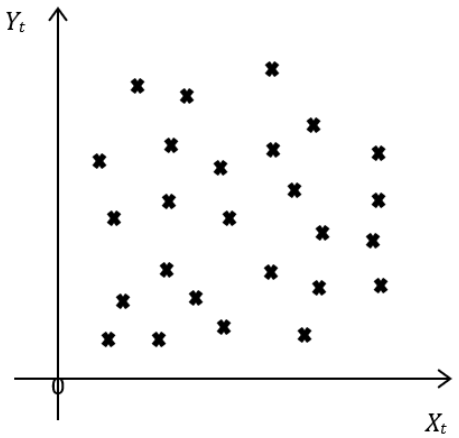
Atendiendo a		Clasificación
La cantidad de variables	Correlación simple	Cuando se estudia el grado de asociación que existe entre un par de variables: una dependiente y otra independiente.
	Correlación múltiple	Cuando se estudia el grado de asociación que simultáneamente existe entre la variable dependiente y dos o más variables independientes.
La forma de la ecuación de la regresión		Según el tipo de ecuación de regresión se tendrá correlación lineal, correlación parabólica, correlación exponencial, correlación logarítmica, entre otros.
La relación entre las variables	Correlación directa o positiva	Cuando por variaciones en la variable independiente se producen variaciones en el mismo sentido en la variable dependiente.
	Correlación inversa o negativa	Cuando por variaciones en la variable independiente se producen variaciones en sentido inverso en la variable dependiente.

Análisis para una variable explicativa – independiente -

Al recopilar la información, los datos pueden resumirse en la siguiente tabla:

Unidad estadística	Y_i	X_i
1	Y_1	X_1
2	Y_2	X_2
3	Y_3	X_3
…	…	…
T	Y_T	X_T

El primer paso que debe llevarse a cabo cuando se analiza la relación existente entre dos variables consiste en graficar los pares ordenados que se obtienen de cada unidad estadística. El gráfico que se obtiene recibe el nombre de nebulosa o diagrama de dispersión:



1. Si hay una correlación lineal perfecta directa, los puntos se ubican sobre una recta de pendiente positiva.

2. Si hay una correlación lineal perfecta inversa, los puntos se ubican sobre una recta de pendiente negativa.

En realidad, no es usual encontrarse con correlaciones perfectas. Al graficar los pares ordenados, se deduce que:

1. Si hay una correlación lineal no perfecta directa los puntos tienden a ubicarse sobre una recta de pendiente positiva.

2. Si hay una correlación lineal no perfecta inversa, los puntos tienden a ubicarse sobre una recta de pendiente negativa.

3. Puede no existir ninguna relación entre las variables ya que a valores pequeños de X algunas veces le corresponden valores pequeños de Y y otras veces valores grandes. Se dice entonces que estas variables no están correlacionadas.

Modelo de regresión lineal

Los métodos de correlación se emplean especialmente cuando interesa conocer el tipo de interrelación existente entre dos o más variables. La finalidad de estudiar la relación existente entre variables es utilizar los resultados en estimaciones o predicciones de una variable particular.

Un objetivo frecuente en la investigación es la especificación de una relación funcional entre dos variables $Y = f(X)$, donde Y es la variable dependiente y X la variable independiente. Al no poder esperar una explicación perfecta, el modelo se escribe como:

$$Y = f(X) + u \tag{9}$$

Donde u es una variable aleatoria llamada residuo o error. El error puede surgir de los errores de medición de Y o de las imperfecciones en la especificación de la función $f(X)$ (por ejemplo, puede haber muchas otras variables que influyan sobre Y además de X), según lo detallado en el capítulo anterior. En consecuencia, al ser u una variable aleatoria, también lo es Y.

De este modo, la forma general del modelo lineal está dada por la siguiente expresión:

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + \cdots + a_kX_{kt} + u_t \tag{10}$$

Donde:

- t: representa el número de observación (t = 1,2…T).
- Y_t es la variable que se desea explicar, que recibe el nombre de *regresando*.
- X_{it} son las variables explicativas o regresores, que se suponen matemáticas, no estocásticas, es decir, medidas sin errores aleatorios (i = 1, 2, … ,k).
- u_t es el residuo de la regresión que puede obedecer a tres causas:
    1. Errores muestrales (de medición)
    2. Errores provenientes de otras variables de menor importancia que las X_{it} consideradas, que no han sido explicitadas en la formulación del modelo y, en consecuencia, no incluidas en su especificación.
    3. Existencia de otra relación funcional, distinta a la lineal.

Modelo lineal general

De acuerdo con lo previamente expuesto, se tiene entonces que:

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + \cdots + a_iX_{it} \cdots + a_kX_{kt} + u_t \qquad (11)$$

es la expresión general del modelo lineal. Es lineal porque tanto la parte funcional del modelo como los coeficientes son lineales. Es aleatorio porque incluye una parte aleatoria, u_t.

El subíndice i hace referencia a la variable de que se trata, mientras que t indica el número de observación correspondiente. Se supone que t = 1, 2, ...,T, donde T representa la última observación que se ha realizado.

Las observaciones surgen de las series estadísticas de las que se dispone. Pueden referirse a:

1. Series de tiempo
2. Series de tipo muestral (por ejemplo, encuestas)

Generalmente, estas series vienen presentadas en cuadros o tablas, del siguiente modo:

t	Y_t	X_{1t}	X_{2t}	**…**	X_{kt}
1	Y_1	X_{11}	X_{21}	…	X_{k1}
2	Y_2	X_{12}	X_{22}	…	X_{k2}
3	Y_3	X_{13}	X_{23}	…	X_{k3}
…	…	…	…	…	…
T	Y_T	X_{1T}	X_{2T}	…	X_{kT}

Debe entonces tenerse presente que:

Los a_i (i = 1,2…k) son parámetros poblacionales “desconocidos”,”no aleatorios”.

u_t es una variable aleatoria cuyo valor se desconoce.

Las X_{it} son variables matemáticas conocidas y las Y_t son variables aleatorias de las que puede obtenerse información. Las variables mencionadas son variables observables.

El principal objetivo es estimar de alguna manera el valor de esos parámetros poblacionales desconocidos a_i (i = 1,2, …, k).

Se tendrá entonces que:

$\hat{a}_0 = \hat{a}_0 (Y_t, X_{1t} \ldots X_{kt})$ es estimador de a_0

$\hat{a}_1 = \hat{a}_1 (Y_t, X_{1t} \ldots X_{kt})$ es estimador de a_1

… …

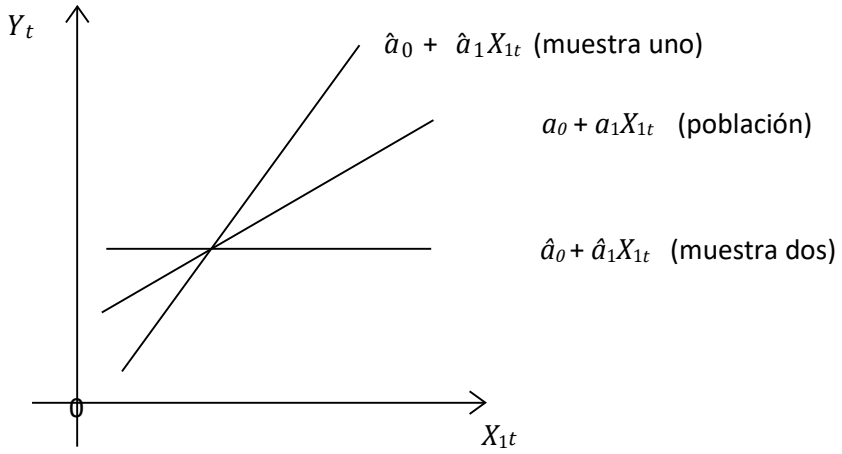
$\hat{a}_k = \hat{a}_k (Y_t, X_{1t} \ldots X_{kt})$ es estimador de a_k

Se debe recordar que la estimación es el resultado numérico que toma el estimador a partir de una muestra dada mientras que el ‘estimador’ es un operador, una

variable aleatoria porque está compuesto por distintas variables aleatorias que provienen de los posibles resultados que puede tomar la muestra; es decir, depende de la muestra, que es aleatoria.

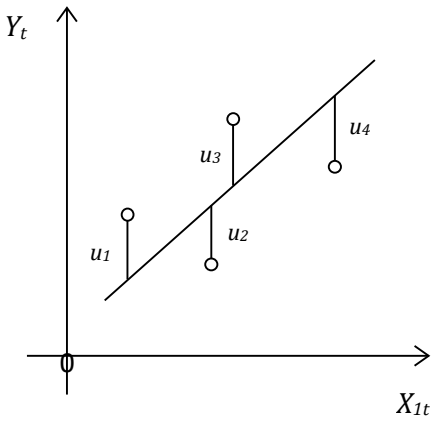
Por ejemplo, en el esquema de regresión para una sola variable explicativa, se tendrán muchas rectas que variarán de acuerdo con las muestras que se seleccionan.

La pregunta que se debiera formular el investigador es: ¿qué relaciones existen entonces entre las a_i con todas las $\hat{a}_i$ correspondientes a las posibles muestras que puedan obtenerse? En el siguiente gráfico pueden observarse dos de las posibles rectas obtenidas a partir de diferentes muestras.



Principio de los mínimos cuadrados clásicos

Si se observa el gráfico para una sola variable explicativa:



Puede concluirse que existen puntos observados sobre y por debajo de la recta de regresión. En consecuencia, existirán desvíos positivos y negativos que, en promedio, han de compensarse; lo que indica que su suma resulta siempre igual a cero. Entonces, si se tiene en cuenta que se debe seleccionar la recta de tal forma que los desvíos sean mínimos, resulta necesario que estos desvíos sean elevados al cuadrado y proceder a minimizar su suma. Lo expuesto previamente puede escribirse como:

$$mín\ \emptyset(a_0, a_1, \dots, a_i, \dots, a_k) = mín \sum_{t=1}^{T} u_t^2 \qquad (12)$$

El desarrollo se realiza para el caso del modelo lineal general. A partir de (11), se reemplaza el valor de u_t por $Y_t - (a_0 + a_1 X_{1t} + \cdots + a_k X_{kt})$.

Se tiene entonces que:

$$mín \sum_{t=1}^{T} u_t^2 = mín \sum_{t=1}^{T} [Y_t - (a_0 + a_1 X_{1t} + \cdots + a_k X_{kt})]^2 \qquad (13)$$

Para minimizar la función, se plantean las condiciones de primer orden, es decir, se obtienen las derivadas primeras respecto a cada uno de los parámetros desconocidos y, luego, se igualan a cero.

Se muestra a continuación el sistema de derivadas obtenido:

$$\begin{cases} \dfrac{\partial \phi}{\partial a_0} = 2\sum_{t=1}^{T}\left[Y_t - (a_0 + a_1 X_{1t} + \ldots + a_k X_{kt})\right](-1) \\ \dfrac{\partial \phi}{\partial a_1} = 2\sum_{t=1}^{T}\left[Y_t - (a_0 + a_1 X_{1t} + \ldots + a_k X_{kt})\right](-X_{1t}) \\ \dfrac{\partial \phi}{\partial a_2} = 2\sum_{t=1}^{T}\left[Y_t - (a_0 + a_1 X_{1t} + \ldots + a_k X_{kt})\right](-X_{2t}) \\ \qquad\qquad \ldots \\ \dfrac{\partial \phi}{\partial a_k} = 2\sum_{t=1}^{T}\left[Y_t - (a_0 + a_1 X_{1t} + \ldots + a_k X_{kt})\right](-X_{kt}) \end{cases}$$

Se igualan todas las ecuaciones a cero. De este modo, los valores de a_0, a_1 ...a_k que hacen igual a cero estas expresiones reciben el nombre de estimadores mínimo-cuadráticos ($\hat{a}_i$), o sea:

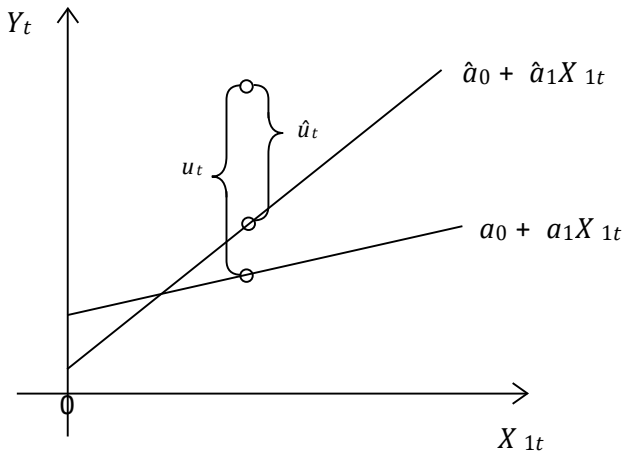
$$\begin{aligned} &2\sum_{t=1}^{T}\left[Y_t - (\hat{a}_0 + \hat{a}_1 X_{1t} + \ldots + \hat{a}_k X_{kt})\right](-1) = 0 \\ &2\sum_{t=1}^{T}\left[Y_t - (\hat{a}_0 + \hat{a}_1 X_{1t} + \ldots + \hat{a}_k X_{kt})\right](-x_{1t}) = 0 \\ &2\sum_{t=1}^{T}\left[Y_t - (\hat{a}_0 + \hat{a}_1 X_{1t} + \ldots + \hat{a}_k X_{kt})\right](-x_{2t}) = 0 \\ &\qquad\qquad \vdots \\ &2\sum_{t=1}^{T}\left[Y_t - (\hat{a}_0 + \hat{a}_1 X_{1t} + \ldots + \hat{a}_k X_{kt})\right](-X_{kt}) = 0 \end{aligned} \qquad (14)$$

Conforme lo expuesto, se tiene que:

$Y_t = a_0 + a_1 X_{1t} + \ldots + a_k X_{kt} + u_t \Rightarrow$ es el modelo poblacional

$Y_t = \hat{a}_0 + \hat{a}_1 X_1 t + \ldots + \hat{a}_k X_{kt} + \hat{u}_t \Rightarrow$ es el modelo estimado

Es decir, para el caso de una sola variable explicativa, el gráfico resultante es:



El gráfico entonces muestra los distintos desvíos – residuos -: poblacionales (u_t desconocidos) y muestrales ($\hat{u}_t$ estimados).

Operando algebraicamente las ecuaciones correspondientes al sistema expuesto en (14), se tiene que:

$$\begin{cases} \sum_{t=1}^{T} Y_t = \sum_{t=1}^{T} (\hat{a}_0 + \hat{a}_1 X_{1t} + \hat{a}_2 X_{2t} + \cdots + \hat{a}_k X_{kt}) \\ \sum_{t=1}^{T} Y_t X_{1t} = \sum_{t=1}^{T} (\hat{a}_0 + \hat{a}_1 X_{1t} + \hat{a}_2 X_{2t} + \cdots + \hat{a}_k X_{kt})\, X_{1t} \\ \sum_{t=1}^{T} Y_t X_{2t} = \sum_{t=1}^{T} (\hat{a}_0 + \hat{a}_1 X_{1t} + \hat{a}_2 X_{2t} + \cdots + \hat{a}_k X_{kt})\, X_{2t} \\ \vdots \\ \sum_{t=1}^{T} Y_t X_{kt} = \sum_{t=1}^{T} (\hat{a}_0 + \hat{a}_1 X_{1t} + \hat{a}_2 X_{2t} + \cdots + \hat{a}_k X_{kt})\, X_{kt} \end{cases}$$

Debe tenerse presente que $\sum_{t=1}^{T} \hat{a}_0 = T\hat{a}_0$ dado que $\hat{a}_0$ adoptará un único valor para la muestra que se seleccione. De este modo, el sistema de ecuaciones anterior puede expresarse del siguiente modo:

$$
\begin{cases}
\sum_{t=1}^{T} Y_t = T\hat{a}_0 + \hat{a}_1 \sum_{t=1}^{T} X_{1t} + \hat{a}_2 \sum_{t=1}^{T} X_{2t} + \cdots + \hat{a}_t \sum_{t=1}^{T} X_{kt} \\
\sum_{t=1}^{T} Y_t X_{1t} = \hat{a}_0 \sum_{t=1}^{T} X_{1t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}^2 + \hat{a}_2 \sum_{t=1}^{T} X_{1t} X_t + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{1t} X_{kt} \\
\sum_{t=1}^{T} Y_t X_{2t} = \hat{a}_0 \sum_{t=1}^{T} X_{1t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t} + \hat{a}_2 \sum_{t=1}^{T} X_{2t} + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{2t} X_{kt} \\
\cdots \\
\sum_{t=1}^{T} Y_t X_{kt} = \hat{a}_0 \sum_{t=1}^{T} X_{kt} + \hat{a}_1 \sum_{t=1}^{T} X_{1t} X_{kt} + \hat{a}_2 \sum_{t=1}^{T} X_{2t} X_{kt} + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{kt}^2
\end{cases}
$$

Es decir, es un sistema de (k+1) ecuaciones con (k+1) incógnitas, $\hat{a}_0$, $\hat{a}_1 \ldots \hat{a}_k$

Matricialmente, este sistema puede escribirse como:

$$
\begin{bmatrix}
\sum_{t=1}^{T} Y_t \\
\sum_{t=1}^{T} Y_t X_{1t} \\
\sum_{t=1}^{T} Y_t X_{2t} \\
\vdots \\
\sum_{t=1}^{T} Y_t X_{kt}
\end{bmatrix}
=
\begin{bmatrix}
T & \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{2t} & \cdots & \sum_{t=1}^{T} X_{kt} \\
\sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 & \sum_{t=1}^{T} X_{1t} X_{2t} & \cdots & \sum_{t=1}^{T} X_{1t} X_{kt} \\
\sum_{t=1}^{T} X_{2t} & \sum_{t=1}^{T} X_{1t} X_{2t} & \sum_{t=1}^{T} X_{2t}^2 & \cdots & \sum_{t=1}^{T} X_{2t} X_{kt} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\sum_{t=1}^{T} X_{kt} & \sum_{t=1}^{T} X_{1t} X_{kt} & \sum_{t=1}^{T} X_{2t} X_{kt} & \cdots & \sum_{t=1}^{T} X_{kt}^2
\end{bmatrix}
\begin{bmatrix}
\hat{a}_0 \\
\hat{a}_1 \\
\hat{a}_2 \\
\vdots \\
\hat{a}_k
\end{bmatrix}
$$

La solución de este sistema de ecuaciones está dada por la siguiente expresión:

$$\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_k \end{bmatrix} = \begin{bmatrix} T & \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{2t} & \cdots & \sum_{t=1}^{T} X_{kt} \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 & \sum_{t=1}^{T} X_{1t}X_{2t} & \cdots & \sum_{t=1}^{T} X_{1t}X_{kt} \\ \sum_{t=1}^{T} X_{2t} & \sum_{t=1}^{T} X_{1t}X_{2t} & \sum_{t=1}^{T} X_{2t}^2 & \cdots & \sum_{t=1}^{T} X_{2t}X_{kt} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{t=1}^{T} X_{kt} & \sum_{t=1}^{T} X_{1t}X_{kt} & \sum_{t=1}^{T} X_{2t}X_{kt} & \cdots & \sum_{t=1}^{T} X_{kt}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} Y_t \\ \sum_{t=1}^{T} Y_t X_{1t} \\ \sum_{t=1}^{T} Y_t X_{2t} \\ \vdots \\ \sum_{t=1}^{T} Y_t X_{kt} \end{bmatrix}$$

Expresión matricial del modelo

Se sabe que:

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + \cdots + a_kX_{kt} + u_t \qquad (t = 1,2,\cdots,T) \qquad (15)$$

Para t = 1,2, ⋯, T, se tiene que:

$$t = 1 \rightarrow Y_1 = a_0 + a_1 X_{11} + a_2 X_{21} + \ldots + a_k X_{k1} + u_1$$

$$t = 2 \rightarrow Y_2 = a_0 + a_1 X_{12} + a_2 X_{22} + \ldots + a_k X_{k2} + u_2$$

$$t = 3 \rightarrow Y_3 = a_0 + a_1 X_{13} + a_2 X_{23} + \ldots + a_k X_{k3} + u_3$$

$$\ldots$$

$$t = T \rightarrow Y_T = a_0 + a_1 X_{1T} + a_2 X_{2T} + \ldots + a_k X_{kT} + u_T$$

Además, cada valor observado puede también escribirse del siguiente modo:

$$Y_1 = \hat{Y}_1 + \hat{u}_1 = \hat{a}_0 + \hat{a}_1 X_{11} + \hat{a}_2 X_{21} + \cdots + \hat{a}_k X_{k1} + \hat{u}_1$$

$$Y_2 = \hat{Y}_2 + \hat{u}_2 = \hat{a}_0 + \hat{a}_1 X_{12} + \hat{a}_2 X_{22} + \cdots + \hat{a}_k X_{k2} + \hat{u}_2$$

$$Y_3 = \hat{Y}_3 + \hat{u}_3 = \hat{a}_0 + \hat{a}_1 X_{13} + \hat{a}_2 X_{23} + \cdots + \hat{a}_k X_{k3} + \hat{u}_3$$

$$\ldots$$

$$Y_T = \hat{Y}_T + \hat{u}_T = \hat{a}_0 + \hat{a}_1 X_{1T} + \hat{a}_2 X_{2T} + \cdots + \hat{a}_k X_{kT} + \hat{u}_T$$

Se definen entonces los siguientes vectores y matrices[3]:

$$\textit{Vector}\ \ Y_{Tx1} \quad \Rightarrow \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_T \end{bmatrix}$$

$$\text{Matriz}\ X_{Tx(k+1)} \quad \Rightarrow \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1T} & X_{2T} & \dots & X_{kT} \end{bmatrix}$$

$$\text{Vector}\ a_{(k+1)x1} \quad \Rightarrow \quad a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Teniendo en cuenta lo expuesto, el vector Y puede escribirse como:

[3] NOTA: El subíndice de cada uno de los vectores o matrices indica el orden de los respectivos vectores o matrices, según se trate.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1T} & X_{2T} & \dots & X_{kT} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_k \end{bmatrix}$$

Que puede expresarse matricialmente como:

$$Y = Xa + u$$

Donde:

- X es la matriz de todos los datos de las series correspondientes a las variables explicativas: $X_{T\times(k+1)}$
- Y es el vector de las observaciones correspondientes a la variable que se desea explicar Y_{Tx1}
- a es el vector de los parámetros desconocidos: $a_{(k+1)\times 1}$.
- u es el vector correspondiente a los residuos; u_{Tx1}

Por otro lado, se tiene que:

$$Y = X\hat{a} + \hat{u}$$

$$\hat{Y} = X\hat{a}$$

Donde:

$$\hat{a}_{(k+1)\times 1} = \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_k \end{bmatrix} \qquad \hat{u}_{(T\times 1)} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix}$$

De este modo, a fin de poder obtener la matriz A:

$$A = \begin{bmatrix} T & \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{2t} & \dots & \sum_{t=1}^{T} X_{kt} \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 & \sum_{t=1}^{T} X_{1t}X_{2t} & \dots & \sum_{t=1}^{T} X_{1t}X_{kt} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{t=1}^{T} X_{kt} & \sum_{t=1}^{T} X_{1t}X_{kt} & \sum_{t=1}^{T} X_{2t}X_{kt} & \dots & \sum_{t=1}^{T} X_{kt}^2 \end{bmatrix}$$

Se debe relacionar la matriz X:

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1T} & X_{2T} & \dots & X_{kT} \end{bmatrix}$$

Con su traspuesta X^t:

$$X^t = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{12} & \cdots & X_{1T} \\ X_{21} & X_{22} & \cdots & X_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kT} \end{bmatrix}$$

En consecuencia, se tiene que:

$$A = X^t X$$

Del mismo modo, se define el vector B como:

$$B = \begin{bmatrix} \sum_{i=1}^{T} Y_t \\ \sum_{t=1}^{T} Y_t X_{1t} \\ \sum_{t=1}^{T} Y_t X_{2t} \\ \vdots \\ \sum_{t=1}^{T} Y_t X_{kt} \end{bmatrix}$$

B es de orden (k+1)×1. Si se tiene en cuenta que X^t es de orden (k+1)×T y que Y es de orden (T × 1), entonces el vector B puede escribirse como:

$$B = \left(X^t Y\right)$$

En resumen, en función de lo previamente expuesto, resulta que:

$$Y = Xa + u$$

$$Y = \hat{Y} + \hat{u}$$

Donde:

$$\hat{Y} = X\hat{a}$$

$$\hat{a} = \left(X^t X\right)^{-1} X^t Y$$

Descomposición de las variaciones. Coeficiente de determinación.

Sea:

$$Y_t = \hat{Y}_t + \hat{u}_t \qquad (16)$$

un término genérico del modelo.

Se sabe además que el promedio de las observaciones resulta equivalente a:

$$\bar{Y} = \bar{\hat{Y}} + \bar{\hat{u}} \tag{17}$$

Donde $\bar{\hat{u}} = 0$ ya que $\sum_{t=1}^{T} \hat{u}_t = 0.$

De este modo, restando miembro a miembro (17) y (16), se tiene que:

$$Y_t - \bar{Y} = \hat{Y}_t - \overline{\hat{Y}} + \hat{u}_t$$

La expresión anterior puede escribirse como:

$$y_t = \hat{y}_t + \hat{u}_t$$

Donde las letras minúsculas representan desvíos, es decir:

$$y_t = Y_t - \bar{Y}$$

$$\hat{y}_t = \hat{Y}_t - \bar{\hat{Y}}$$

Se elevan al cuadrado ambos miembros de la igualdad y, se suman para todos los valores de *t*:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} (\hat{y}_t + \hat{u}_t)^2$$

Se opera algebraicamente:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{u}_t^2 + 2\sum_{t=1}^{T} \hat{y}_t \hat{u}_t \tag{18}$$

El tercer término del lado derecho de la igualdad puede escribirse como $X^t Y - X^t Y$ a partir del desarrollo que se muestra a continuación:

Se parte de:

$$\hat{u} = Y - \hat{Y} = Y - X\hat{a} = Y - X\left(X^tX\right)^{-1}X^tY$$

Pre multiplicando ambos miembros de la igualdad por X^t, se tiene que:

$$X^t\hat{u} = X^tY - X^tX\left(X^tX\right)^{-1}X^tY = X^tY - X^tY$$

O sea:

$$X^t\hat{u} = 0$$

De este modo, la expresión (18) resultante de la descomposición de las variaciones se reduce a:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{u}_t^2 \qquad (19)$$

Donde:

- $\sum_{t=1}^{T} y_t^2$: recibe el nombre de variación total
- $\sum_{t=1}^{T} \hat{y}_t^2$: recibe el nombre de variación explicada o de regresión
- $\sum_{t=1}^{T} \hat{u}_t^2$: recibe el nombre de variación no explicada o residual

Dividiendo (19) por T se obtiene que:

$$\frac{\sum_{t=1}^{T} y_t^2}{T} = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{T} + \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T}$$

$$S_y^2 = S_{\hat{y}}^2 + S_{\hat{u}}^2$$

Donde:

- S_y^2: varianza de la variable Y observada
- $S_{\hat{y}}^2$: varianza de la parte funcional (da una idea de la bondad del ajuste)
- $S_{\hat{u}}^2$: varianza residual, lo que queda sin explicar

A medida que el ajuste va mejorando, gana fuerza la parte funcional.

A partir de la expresión obtenida en (19) que se transcribe a continuación:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{u}_t^2$$

Se dividen ambos miembros de la igualdad por el lado izquierdo de la misma, con lo cual, se tiene que:

$$1 = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2} + \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

A partir de la igualdad anterior, pueden obtenerse las siguientes expresiones del **coeficiente de determinación R^2**.

$$R^2 = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

El coeficiente de determinación indica el porcentaje de las variaciones de la variable a explicar que es explicado en forma conjunta por las variaciones de las variables seleccionadas como explicativas. Este coeficiente adopta valores positivos entre 0 y 1; a partir de estos resultados y, siempre que se comparen modelos con igual cantidad de variables explicativas, se seleccionará aquel modelo cuyo coeficiente de determinación arroje un mayor valor.

Para poder entonces enunciar las propiedades estadísticas del modelo lineal general $Y_t = a_0 + a_1 X_{1t} + a_2 X_{2t} + \cdots + a_k X_{kt} + u_t$, resulta necesario definir supuestos acerca del término aleatorio.

Capítulo 3

Análisis estadístico

Supuestos de Gauss Markov

Se detallan a continuación los supuestos que se utilizan para resolver el modelo;

1. En promedio los errores se compensan:

$$E(u_t) = 0 \qquad t = 1,2,\cdots,T$$

$$E(u_1) = E(u_2) = \cdots = E(u_T) = 0$$

Que puede escribirse en notación matricial como:

$$E(u) = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

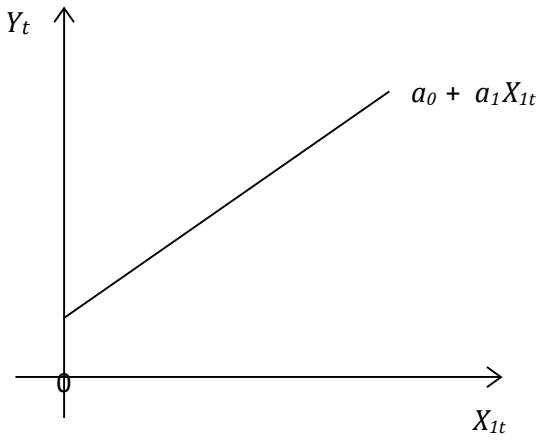
En el caso de una variable explicativa, lo expuesto se traduce en:

$$Y_t = a_0 + a_1 X_{1t} + u_t$$

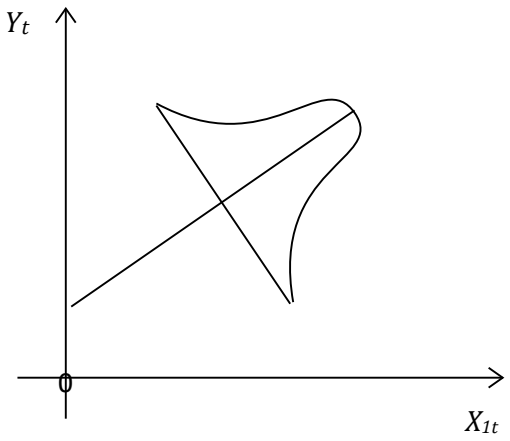
$$E(Y_t/X_{1t}) = E[a_0 + a_1 X_{1t} + u_t/X_{1t}]$$

$$E(Y_t/X_{1t}) = a_0 + a_1 X_{1t} + \underbrace{E(u_t/X_{1t})}_{=0}$$

Que puede graficarse del siguiente modo:



La recta de regresión da, entonces, el valor esperado (promedio). El residuo u_t va a oscilar en torno a ese promedio. Se grafica la distribución de u_t en torno al valor medio que está sobre la recta.



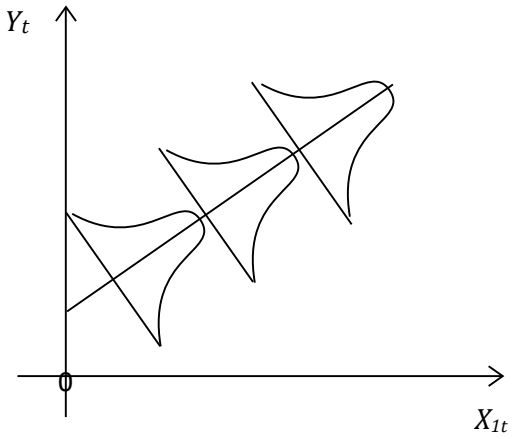
2. Debe definirse la matriz de varianzas y covarianzas de los residuos. A tal fin, el supuesto establece que:

$$E(u_i u_j) = \begin{cases} \sigma^2 \text{ (constante)} & i = j \\ 0 & i \neq j \end{cases}$$

$E(u_i u_j) = \sigma^2$ (constante) hace referencia a la propiedad de homocedasticidad mientras que $E(u_i u_j) = 0$ a la de independencia.

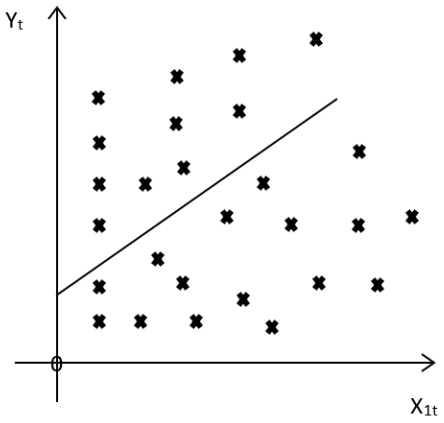
a) Independencia quiere decir que los errores para un periodo t dado son independientes de los errores para otro periodo $t + h$ o, que los errores de dos observaciones realizadas en un mismo momento son independientes entre sí (por ejemplo, el consumo de una familia no tiene nada que ver con el consumo correspondiente a otra familia). Lo importante es que no puedan predecirse los errores subsiguientes.

b) Homocedasticidad: los errores están distribuidos uniformemente, con igual varianza, en torno a la recta de represión. El siguiente gráfico representa lo expuesto para la variable consumo en función de la variable ingreso, $C = f(Y)$.

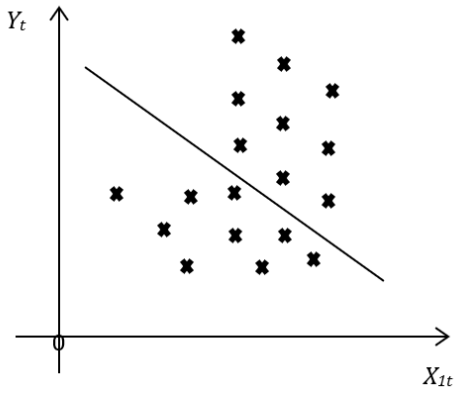


Sin embargo, en la realidad la variabilidad en torno al ingreso alto será mayor que en el caso de un ingreso bajo donde se verifica que la variabilidad será menor.

Lo expuesto previamente se muestra en el siguiente gráfico en el cual se observa que la varianza aumenta a medida que lo hace el ingreso. Este es un caso de *heterocedasticidad*.



Puede darse la heterocedasticidad al revés, la varianza σ^2 disminuye a medida que aumenta el número de horas de práctica, por ejemplo, en un modelo de aprendizaje. Lo expuesto se muestra en el siguiente gráfico:



En conclusión, el supuesto puede resumirse del siguiente modo:

- Si $i = j \quad E\left(u_1^2\right) = E\left(u_2^2\right) = \cdots = E\left(u_T^2\right) = \sigma^2$
- Si $i \neq j \quad E(u_1 u_T) = E(u_2 u_T) = \cdots = E(u_{T-1} u_T) = 0$

Lo previamente señalado puede ser expresado en forma matricial como:

$$\sum_{uu} = E\left(uu^t\right) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}\begin{bmatrix} u_1 & u_2 & \cdots & u_T \end{bmatrix}$$

$$\sum_{uu} = \begin{bmatrix} E\left(u_1^2\right) & E\left(u_1 u_2\right) & \ldots & E\left(u_1 u_T\right) \\ E\left(u_1 u_2\right) & E\left(u_2^2\right) & \cdots & E\left(u_2 u_T\right) \\ \vdots & \vdots & \vdots & \vdots \\ E\left(u_1 u_T\right) & E\left(u_2 u_T\right) & \ldots & E\left(u_T^2\right) \end{bmatrix}$$

Σ_{uu} se denomina matriz de varianzas y covarianzas de los residuos. Teniendo en cuenta este segundo supuesto, la mencionada matriz puede expresarse como:

$$E\left(uu^t\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I_T$$

[4]

En resumen, el segundo supuesto puede escribirse como:

$$E\left(uu^t\right) = \sigma^2 I_T$$

3. Este supuesto se refiere a las variables explicativas, las mencionadas variables son variables matemáticas independientes del término aleatorio, lo cual se escribe como:

$$E\,(X_{it}u_t) = X_{it}E\,(u_t) = 0$$

4. Las variables explicativas son independientes entre sí.

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1T} & X_{2T} & \dots & X_{kT} \end{bmatrix}$$

[4] NOTA: El término σ^2 siempre hace referencia a σ_u^2

Las columnas correspondientes a la matriz son independientes entre sí, es decir, $r(X) = k+1$ (el rango de la matriz es $(k+1)$). Este supuesto asegura una única solución del modelo.

La violación de este supuesto recibe el nombre de multicolinealidad. La implicancia que ejerce sobre el modelo es que no agrega nada la introducción de una nueva variable. No se puede saber el efecto individual de cada variable sobre el fenómeno.

Si hay multicolinealidad, entonces $r(X) < k + 1$. Si se supone que $r(X) = k$, entonces:

$$r(X) = r\left(X^t\right) \Rightarrow r\left(X^t X\right) \leq k$$

Ya que:

$$r(X^t X) \leq min\left[r(X), r\left(X^t\right)\right]$$

Pero $\left(X^t X\right)$ es de orden $(k + 1)$, con lo cual, como el rango es menor al orden de $\left(X^t X\right)$, esta matriz es singular, es decir, el determinante de $\left(X^t X\right) = 0.$

De este modo, no existiría inversa y no podrían calcularse los estimadores.

5. Supuesto de normalidad

$$u \sim N\left[E(u) = 0;\ E(uu^t) = \sigma_u^2\ (X^t X)^{-1}\right]$$

Debe tenerse en cuenta que Y_t es una variable aleatoria, con lo cual, los estimadores mínimo- cuadráticos $\hat{a}_i$ son variables aleatorias y, en consecuencia, tienen un valor esperado, una varianza y una distribución. A continuación, se procede entonces a realizar este análisis.

Propiedades de los estimadores mínimo cuadráticos

1. Bajo los supuestos de Gauss-Markov, los estimadores mínimo-cuadráticos son estimadores insesgados de los parámetros poblacionales. Es decir:

$$\begin{cases} E\left(\hat{a}_0\right) = a_0 \\ E\left(\hat{a}_1\right) = a_1 \\ \quad\quad\vdots \\ E\left(\hat{a}_k\right) = a_k \end{cases}$$

Que, en forma matricial, puede escribirse como:

$$E(\hat{a}) = a$$

Demostración:

Se tiene que: $$\hat{a} = \left(X^t X\right)^{-1} X^t Y.$$

Reemplazando $Y = Xa + u$ en la ecuación anterior obtiene que:

$$\hat{a} = \left(X^t X\right)^{-1} X^t (Xa + u)$$

Al distribuir, se llega a la siguiente expresión:

$$\hat{a} = \underbrace{\left(X^t X\right)^{-1} X^t X}_{I} a + \left(X^t X\right)^{-1} X^t u = a + \left(X^t X\right)^{-1} X^t u \qquad (20)$$

Aplicando ahora el operador esperanza a (20), se tiene que:

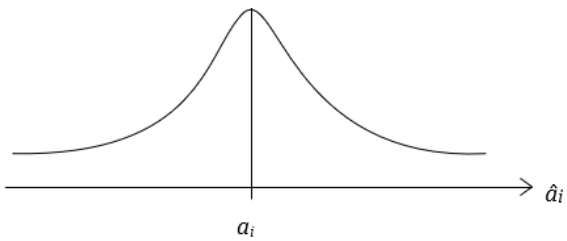
$$E(\hat{a}) = E\left[a + \left(X^t X\right)^{-1} X^t u\right]$$

$$E(\hat{a}) = E(a) + \left(X^t X\right)^{-1} X^t \underbrace{E(u)}_{0}$$

Es decir, los estimadores minimo-cuadráticos $\hat{a}$ son estimadores insesgados de los parámetros a:

$$E(\hat{a}) = a$$

Se puede graficar lo expuesto del siguiente modo:



2. Varianza de los estimadores mínimo-cuadráticos $\hat{a}$

Cálculo de la matriz de varianzas y covarianzas de los estimadores.

Se sabe que:

$$var(\hat{a}_i) = E\,[\hat{a}_i - a_i]^2$$

$$\text{cov}(\hat{a}_i\hat{a}_j) = E\,[(\hat{a}_i - a_i)(\hat{a}_j - a_j)] = \sigma\,(\hat{a}_i\hat{a}_j)$$

Por otro lado, se tiene que $\sum_{uu} = E\left(uu^t\right)$

La matriz de varianzas y covarianzas de los estimadores se calcula del siguiente modo:

$$\sum_{\hat{a}\hat{a}} = E\left[(\hat{a}-a)(\hat{a}-a)^t\right] = E\begin{bmatrix}\hat{a}_0 - a_0\\ \hat{a}_1 - a_1\\ \vdots\\ \hat{a}_k - a_k\end{bmatrix}\left[(\hat{a}_0 - a_0)\,(\hat{a}_1 - a_1)\ldots(\hat{a}_k - a_k)\right] \quad (21)$$

Realizando el producto entre matrices, la matriz resultante es:

$$\sum_{\hat{a}\hat{a}} = \begin{bmatrix} E\,(\hat{a}_0 - a_0)^2 & E\,(\hat{a}_0 - a_0)\,(\hat{a}_1 - a_1) & \cdots & E\,(\hat{a}_0 - a_0)\,(\hat{a}_k - a_k)\\ E\,(\hat{a}_0 - a_0)\,(\hat{a}_1 - a_1) & E\,(\hat{a}_1 - a_1)^2 & \cdots & E\,(\hat{a}_1 - a_1)\,(\hat{a}_k - a_k)\\ \vdots & \vdots & \vdots & \vdots\\ E\,(\hat{a}_0 - a_0)\,(\hat{a}_k - a_k) & E\,(\hat{a}_1 - a_1)\,(\hat{a}_k - a_k) & \cdots & E\,(\hat{a}_k - a_k)^2 \end{bmatrix}$$

Que es equivalente a

$$\sum_{\hat{a}\hat{a}} = \begin{bmatrix} \text{var}\,(\hat{a}_0) & \text{cov}\,(\hat{a}_0\hat{a}_1) & \ldots & \text{cov}\,(\hat{a}_0\hat{a}_k)\\ \text{cov}\,(\hat{a}_1\hat{a}_0) & \text{var}\,(\hat{a}_1) & \ldots & \text{cov}\,(\hat{a}_1\hat{a}_k)\\ \vdots & \vdots & \vdots & \vdots\\ \text{cov}\,(\hat{a}_k\hat{a}_0) & \text{cov}\,(\hat{a}_k\hat{a}_1) & \ldots & \text{var}\,(\hat{a}_k) \end{bmatrix}$$

Por otro lado, se sabe que:

$$\hat{a} = \left(X^tX\right)^{-1}X^tY$$

Donde:

$$X = Xa + u$$

De acuerdo con lo obtenido en (20):

Con lo cual,

$$\hat{a} = a + \left(X^t X\right)^{-1} X^t u$$

$$\hat{a} - a = \left(X^t X\right)^{-1} X^t u$$

Asimismo, teniendo en cuenta los supuestos de Gauss-Markov,

$$E(u) = 0$$

$$\sum_{uu} = E(uu^t) = \sigma^2 I_T$$

Y recordando, además, algunas propiedades de las matrices:

- $(AB)^t = B^t A^t$
- $\left(A^{-1}\right)^t = A^{-1}$
- $A^t = A \rightarrow$ simétrica
- Si A es simétrica, entonces A^{-1} también lo es [5]

Aplicando lo expuesto previamente en (21), se tiene que:

$$\sum_{\hat{a}\hat{a}} = E\left[(\hat{a} - a)(\hat{a} - a)^t\right] = E\left[\left(X^t X\right)^{-1} X^t u \left(\left(X^t X\right)^{-1} X^t u\right)^t\right]$$

$$\sum_{\hat{a}\hat{a}} = E\left[\left(X^t X\right)^{-1} X^t u u^t X \left(X^t X\right)^{-1}\right] = \left(X^t X\right)^{-1} X^t E\left(uu^t\right) X \left(X^t X\right)^{-1}$$

$$\sum_{\hat{a}\hat{a}} = \left(X^t X\right)^{-1} X^t \sigma_u^2 X (X^t X)^{-1}$$

σ_u^2 es un escalar y puede salir fuera de la matriz. De este modo, se tiene que:

[5] NOTA: Esta última propiedad se aplica a la matriz simétrica, $(X^t X)^{-1}$, es decir: $[(X^t X)^{-1}]^t = (X^t X)^{-1}$

$$\sum_{\hat{a}\hat{a}} = \sigma_u^2 \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1}$$

Que se reduce a la siguiente expresión

$$\sum_{\hat{a}\hat{a}} = \sigma_u^2 \left(X^t X\right)^{-1}$$

La matriz resultante puede escribirse del siguiente modo:

$$\sum_{\hat{a}\hat{a}} = \sigma_u^2 \begin{bmatrix} c^{00} & c^{01} & \cdots & c^{0k} \\ c^{10} & c^{11} & \cdots & c^{1k} \\ \vdots & \vdots & \vdots & \vdots \\ c^{0k} & c^{1k} & \cdots & c^{kk} \end{bmatrix} = \left\{\sigma_u^2 c^{ii}\right\}$$

Donde c^{ij} es el *ij – ésimo* elemento de la matriz $(X^t X)^{-1}$.

En resumen,

- $var(\hat{a}_i) = E(\hat{a}_i - a_i)^2 = \sigma_u^2 c^{ii}$ donde c^{ii} es el elemento i *– ésimo* de la diagonal principal.
- $cov\ (\hat{a}_i \hat{a}_j) = E\ [(\hat{a}_i - a_i)(\hat{a}_j - a_j)] = \sigma_u^2 c^{ij}$

En la matriz $(X^t X)^{-1}$ se encuentran en la diagonal principal, los elementos que acompañan a las varianzas de los estimadores y, fuera de ella, los correspondientes a las covarianzas. El problema es que la varianza σ_u^2 es desconocida.

Construcción de un estimador insesgado de σ_u^2

Se sabe que:

$$\hat{u} = Y - \hat{Y} = Y - X\hat{a} = Y - X\left(X^tX\right)^{-1}X^tY = \left[I - X\left(X^tX\right)^{-1}X^t\right]Y$$

Se define la matriz M como;

$$M = I - X\left(X^tX\right)^{-1}X^t$$

De este modo:

$$\hat{u} = MY = M(Xa + u) = MXa + Mu \qquad (22)$$

M es una matriz simétrica, es decir, su traspuesta es M, $M^t = M$.

Puede calcularse el cuadrado de la matriz como:

$$M^2 = \left[I - X\left(X^tX\right)^{-1}X^t\right]\left[I - X\left(X^tX\right)^{-1}X^t\right]$$

Operando la expresión:

$$M^2 = I + X\left(X^tX\right)^{-1}X^tX\left(X^tX\right)^{-1}X^t - X\left(X^tX\right)^{-1}X^t - X\left(X^tX\right)^{-1}X^t$$

$$M^2 = I - X\left(X^tX\right)^{-1}X^t = M$$

Se observa que el cuadrado de la matriz es equivalente a la matriz original. Este tipo de matrices, en las que se verifican las dos propiedades expuestas, recibe el nombre de matrices *idempotentes*.

Otra propiedad que las caracteriza es:

$$MX = \left[I - X\left(X^tX\right)^{-1}X^t\right]X = X - X\left(X'X\right)^{-1}X^tX = X - X = 0$$

De este modo, conforme lo previamente expuesto, surge que:

$$MXa = 0$$

Con lo cual, la expresión (22) se reduce a:

$$\hat{u} = MXa + Mu = Mu$$

Es decir, la expresión final está dada por:

$$\hat{u} = Mu$$

De este modo,

$$\sum_{t=1}^{T} \hat{u}^2 = \hat{u}^t \hat{u} = (Mu)^t Mu = u^t M^t Mu = u^t MMu$$

$$\sum_{t=1}^{T} \hat{u}^2 = u^t M^2 u = u^t Mu$$

$$\sum_{t=1}^{T} \hat{u}^2 = u^t Mu = tra\left(u^t Mu\right)$$

Al ser la traza conmutativa puede escribirse:

$$\sum_{t=1}^{T} \hat{u}^2 = tra\left(uu^t M\right)$$

Se debe recordar la definición de traza y sus propiedades.

Sea la matriz A:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

La traza resulta equivalente a:

- $tra(A) = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^{n} a_{ii}$

Se verifican además las siguientes propiedades:

- Tra(kA)= k*tra(A)
- tra (A B) =tra (B A)
- tra (A+B) =tra (A)+tra (B)

Se calcula entonces la esperanza matemática de $\sum_{t=1}^{T} \hat{u}_t^2$ teniendo en cuenta las propiedades mencionadas:

$$E\left(\sum_{t=1}^{T} \hat{u}_t^2\right) = E\left[tr\left(uu^t M\right)\right] = tr\left[E(uu^t M)\right] = tr\left[ME(uu^t)\right] = tr\left(\sigma_u^2 M\right)$$

Con lo cual:

$$E\left[\sum_{t=1}^{T}\hat{u}_t^2\right] = \sigma_u^2 tr(M)$$ ya que σ_u^2 es constante

Donde:

$$tr(M) = tr\left[I - X\left(X^tX\right)^{-1}X^t\right] = tr\left[I_T\right] - tr\left[X\left(X^tX^{-1}\right)X^t\right]$$

Al ser la traza conmutativa, se tiene que:

$$tr(M) = tr\left(I_T\right) - tr\left[X^tX\left(X^tX\right)^{-1}\right] = tr\left(I_T\right) - tr\left(I_{k+1}\right) = T - (k+1)$$

La expresión final resultante es, entonces;

$$E\left[\sum_{t=1}^{T}\hat{u}_t^2\right] = \sigma_u^2[T - (k+1)]$$

De allí que:

$$E\left[\frac{\sum_{t=1}^{T}\hat{u}^2t}{T-(k+1)}\right] = \sigma_u^2$$

Se concluye entonces que para construir un estimador insesgado de σ_u^2 se debe dividir la suma de los desvíos cuadráticos por T-(k+1) - los grados de libertad-.

Entonces, la expresión:

$$\hat{\sigma}_u^2 = \frac{\sum_{t=1}^{T}\hat{u}_t^2}{T-(k+1)}$$

Es el estimador insesgado de σ_u^2.

De este modo, la matriz de varianzas y covarianzas estimada de los estimadores mínimos cuadráticos es:

$$\sum_{\hat{a}\hat{a}}^{\wedge} = \hat{\sigma}_u^2\left(X^tX\right)^{-1}$$

¿Qué sucede con los estimadores $\hat{a}_i$ cuando se agrega el supuesto de Gauss Markov referente a normalidad?

Distribución de los estimadores mínimos cuadráticos

Bajo los supuestos de Gauss Markov:

$$u \sim N\left[0; \sigma_u^2 I_t\right]$$

Se tiene que la expresión correspondiente a los estimadores mínimo-cuadráticos está dada por:

$$\hat{a} = \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_k \end{bmatrix} = a + \left(X^t X\right)^{-1} X^t u$$

Puede escribirse entonces que:

$$\hat{a}_i = a_i + \underbrace{m_{i1}u_1 + m_{i2}u_2 + \cdots + m_{iT}u_T}_{\text{combinación lineal de } u_j}$$

Se sabe que toda combinación de variables aleatorias normales independientes da como resultado una variable con distribución normal. Entonces:

$$\hat{a}_i \sim N\left[a_i; \sigma_u^2 \left(X^t X\right)_{ii}^{-1}\right]$$

$$\hat{a_i} \sim N\left[a_i; \sigma_u^2 c^{ii}\right]$$

Como σ_u^2 es desconocido no puede ser utilizada la distribución normal. Es necesario entonces recurrir a la distribución t de Student a cuyo efecto, con carácter previo debe definirse la correspondiente distribución χ^2.

Debe recordarse que una distribución chi cuadrado se define como una suma de variables aleatorias normales, estandarizadas, independientes y elevadas al cuadrado.

Se tiene entonces:

$$\chi^2_{(n)} = \sum_{i=1}^{n} X_i^2$$ donde $X_i \sim N(0,1)$ y son independientes entre sí

Esta distribución toma únicamente valores positivos. El valor esperado de esta variable coincide con sus grados de libertad y su varianza es equivalente al duplo de sus grados de libertad.

Además, la suma de dos chi-cuadrados independientes entre sí da por resultado otra chi-cuadrado cuyos grados de libertad resultan de la suma de los grados de libertad de las chi-cuadrados sumandos.

Es decir,

$$E\left[\chi^2(n)\right] = n \qquad \sigma^2\left[\chi^2(n)\right] = 2n$$

$\left.\begin{array}{l} u \sim \chi^2_n \\ v \sim \chi^2_m \end{array}\right\}$ si son independientes, entonces, $\chi^2_u + \chi^2_v = \chi^2_{n+m}$

Generalización de Cochrane al Lema de Fisher

Conforme la relación obtenida al desarrollar la descomposición de las variaciones en el capítulo anterior, se tiene que:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{u}_t^2$$

Dividiendo ambos miembros por σ_u^2, se obtiene la siguiente expresión:

$$\underbrace{\frac{\sum_{i=1}^{T} y_t^2}{\sigma_u^2}}_{\chi^2_{T-1}} = \underbrace{\frac{\sum_{i=1}^{T} \hat{y}_t^2}{\sigma_u^2}}_{\chi^2_k} + \underbrace{\frac{\sum_{i=1}^{T} \hat{u}_t^2}{\sigma_u^2}}_{\chi^2_{T-(K+1)}}$$

Se observa que las distribuciones de todos los miembros de la igualdad responden a la forma de una distribución chi-cuadrado dado que se trata de una suma de variables con distribución normal, independientes, estandarizadas y elevadas al cuadrado. Debe analizarse entonces el porqué de los grados de libertad.

- La chi-cuadrado del lado izquierdo de la igualdad tiene $T - 1$ grados de libertad: a la totalidad de sumandos T se le resta 1 porque se ha tenido que calcular la media de las observaciones en función de los T datos.
- La correspondiente al primer sumando del segundo miembro tiene k grados de libertad: ello se debe a que las variables independientes son los k estimadores de los parámetros a_i con (i = 1,2, …, k) dado que $\hat{a}_0$ puede obtenerse como combinación lineal de los estimadores previamente mencionados.

- La correspondiente al segundo sumando tiene $T - k - 1$ grados de libertad: ello se debe a que, si bien se cuenta con T estimaciones de Y, para poder realizarlas hubo que calcular con carácter previo a los $k + 1$ estimadores de a_i.

  Además, debe destacarse que los dos sumandos del segundo miembro son independientes entre sí (*recordar la propiedad mencionada de la chi cuadrado al respecto*).

Teniendo ahora en cuenta la propiedad de valor esperado de una variable con distribución chi-cuadrado, se concluye:

$$E\left[\frac{\sum_{t=1}^{T}\hat{u}_t^2}{\sigma_u^2}\right] = T - k - 1$$

De donde:

$$E\left[\frac{\sum_{t=1}^{T}\hat{u}_t^2}{T - k - 1}\right] = \sigma_u^2$$

Al trabajar con la traza se llega a este mismo resultado sin suponer que u_t se distribuye normalmente; con lo cual, la expresión obtenida a partir de la deducción con la traza es válida para cualquier distribución.

El desarrollo efectuado en este punto es sólo válido si la distribución es normal.

Se tenía entonces que la distribución de los estimadores $\hat{a}_i$ estaba dada por:

$$\frac{\hat{a}_i - a_i}{\sigma_u\sqrt{c^{ii}}} \sim N(0,1)$$

donde σ_u es desconocida

Se define ahora la variable t *de student* como el cociente entre una normal estandarizada y la raíz cuadrada de una chi-cuadrado dividida por sus respectivos grados de libertad.[6]

[6] NOTA: la chi cuadrado es la que surge del segundo sumando de la expresión obrante en la página 37.

De este modo, la expresión que se obtiene al reemplazar la normal estandarizada y la chi cuadrado, es la siguiente:

$$t_{(T-k-1)} = \frac{\frac{\hat{a}_i - a_i}{\sigma_u \sqrt{c^{ii}}}}{\sqrt{\frac{1}{T-k-1}\frac{\sum_{t=1}^{T}\hat{u}^2}{{\sigma_u}^2}}} = \frac{\hat{a}_i - a_i}{\hat{\sigma}_u \sqrt{c^{ii}}} = \frac{\hat{a}_i - a_i}{\hat{\sigma}(\hat{a}_i)}$$

Donde:

$$\hat{\sigma_u}^2 = \frac{\sum \hat{u}_t^2}{T-k-1}$$

Distribución F de Snedecor

Debe recordarse la definición de la distribución F de Snedecor, como cociente entre dos chi cuadrado independientes entre sí, divididas por sus respectivos grados de libertad:

$$F_n^m = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

Entonces, en el caso analizado, en el numerador se tomará la chi cuadrado con k grados de libertad:

$$\chi_k^2 = \frac{\sum_{t=1}^{T}\hat{y}_t^2}{\sigma_u^2}$$

Mientras que en el denominador se tomará la chi cuadrado con $T-k-1$ grados de libertad:

$$\chi_{T-k-1}^2 = \frac{\sum_{t=1}^{T}\hat{u}_t^2}{\sigma_u^2}$$

De este modo, la distribución F resultante está dada por:

$$F_{T-k-1}^k = \frac{\frac{\sum_{t=1}^{T}\hat{y}_t^2}{\sigma_u^2} \times \frac{1}{k}}{\frac{\sum_{t=1}^{T}\hat{u}_t^2}{\sigma_u^2} \times \frac{1}{T-k-1}} = \frac{\frac{\sum_{t=1}^{T}\hat{y}_t^2}{k}}{\frac{\sum_{t=1}^{T}\hat{u}_t^2}{T-k-1}}$$

Que puede reordenarse del siguiente modo;

$$F_{T-k-1}^{k} = \frac{T-k-1}{k} \times \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} \hat{u}_t^2} = \frac{T-k-1}{k} \times \frac{R^2}{1-R^2}$$

donde R^2:

$$R^2 = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2} = 1 - \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

Aplicaciones a la inferencia estadística

Test de hipótesis para parámetros individuales

Este tipo de prueba sirve para establecer si una variable determinada del modelo es significativa para explicarlo. A tal fin, se plantea la siguiente hipótesis nula:

$$H_0 : a_i = 0$$

donde a_i es el parámetro que acompaña a la variable X_{it} de la cual se quiere probar su significatividad. El planteo de esta hipótesis señala que es indiferente la inclusión de X_{it} en el modelo ya que si se observa la forma de éste:

$$Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + \ldots + a_iX_{it} + \ldots + a_kX_{kt} + u_t$$

Al plantear la mencionada hipótesis, ese sumando puede no incluirse dado que $a_iX_{it} = 0$.

O sea, plantear $H_{0:}\ a_i = 0$ significa decir que la variable X_{it} que acompaña al parámetro a_i no es significativa.

Se sabe que toda hipótesis nula debe ser contrastada a una hipótesis alternativa.

De este modo, las tres hipótesis alternativas posibles son las tradicionales: mayor, menor o distinto. El interrogante que debe hacerse quien plantea la hipótesis es en qué casos se plantea cada una de ellas.

En general, tratándose de problemas económicos, las mismas serán planteadas en función de lo que se espera de la relación entre las variables analizadas.

Por ejemplo, si se explica el consumo en función del ingreso se espera una relación directa de donde, si se desea probar la significatividad del ingreso para explicar al consumo, se ha de plantear una hipótesis alternativa de mayor.

En los casos en que se desconozca el tipo de relación que vincula a las variables, se utilizaría una hipótesis alternativa de distinto.

Para poder llevar a cabo el test de hipótesis se necesita conocer la distribución del estimador correspondiente. De acuerdo con lo desarrollado previamente, se tiene que la distribución del estimador es t_{T-k-1}.

De este modo, el estadístico adopta la siguiente forma:

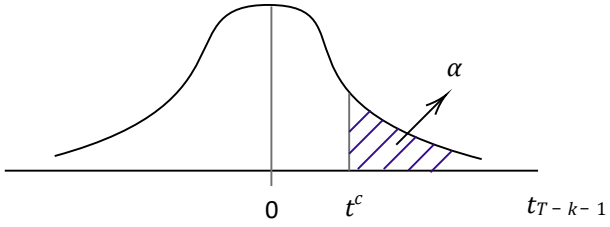
$$\frac{\hat{a}_i - a_i}{\hat{\sigma}\,(\hat{a}_i)} \sim t_{T-k-1}$$

Entonces, es posible plantear los siguientes test de hipótesis:

1.

$H_0 : a_i = 0$

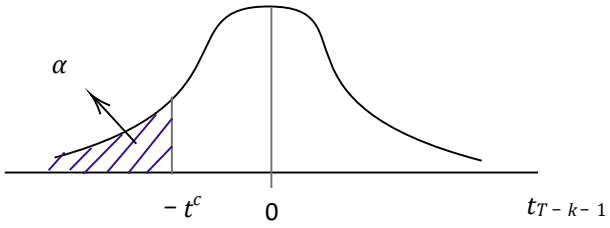
$H_1 : a_i > 0$



2.

$H_0 : a_i = 0$

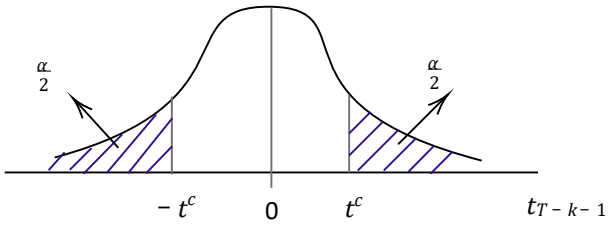
$H_1 : a_i < 0$



3.

$H_0 : a_i = 0$

$H_1 : a_i \neq 0$



α es el nivel de riesgo con el cual se ha de trabajar. El valor asignado al mismo permite establecer el/los puntos/s crítico/s.

t^c representa el punto crítico que, dependiendo de su ubicación, puede ser positivo o negativo.

Una vez establecido el punto crítico, se está en condiciones de definir la regla de decisión.

A título de ejemplo, si se ha seleccionado el caso 1 correspondiente a una hipótesis alternativa de mayor, la regla de decisión será:

Si el valor del estadístico de prueba es superior al punto crítico t^c, se rechaza la hipótesis nula, caso contrario, no existen pruebas para rechazarla.

Quedaría entonces establecer la decisión a adoptar. A tal fin, es necesario haber seleccionado una muestra y realizado una regresión a partir de la cual se haya obtenido el valor de $\hat{a}_i$.

De este modo, el estadístico de prueba t^* resulta equivalente a:

$$t^* = \frac{\hat{a}_i - a_i}{\hat{\sigma}\left(\hat{a}_i\right)}$$

que, en el caso de la hipótesis planteada, se reduce a:

$$t^* = \frac{\hat{a}_i}{\hat{\sigma}\left(\hat{a}_i\right)}$$

dado que se ha planteado

$$H_0 : a_i = 0$$

Entonces, conforme lo mencionado con carácter previo, para una hipótesis alternativa de mayor:

Si $t^* > t^c$ se rechaza H_0

Conclusión: con un nivel de riesgo de α se rechaza la H_0 que plantea que la variable X_{it} no es significativa para explicar al modelo. Es decir, se está diciendo que la variable sirve para explicar Y_t.

Si $t^* < t^c$ no existen pruebas para rechazar H_0

Conclusión: con un nivel de confianza de $(1 - \alpha)$ no existen pruebas para rechazar la H_0 que plantea que la variable X_{it} no es significativa para explicar el modelo. Es decir, esta variable no sirve.

Construcción de intervalos de confianza para los parámetros a_i

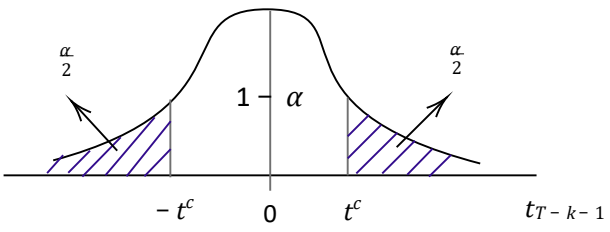
A fin de poder construir un intervalo de confianza para un parámetro determinado es necesario establecer cuál es el estimador y qué distribución tiene el mismo.

El parámetro a estimar es, en este caso, a_i y su estimador $\hat{a}_i$. La distribución de $\hat{a}_i$ es:

$$\frac{\hat{a}_i - a_i}{\hat{\sigma}\,(\hat{a}_i)} \sim t_{+-k-1}$$

que recibe el nombre de función pivotal.

Se trabaja con un nivel de confianza de $(1 - \alpha)$; como se muestra en el siguiente gráfico:



De la observación del gráfico surge que:

$$P\,[-t^c < t_{T-k-1} < t^c] = 1 - \alpha$$

Reemplazando t_{T-k-1} por la función pivotal, se tiene que:

$$P\left[-t^c < \frac{\hat{a}_i - a_i}{\hat{\sigma}\,(\hat{a}_i)} < t^c\right] = 1 - \alpha$$

Dado que el objetivo es construir un intervalo de confianza para el parámetro a_i, se debe despejar el mismo.

Multiplicando ambos miembros de la desigualdad por $\hat{\sigma}(\hat{a}_i)$, se obtiene:

$$P\,[-t^c \cdot \hat{\sigma}(\hat{a}_i) < \hat{a}_i - a_i < t^c \cdot \hat{\sigma}(\hat{a}_i)] = 1 - \alpha$$

Restando $\hat{a}_i$ en todos los miembros:

$$P\,[-t^c \cdot \hat{\sigma}(\hat{a}_i) - \hat{a}_i < -a_i < t^c \cdot \hat{\sigma}(\hat{a}_i) - \hat{a}_i] = 1 - \alpha$$

Multiplicando por (−1) los miembros de la desigualdad, se obtiene:

$$P\,[\hat{a}_i - t^c \cdot \hat{\sigma}(\hat{a}_i) < a_i < \hat{a}_i + t^c \cdot \hat{\sigma}(\hat{a}_i)] = 1 - \alpha$$

En general, el intervalo resultante se escribe como:

$$[\hat{a}_i \mp t^c \cdot \hat{\sigma}\,(\hat{a}_i)]$$

donde t^c es un valor de abscisa correspondiente a la distribución t_{T-k-1} que se obtiene a partir del nivel de riesgo establecido en la determinación del intervalo de confianza.

Compatibilidad de conclusiones

El intervalo obtenido puede contener o no al valor cero. Si lo contiene, la conclusión es que la variable X_{it} que acompaña al parámetro a_i no es significativa, es decir, es la misma conclusión a la que se llega cuando no se rechaza la hipótesis nula planteada en el test para parámetros individuales.

Si el intervalo no contiene el valor cero, entonces la variable X_{it} que acompaña al parámetro a_i es significativa para explicar el modelo. Es decir, es la misma conclusión a la que se llega cuando se rechaza la hipótesis nula planteada en el test para parámetros individuales.

Test de hipótesis para variables conjuntas

Se plantea la siguiente hipótesis nula:

$H_0 : a_1 = a_2 = \cdots = a_k = 0$

Este planteo significa que las variables explicativas no son significativas en forma conjunta para explicar el modelo.

La hipótesis alternativa es:

H_1 : por lo menos una de las variables es significativa

Como en toda prueba de hipótesis, debe definirse el estadístico a utilizar y su distribución.

De acuerdo con la hipótesis planteada, surge que el estadístico a utilizar debe comparar lo explicado con lo no explicado en el modelo por las variables seleccionadas.

Ese estadístico es el que ya ha sido desarrollado en la sección anterior - pag.40 - y está dado por:

$$\frac{\frac{\sum_{t=1}^{T} \hat{y}_t^2}{k}}{\frac{\sum_{t=1}^{T} \hat{u}_t^2}{T-k-1}} \sim F_{(k,T-k-1)}$$

Que puede expresarse también como:

$$\frac{R^2}{1-R^2} \cdot \frac{T-k-1}{k} \sim F_{(k,T-k-1)}$$

Entonces, la distribución a utilizar es $F_{(k,T-k-1)}$.

Debe determinarse ahora la zona de rechazo en función de las hipótesis planteadas.

En este tipo de test, la zona de rechazo se establece siempre del lado derecho. ¿Por qué?

Si se considera la situación límite de $R^2 = 1$, ese valor de R^2 indica que las variables seleccionadas explican en forma conjunta el 100% de las variaciones del modelo.

Recordando que $H_0 : a_1 = a_2 = \cdots = a_k = 0$, significa que las variables seleccionadas no son significativas en forma conjunta para explicar al modelo; se tiene que debería rechazarse entonces la hipótesis nula cuando $R^2 = 1$.

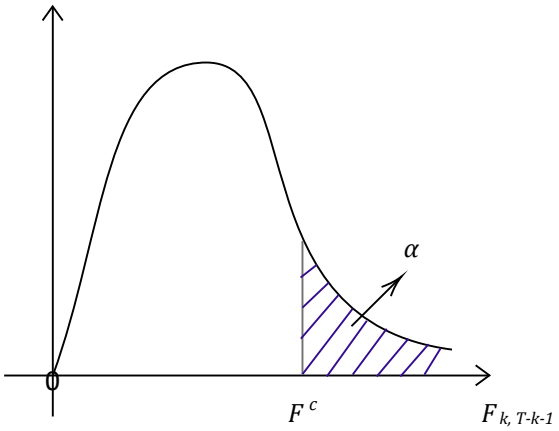
Además, si $R^2 = 1$ el valor del estadístico propuesto tiende a ∞.

De lo expuesto se infiere que el rechazo de la hipótesis nula tendrá lugar para valores altos de la variable F; con lo cual, la zona de rechazo debe ubicarse del lado derecho.

En resumen,

$H_0 : a_1 = a_2 = \cdots = a_k = 0$

H_1 : por lo menos una es significativa



donde F_c es un valor de abscisa correspondiente a la distribución $F_{(k,\, T-k-1)}$ que se obtiene a partir del nivel de riesgo establecido.

El estadístico de prueba es:

$$F^* = \frac{R^2}{1 - R^2} \cdot \frac{T - k - 1}{k}$$

Entonces, la regla de decisión es:

Si $F^* > F_c$ se rechaza H_0

Con un nivel de riesgo de α se rechaza la hipótesis nula que plantea que las variables explicativas no son significativas en forma conjunta para explicar al modelo. Es decir, al menos una de las variables es significativa.

Si $F^* < F_c$ no se rechaza H_0

Con un nivel de confianza de (1 – α) no existen pruebas para rechazar la hipótesis nula que plantea que las variables explicativas no son significativas en forma conjunta para explicar al modelo.

Compatibilidad con los test de hipótesis para parámetros individuales

Si se rechaza este test de hipótesis de variables conjuntas quiere decir que al menos una de las variables es significativa para explicar al modelo. Es decir, se debe al menos haber rechazado uno de los test individuales.

En el caso de no haber rechazado el test de variables conjuntas, no se tiene que haber rechazado ninguno de los test individuales.

Coeficiente de determinación corregido o ajustado

Este coeficiente se obtiene a partir del ajuste del coeficiente de determinación por sus grados de libertad como se muestra a continuación:

$$R_c^2 = 1 - \frac{\frac{\sum_{t=1}^{T} \hat{u}_t^2}{T-k-1}}{\frac{\sum_{t=1}^{T} y_t^2}{T-1}} \tag{23}$$

La expresión anterior puede ser escrita de estas dos formas:

$$R_c^2 = 1 - \frac{T-1}{T-k-1} \cdot \left(1 - R^2\right)$$

$$R_c^2 = R^2 - \frac{k}{T-k-1} \cdot \left(1 - R^2\right) \tag{24}$$

A partir del análisis de la segunda expresión surge que $R_c^2 < R^2$, es decir, el valor del coeficiente de determinación corregido es siempre inferior al del coeficiente de determinación que le da origen.

¿Por qué se calcula el coeficiente de determinación corregido? El valor del coeficiente de determinación aumenta a medida que se agregan variables explicativas al modelo, aunque éstas no sean significativas; en cambio el valor del coeficiente de determinación corregido aumenta sólo cuando la variable agregada es significativa.

Diferencia entre dos errores de regresión

Sean:

$\sum_{i=1}^{T} \hat{u}_t^2\ (Y, X_1)$, los residuos obtenidos a partir de la explicación de Y por X_1

$\sum_{i=1}^{T} \hat{u}_t^2\ (Y, X_1 X_2)$, los residuos obtenidos a partir de la explicación de Y por X_1 y X_2 en forma conjunta.

Se define el coeficiente de correlación parcial $R_{Y\ X_2 \cdot X_1}$ como el porcentaje de la varianza del regresando no explicada por X_1 que será explicada con la introducción de X_2. Es decir, mide en qué porcentaje mejora el modelo el hecho de agregar X_2 como variable explicativa.

Su fórmula está dada por la siguiente expresión:

$$R_{Y\ X_2 \cdot X_1} = 1 - \frac{\sum_{i=1}^{T} \hat{u}_t^2 (Y, X_1 X_2)}{\sum_{i=1}^{T} \hat{u}_t^2 (Y, X_1)}$$

La comparación entre los coeficientes de correlación parcial permite decidir qué variable agregar al modelo ya construido. Se elige agregar aquella variable a la cual le corresponda el coeficiente de correlación parcial más alto.

En general, para el caso en que se introduzca una segunda variable X_i permaneciendo constante X_j se aplica la siguiente fórmula:

$$R_{Y\ X_i \cdot X_j} = 1 - \frac{\sum_{i=1}^{T} \hat{u}_t^2 (Y, X_i X_j)}{\sum_{i=1}^{T} \hat{u}_t^2 (Y, X_j)}$$

Si hubiere dos variables explicativas X_i y X_j y se desea agregar una tercera X_k, se tiene que la fórmula a aplicar es:

$$R_{Y\,X_k \cdot X_i \cdot X_j} = 1 - \frac{\sum_{i=1}^{T} \hat{u}_t^2(Y, X_i X_j X_k)}{\sum_{i=1}^{T} \hat{u}_t^2(Y,\ X_i X_j)}$$

Puede generalizarse la fórmula para cualquier número de variables explicativas. Los coeficientes de correlación parcial sirven para decidir qué variable agregar, pero no para determinar si ésta es significativa. Para ello, habrá que realizar los test de hipótesis para parámetros individuales.

Capítulo 4

Un caso particular

Sin perjuicio que las herramientas desarrolladas en los capítulos anteriores sean aplicables a modelos con una sola variable explicativa, se desarrollarán en el presente capítulo aquellas que sólo se aplican a este caso particular.

Determinación de los estimadores

A partir del sistema de ecuaciones:

$$\begin{cases} \sum_{t=1}^{T} Y_t = T\hat{a}_0 + \hat{a}_1 \sum_{t=1}^{T} X_{1t} \\ \sum_{t=1}^{T} Y_t X_{1t} = \hat{a}_0 \sum_{t=1}^{T} X_{1t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}^2 \end{cases}$$

Puede obtenerse el valor de $\hat{a}_1$ mediante la aplicación del método de determinantes del siguiente modo:

$$\hat{a}_1 = \frac{\Delta \hat{a}_1}{\Delta} \tag{25}$$

Donde:

- $\Delta_{\hat{a}_1}$ es el determinante correspondiente a la incógnita que se desea calcular; en este caso, $\hat{a}_1$.
- Δ es el determinante del sistema.

El determinante del sistema se obtiene tomando los coeficientes que acompañan a las incógnitas; es decir,

$$\Delta = \begin{bmatrix} T & \sum_{t=1}^{T} X_{1t} \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 \end{bmatrix}$$

Se resuelve el mismo y se obtiene:

$$\Delta = T \sum_{t=1}^{T} X_{1t}^2 - \left(\sum_{t=1}^{T} X_{1t} \right)^2$$

Dividiendo ambos miembros por T^2 y reordenando, se llega a:

$$\frac{\Delta}{T^2} = \sum_{t=1}^{T} \frac{X_{1t}^2}{T} - \left(\sum_{t=1}^{T} \frac{X_{1t}}{T} \right)^2$$

O sea, las expresiones expuestas responden a las fórmulas de momentos absolutos como se muestra a continuación:

$$\frac{\Delta}{T^2} = m_2 - m_1^2$$

Es decir, es la fórmula correspondiente a la varianza.

$$\frac{\Delta}{T^2} = S_{x_1}^2$$

De donde, el determinante resulta equivalente a:

$$\Delta = T^2 \cdot S_{x_1}^2$$

El determinante de la incógnita $\Delta_{\hat{a}_1}$ surge de reemplazar en el determinante del sistema la columna de los términos correspondientes a la incógnita por la de los términos independientes. Es decir,

$$\Delta_{\hat{a}_1} = \begin{bmatrix} T & \sum_{t=1}^{T} Y_t \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} Y_t X_{1t} \end{bmatrix}$$

Al resolver el determinante se obtiene la siguiente expresión:

$$\Delta_{\hat{a}_1} = T \cdot \sum_{t=1}^{T} Y_t X_{1t} - \sum_{t=1}^{T} X_{1t} \cdot \sum_{t=1}^{T} Y_t$$

Dividiendo ambos miembros por T^2, se tiene que:

$$\frac{\Delta_{\hat{a}_1}}{T^2} = \frac{\sum_{t=1}^{T} Y_t X_{1t}}{T} - \frac{\sum_{t=1}^{T} X_{1t}}{T} \cdot \frac{\sum_{t=1}^{T} Y_t}{T}$$

O sea, aparecen ahora el momento absoluto mixto y el correspondiente a cada una de las variables:

$$\frac{\Delta_{\hat{a}_1}}{T^2} = m_{YX_1} - m_{1X_1} \cdot m_{1X_2}$$

El desarrollo anterior es equivalente al de la covarianza entre ambas variables, es decir:

$$\frac{\Delta_{\hat{a}_1}}{T^2} = cov\,(Y X_1)$$

Que puede escribirse como:

$$\Delta_{\hat{a}_1} = T^2 cov\,(Y\,X_1)$$

De este modo, la expresión (25) resulta equivalente a:

$$\hat{a}_1 = \frac{\Delta_{\hat{a}_1}}{\Delta} = \frac{T^2 \cdot cov\,(Y X_1)}{T^2 \cdot S_{X_1}^2}$$

La expresión final para $\hat{a}_1$ es entonces,

$$\hat{a}_1 = \frac{\mathrm{cov}\,(Y X_1)}{S_{X_1}^2}$$

A partir ahora de la primera ecuación del sistema de ecuaciones que, se transcribe a continuación:

$$\sum_{t=1}^{T} Y_t = T\hat{a}_0 + \hat{a}_1 \cdot \sum_{t=1}^{T} X_{1t}$$

Se dividen ambos miembros por T, con lo cual:

$$\frac{\sum_{t=1}^{T} Y_t}{T} = \hat{a}_0 + \hat{a}_1 \cdot \frac{\sum_{t=1}^{T} X_{1t}}{T}$$

O sea:

$$\bar{Y} = \hat{a}_0 + \hat{a}_1 \bar{X}_1$$

De este modo, $\hat{a}_0$ resulta equivalente a:

$$\hat{a}_0 = \bar{Y} - \hat{a}_1 \bar{X}_1$$

Determinación de las varianzas y covarianza de los estimadores

A partir del desarrollo realizado para el modelo lineal general en relación a la matriz de varianzas y covarianzas, se tiene que:

$$\sum_{\hat{a}\hat{a}} = \left(X^t X\right)^{-1} \cdot \hat{\sigma}_u^2$$

En el caso de una variable explicativa, la matriz $(X^t X)^{-1}$ es equivalente a:

$$\left(X^tX\right)^{-1} = \begin{bmatrix} T & \sum_{t=1}^{T} X_{1t} \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 \end{bmatrix}^{-1}$$

El determinante asociado a la matriz está dado por:

$$\Delta\left(X^tX\right) = T\sum_{t=1}^{T} X_{1t}^2 - \left(\sum_{t=1}^{T} X_{1t}\right)^2$$

Que ya ha sido previamente resuelto y cuyo resultado se transcribe a continuación:

$$\Delta\left(X^tX\right) = T^2 \cdot S_{X_1}^2$$

Donde

$$S_{X_1}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(X_{1t} - \bar{X}_1\right)^2$$

Con lo cual:

$$\Delta\left(X^tX\right) = T\sum_{t=1}^{T}\left(X_{1t} - \bar{X}_1\right)^2$$

La matriz adjunta de la matriz $\left(X^tX\right)$ está dada por:

$$Adj\left(X^tX\right) = \begin{bmatrix} \sum_{t=1}^{T} X_{1t}^2 & -\sum_{t=1}^{T} X_{1t} \\ -\sum_{t=1}^{T} X_{1t} & T \end{bmatrix}$$

Se tratará de obtener $\sum_{t=1}^{T} X_{1t}^2$ a partir de los momentos centrados del siguiente modo:

La varianza $S_{X_1}^2$ puede escribirse, por su relación con los momentos absolutos, como:

$$S_{X_1}^2 = m_2 - \bar{X}_1^2$$

De donde:

$$m_2 = S_{x_1}^2 + \bar{X}_1{}^2$$

Teniendo en cuenta que:

$$m_2 = \frac{\sum_{t=1}^{T} X_{1t}^2}{T}$$

Puede establecerse que:

$$\sum_{t=1}^{T} X_{1t}^2 = T\left[S_{X_1}^2 + \bar{X}_1^2\right]$$

Al distribuir T en el segundo miembro, se llega a la siguiente igualdad:

$$\sum_{t=1}^{T} X_{1t}^2 = \sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2 + T\bar{X}_1^2$$

De este modo, la matriz adjunta de $\left(X^t X\right)$puede expresarse como:

$$Adj\left(X^t X\right) = \begin{bmatrix} \sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2 + T\bar{X}_1^2 & -\sum_{t=1}^{T} X_{1t} \\ -\sum_{t=1}^{T} X_{1t} & T \end{bmatrix} \qquad (26)$$

La matriz inversa de $\left(X^t X\right)$se define como:

$$\left(X^t X\right)^{-1} = \frac{Adj\left(X^t X\right)}{\Delta\left(X^t X\right)}$$

Se reemplaza la matriz adjunta por el resultado obtenido en (26); con lo cual:

$$\left(X^t X\right)^{-1} = \frac{\begin{bmatrix} \sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2 + T\bar{X}_1^2 & -\sum_{t=1}^{T} X_{1t} \\ -\sum_{t=1}^{T} X_{1t} & T \end{bmatrix}}{T\sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2}$$

De este modo, teniendo en cuenta que:

$$\sum_{\hat{a}\hat{a}} = \hat{\sigma}_u^2 \left(X^t X\right)^{-1}$$

Y que para el cálculo de las varianzas de los estimadores mínimo cuadráticos deben tomarse los elementos ubicados en le diagonal principal, se obtiene la siguiente expresión para la varianza estimada de $\hat{a}_0$:

$$\hat{\sigma}^2\left(\hat{a}_0\right) = \left[\frac{\sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2 + T\bar{X}_1^2}{T\sum_{t=1}^{T} \left(X_{1t} - \bar{X}_1\right)^2}\right] \cdot \hat{\sigma}_u^2$$

Que resulta equivalente a:

$$\hat{\sigma}^2(\hat{a}_0) = \left[\frac{1}{T} + \frac{\bar{X}_1^2}{\sum_{t=1}^{T}(X_{1t} - \bar{X}_1)^2}\right] \cdot \hat{\sigma}_u^2$$

Y esta otra expresión para la varianza estimada de a_1:

$$\hat{\sigma}^2(\hat{a}_1) = \frac{T}{T\sum_{t=1}^{T}(X_{1t} - \bar{X}_1)^2} \cdot \hat{\sigma}_u^2$$

Que puede escribirse como:

$$\hat{\sigma}^2(\hat{a}_1) = \frac{\hat{\sigma}_{\hat{u}}^2}{\sum_{t=1}^{T}(X_{1t} - \bar{X}_1)^2}$$

En forma análoga puede obtenerse el valor de la covarianza como:

$$\hat{\text{cov}}(\hat{a}_0\hat{a}_1) = \left[\frac{-\sum_{t=1}^{T} X_{1t}}{T\sum_{t=1}^{T}(X_{1t} - \bar{X}_1)^2}\right] \cdot \hat{\sigma}_u^2$$

Que puede escribirse del siguiente modo:

$$\hat{cov}(\hat{a_0}\hat{a_1}) = \frac{-\bar{X}_1}{\sum_{t=1}^{T}(X_{1t} - \bar{X}_1)^2} \cdot \hat{\sigma}_u^2$$

Cálculo de la varianza de $\hat{Y}$

Se sabe que

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X$$

Con lo cual, el valor medio de $\hat{Y}$ está dado por:

$$E(\hat{Y}) = E(\hat{a}_0 + \hat{a}_1 X_1)$$

Aplicando las propiedades de esperanza a la expresión anterior, se llega a:

$$E(\hat{Y}) = E(\hat{a}_0) + X_1 E(\hat{a}_1)$$

X_1 sale fuera del operador esperanza dado que es una variable matemática.

Teniendo en cuenta que los $\hat{a}_i$ (i = 0,1) son estimadores insesgados, respectivamente, de los parámetros a_0 y a_1, se obtiene:

$$E(\hat{Y}) = a_0 + a_1 X_1$$

De este modo, la variable centrada $\hat{Y} - E(\hat{Y})$ resulta:

$$\hat{Y} - E(\hat{Y}) = \hat{a}_0 + \hat{a}_1 X_1 - a_0 - a_1 X_1$$

Que puede escribirse como:

$$\hat{Y} - E(\hat{Y}) = (\hat{a}_0 - a_0) + (\hat{a}_1 - a_1)\, X_1 \qquad (27)$$

La varianza de $\hat{Y}$ se define como:

$$\text{var}(\hat{Y}) = E\left[[\hat{Y} - E(\hat{Y})]^2\right]$$

Al reemplazar el corchete por la expresión obtenida en (27), se obtiene que:

$$\text{var}(\hat{Y}) = E\left[(\hat{a}_0 - a_0) + (\hat{a}_1 - a_1)\, X_1\right]^2$$

Se desarrolla el binomio al cuadrado:

$$\text{var}(\hat{Y}) = E\left[(\hat{a}_0 - a_0)^2 + 2\,(\hat{a}_0 - a_0)\,(\hat{a}_1 - a_1)\, X_1 + (\hat{a}_1 - a_1)^2 \cdot X_1^2\right.$$

Se aplican las propiedades del operador esperanza y se llega a la siguiente expresión:

$$V(\hat{Y}) = E\left[\hat{a}_0 - a_0\right)^2] + 2X_1 E\left[(\hat{a}_0 - a_0))\,(\hat{a_1} - a_1)\right] + E\,(\hat{a}_1 - a_1)^2 \cdot X_1^2$$

Que se reduce a:

$$V(\hat{Y}) = \hat{\sigma}^2(\hat{a}_0) + 2X_1 c\hat{o}v(\hat{a_0}\hat{a_1}) + \hat{\sigma}^2(\hat{a_1})X_1^2$$

Si se reemplazan las expresiones obtenidas para las varianzas y covarianza de los estimadores, se llega a:

$$V(\hat{Y}) = \hat{\sigma}_u^2 \cdot \left[\frac{1}{T} + \frac{\bar{X}_1^2}{\sum_{t=1}^{T}\left(X_{1t} - \bar{X}_1\right)^2}\right] + 2X_1 \left[-\frac{\bar{X}_1}{\sum_{t=1}^{T}\left(X_{1t} - \bar{X}_1\right)^2}\right] \cdot \hat{\sigma}_u^2 + X_1^2 \cdot \frac{\hat{\sigma}_u^2}{\sum_{t=1}^{T}\left(X_{1t} - \bar{X}_1^2\right)^2}$$

Operando algebraicamente la expresión anterior, se obtiene que:

$$V(\hat{Y}) = \hat{\sigma}_{\hat{u}}^2 \left[\frac{1}{T} + \frac{(\bar{X}_1^2 - 2X_1\bar{X}_1 + X_1^2)}{\sum_{t=1}^{T} (X_1 - \bar{X}_1)^2} \right]$$

Que puede resumirse del siguiente modo:

$$V(\hat{Y}) = \hat{\sigma}_{\hat{u}}^2 \left[\frac{1}{T} + \frac{(X_1 - \bar{X}_1)^2}{\sum_{t=1}^{T} (X_{1t} - \bar{X}_1)^2} \right]$$

A partir de esta expresión correspondiente a la varianza de $\hat{Y}$, pueden obtenerse las siguientes conclusiones. La variabilidad de la estimación es menor:

1. A medida que aumenta al tamaño de la muestra
2. Cuanto más próximo sea el valor a estimar al promedio de los datos

Puede a partir de estas relaciones construirse un intervalo de confianza para la estimación; que adoptará la siguiente forma:

$$\left[\hat{Y} \mp t_{T-k-1}^{*} \cdot \hat{\sigma}(\hat{Y}) \right]$$

Coeficiente de correlación

Este coeficiente –ρ o *r*- es una medida que se utiliza para el caso que el modelo esté explicado por una sola variable.

Mide la intensidad de la relación, siendo su intervalo de existencia: $-1 \leq \rho \leq 1$.

Se dice que la relación es más intensa cuanto más próximo a uno, en valor absoluto, resulte el mismo.

La fórmula de cálculo del coeficiente de correlación está dada por la siguiente expresión:

$$\rho = \frac{cov\,(Y X_1)}{S_{X_1} \cdot S_Y}$$

De modo que el signo del coeficiente de correlación estará determinado por el de la covarianza que es quien define el sentido de la relación.

Relación entre el coeficiente de correlación y la pendiente

De acuerdo con el desarrollo efectuado en el presente capítulo, la fórmula de la pendiente está dada por:

$$\hat{a}_1 = \frac{cov(YX_1)}{S_{X_1}^2}$$

Si se parte de la fórmula correspondiente al coeficiente de correlación:

$$\rho = \frac{cov\,(YX_1)}{S_{X_1} \cdot S_Y}$$

Y se multiplica y divide la expresión por S_{X1}, se tiene que:

$$\rho = \frac{cov(YX_1)}{S_{X_1} S_Y} \cdot \frac{S_{X_1}}{S_{X_1}}$$

Que puede escribirse como:

$$\rho = \frac{cov\,(YX_1)}{S_{X_1}^2} \cdot \frac{S_{X_1}}{S_Y}$$

De este modo, se obtiene la siguiente relación:

$$\rho = \hat{a}_1 \cdot \frac{S_{X_1}}{S_Y}$$

A partir de esta expresión se obtiene como conclusión que el signo del coeficiente de correlación debe coincidir con el signo de la pendiente.

Ello es así porque se está trabajando sobre un mismo conjunto de datos para la obtención de ambas medidas; es decir, la relación es la misma.

Otra recta de regresión:

A veces, el investigador puede estar interesado en conocer también la recta que exprese a la variable X_1 en función de la variable Y. Se presenta entonces la siguiente ecuación:

$$\hat{X}_{1t} = \hat{b}_0 + \hat{b}_1 Y_t$$

Donde $\hat{b}_0$ y $\hat{b}_1$ son respectivamente, la ordenada al origen y la pendiente estimadas de esta nueva recta.

Las fórmulas correspondientes a las mismas están dadas por las siguientes expresiones:

Pendiente

$$\hat{b}_1 = \frac{cov(X_1 Y)}{S_Y^2}$$

Intersección

$$\hat{b}_0 = \bar{X}_1 - \hat{b}_1 \bar{Y}$$

Resulta interesante resaltar las siguientes relaciones que se verifican:

- El signo de las pendientes de ambas rectas y del coeficiente de correlación coincide dado que se está trabajando con la misma base de datos.
- El producto de las pendientes de las respectivas rectas de regresión es igual al coeficiente de determinación:

$$\hat{a}_1 \times \hat{b}_1 = R^2$$

Capítulo 5

Modelos no lineales

Se procede a continuación a indicar cómo pueden ser resueltos algunos modelos que no responden a la forma lineal empleando los conceptos desarrollados para modelos lineales.

Parábola

Sea la función:

$$Y_t = a_0 + a_1 X_{1t} + a_2 X_{1t}^2 + u_t \tag{28}$$

Debe tenerse presente que el método de los mínimos cuadrados clásicos solo puede ser aplicado a modelos lineales. Es por ello que la primera tarea consiste en linealizar estos modelos.

Al aplicar el criterio de los mínimos cuadrados clásicos, se obtiene el siguiente sistema de ecuaciones:

$$\begin{cases} \sum_{t=1}^{T} Y_t = T\hat{a}_0 + \hat{a}_1 \sum_{t=1}^{T} X_{1t} + \hat{a}_2 \sum_{t=1}^{T} X_{1t}^2 \\ \sum_{t=1}^{T} Y_t X_{1t} = \hat{a}_0 \sum_{t=1}^{T} X_{1t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}^2 + \hat{a}_2 \sum_{t=1}^{T} X_{1t}^3 \\ \sum_{t=1}^{T} Y X_{1t}^2 = \hat{a}_0 \sum_{t=1}^{T} X_{1t}^2 + \hat{a}_1 \sum_{t=1}^{T} X_{1t}^3 + \hat{a}_2 \sum_{t=1}^{T} X_{1t}^4 \end{cases} \tag{29}$$

Que puede resolverse por cualquier método de resolución de sistemas de ecuaciones. Son de aplicación los supuestos de Gauss-Markov para la determinación de los coeficientes de regresión, distribución de los estimadores, intervalos de confianza y test de hipótesis.

Modelo exponencial

La forma del modelo está dada por la siguiente expresión:

$$Y_t = e^{a_0} \cdot {a_1}^{X_{1t}} \cdot {a_2}^{X_{2t}} \cdot \ldots \cdot {a_k}^{X_{kt}} \cdot e^{u_t} \tag{30}$$

Aplicando logaritmos naturales, se tiene que:

$$\ln Y_t = a_0 + X_{1t} \ln a_1 + X_{2t} \ln a_2 + \ldots + X_{kt} \ln a_k + u_t \qquad (31)$$

Haciendo las siguientes transformaciones:

$$Z_t = \ln Y_t$$

$$A_0 = \ln a_0$$

$$A_i = \ln a_i \qquad (i = 1,2\ldots k)$$

Y reemplazando las mismas en (31), se tiene que:

$$Z_t = A_0 + A_1 X_{1t} + A_2 \cdot X_{2t} + \ldots + A_k \cdot X_{kt} + u_t \qquad (32)$$

Se observa que se ha obtenido un modelo linealizado en las variables Z_t y X_{it} que se resuelve del mismo modo que el desarrollado en el caso del modelo lineal, pudiéndose obtener de este modo los distintos valores de A_i.

Los supuestos de Gauss-Markov se verifican para el modelo linealizado con el cual ha de realizarse toda la represión.

En este caso, la interpretación del coeficiente A_i es la siguiente:

Una variación de una unidad en la variable X_{it} produce una variación porcentual, en el (mismo o contrario, según el signo de A_i) sentido, de A_i en la variable Y_t, siempre que el resto de las variables permanezca constante.

Para volver a la expresión del modelo original habrá que aplicar antilogaritmos.

Modelo potencial

La forma del modelo está dada por la siguiente expresión:

$$Y_t = e^{a_0} \cdot X_{1t}^{a_1} \cdot X_{2t}^{a_2} \ldots X_{kt}^{a_k} \cdot e^{u_t} \qquad (33)$$

Aplicando logaritmos naturales, se tiene que:

$$\ln Y_t = a_0 + a_1 \ln X_{1t} + a_2 \ln X_{2t} + \ldots + a_k \ln X_{kt} + u_t \qquad (34)$$

Haciendo las siguientes transformaciones:

$$Z_t = \ln Y_t$$

$$X^*_{it} = \ln X_{it}$$

Y reemplazando las mismas en (34), se tiene que:

$$Z_t = a_0 + a_1 X^*_{1t} + a_2 X^*_{2t} + \ldots + a_k X^*_{kt} + u_t \qquad (35)$$

Resultando así un sistema lineal en las variables Z_t y X^*_{it} siendo válidas las consideraciones efectuadas para el modelo exponencial.

En este caso, la interpretación del coeficiente a_i resulta la siguiente: una variación porcentual de una unidad en la variable X_{it} produce una variación porcentual, en el (mismo o contrario, según el signo de a_i) sentido de a_i unidades en la variable Y_t, siempre que el resto de las variables permanezca constante. En el caso del modelo potencial, ambas variaciones son porcentuales; es decir, responden al concepto económico de elasticidad.

Capítulo 6

Distintos métodos de resolución aplicando Excel

Se supone que se desean explicar las ventas mensuales en miles de unidades en una fábrica de automóviles a través del siguiente modelo de represión múltiple:

$$Y_t = a_0 + a_1 X_{1t} + a_2 X_{2t} + u_t$$

Donde:

- Y_t: Ventas mensuales en miles de unidades
- X_{1t}: Precio del producto (en millones de pesos)
- X_{2t}: Gastos mensuales por capacitación del personal (en millones de pesos)

A tal fin se dispone de la siguiente información:

Y	X_1	X_2
5,298	0,85	2,4
8,100	0,92	3,1
4,506	0,97	1,4
4,816	1,10	2,3
9,768	0,98	4,5
6,486	1,15	2,8
2,022	1,15	1,4
4,676	1,20	1,9
5,524	1,20	2,5
4,152	1,22	2,6

El problema consiste entonces en estimar los parámetros a_i por alguna técnica de represión. El método que se utilizará es el de los mínimos cuadrados clásicos.

Primer método: Método de determinantes para resolución de sistemas de ecuaciones

El sistema de ecuaciones resultante para k variables explicativas es:

$$\begin{cases} \sum_{t=1}^{T} Y_t = \hat{a}_0 T + \hat{a}_1 \sum_{t=1}^{T} X_{1t} + \hat{a}_2 \sum_{t=1}^{T} X_{2t} + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{kt} \\ \sum_{t=1}^{T} Y_t X_{1t} = \hat{a}_0 \sum_{t=1}^{T} X_{1t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}^2 + \hat{a}_2 \sum_{t=1}^{T} X_{1t}X_{2t} + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{1t}X_{kt} \\ \sum_{t=1}^{T} Y_t X_{2t} = \hat{a}_0 \sum_{t=1}^{T} X_{2t} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}X_{2t} + \hat{a}_2 \sum_{t=1}^{T} X_{2t}^2 + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{2t}X_{kt} \\ \vdots \\ \sum_{t=1}^{T} Y_t X_{kt} = \hat{a}_0 \sum_{t=1}^{T} X_{kt} + \hat{a}_1 \sum_{t=1}^{T} X_{1t}X_{kt} + \hat{a}_2 \sum_{t=1}^{T} X_{2t}X_{kt} + \cdots + \hat{a}_k \sum_{t=1}^{T} X_{kt}^2 \end{cases}$$

En el ejemplo planteado sólo hay dos variables explicativas con lo cual, se debe resolver un sistema de tres ecuaciones con tres incógnitas.

Todos los valores expuestos en este sistema de ecuaciones pueden obtenerse a partir de una planilla de Excel mediante la aplicación de alguna de las funciones con que la misma cuenta.

Con el objeto de poder aplicar el método de determinantes para la resolución del sistema, resulta necesario en primer término calcular la suma de las columnas correspondientes a los datos.

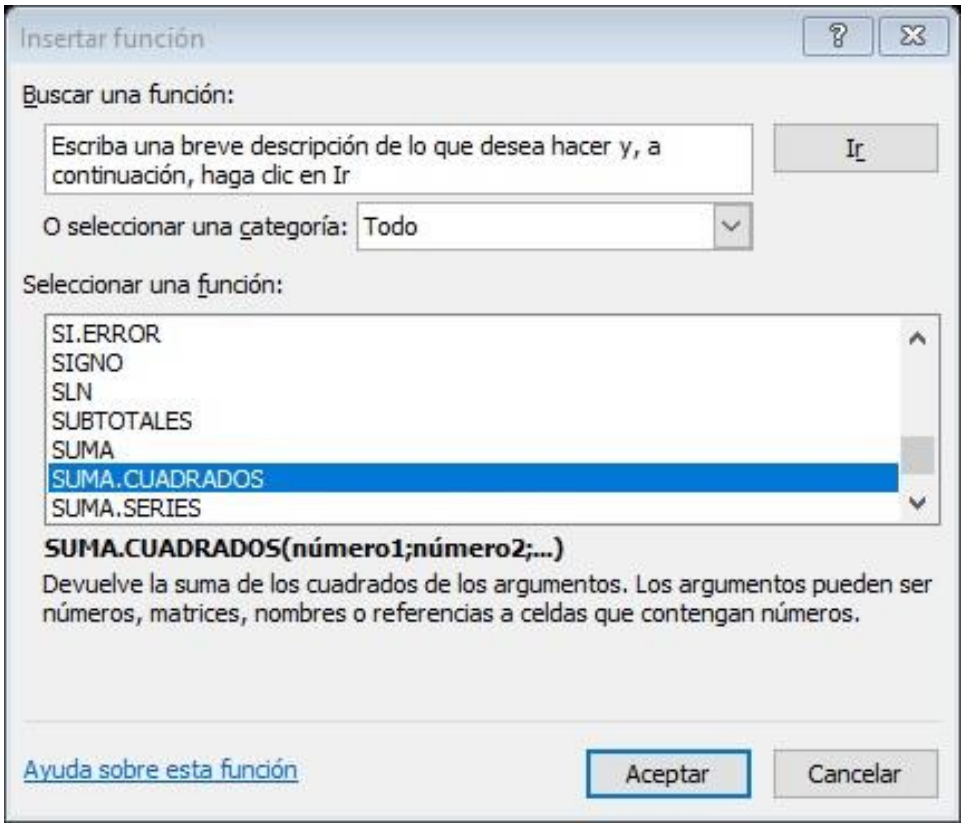
A tal fin, en la planilla de Excel se aplica:

Para sumar la columna de los valores correspondientes a la variable Y, por ejemplo, se aplica $Suma\ (F9{:}\ F18)$ indica que están sumando todos los valores que se encuentran en la columna F desde el renglón 9 al 18, ambos inclusive.

- Función Suma ∑: A partir de su aplicación se obtendrá la suma de las columnas correspondientes a la variable Y que se desea explicar y a las k variables explicativas X. Para obtenerlas, se debe marcar la columna correspondiente a los datos que se desean sumar del siguiente modo:

Para sumar las columnas de los valores correspondientes a los cuadrados de los valores de las variables explicativas, se aplica la función $SUMA.CUADRADOS$

- Función Suma Cuadrados: A partir de su aplicación se obtendrá la suma de los cuadrados correspondientes a cada uno de los números insertos en cada una de las columnas.

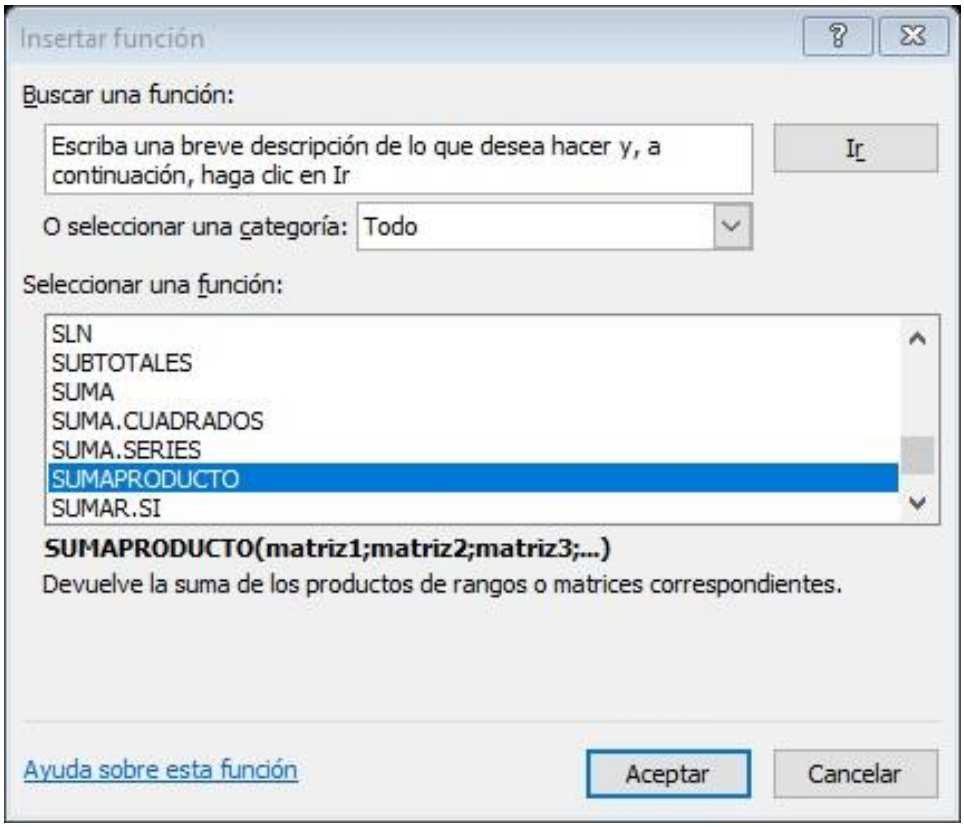


Para obtenerlas, se debe marcar la columna correspondiente a los datos cuyos cuadrados se desean sumar del siguiente modo:

$SUMA.CUADRADOS(G9{:}\ G18)$ indica que está sumando todos los valores –cada uno de ellos elevados al cuadrado – correspondientes a la variable X_1 que se encuentran en la columna G desde el renglón 9 al 18, ambos inclusive. [7]

- Función Suma Producto: A partir de su aplicación se obtendrá la suma de los productos entre cada uno de los números insertos en la columna correspondiente a la variable Y que se desea explicar y cada una de las k variables explicativas X, o entre los correspondientes a dos variables explicativas. [8]

[7] El resultado de la suma aparece en la columna I
[8] En un sistema de k+1 incógnitas se tienen $k\times(k+1)/2$ combinaciones de sumaproducto

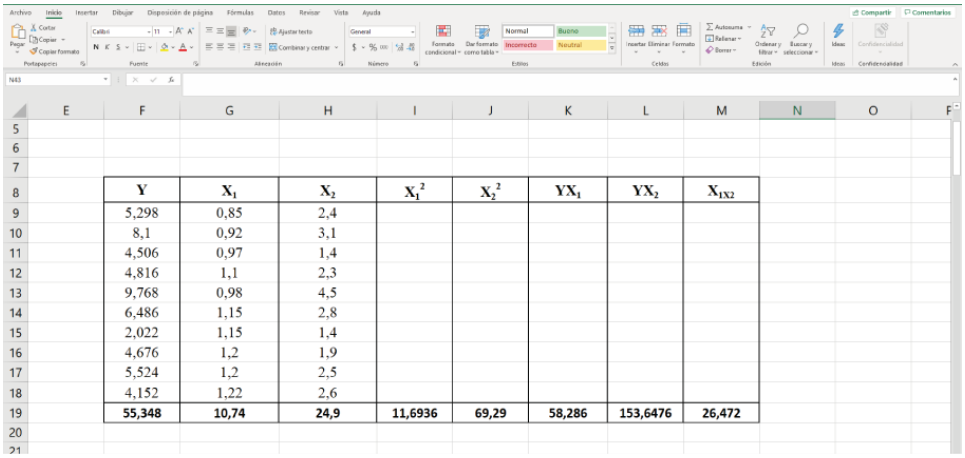


Para obtenerlas, se debe marcar la columna correspondiente a cada uno de los datos que se desean sumar del siguiente modo:

$SUMAPRODUCTO\ (F9{:}\ F18;\ G9{:}\ G18)$ indica que están sumando todos los productos entre los valores insertos en las columnas F y G que se encuentran en las citadas columnas desde el renglón 9 al 18, ambos inclusive. Es el producto entre los valores de Y e X_1 [9].

Se muestra entonces la planilla de Excel obtenida con los correspondientes resultados:

[9] El resultado de la suma de esos productos aparece en la columna K



Y	X_1	X_2	X_1^2	X_2^2	YX_1	YX_2	X_{1X2}
5,298	0,85	2,4					
8,1	0,92	3,1					
4,506	0,97	1,4					
4,816	1,1	2,3					
9,768	0,98	4,5					
6,486	1,15	2,8					
2,022	1,15	1,4					
4,676	1,2	1,9					
5,524	1,2	2,5					
4,152	1,22	2,6					
55,348	**10,74**	**24,9**	**11,6936**	**69,29**	**58,286**	**153,6476**	**26,472**

El resumen de los resultados obtenidos es el siguiente:

- $SUMA(F9:F18) = 55{,}348$
- $SUMA(G9:G18) = 10{,}74$
- $SUMA(H9:H18) = 24{,}90$

- $SUMA.CUADRADOS(I9:I18) = 11{,}6936$
- $SUMA.CUADRADOS(J9:J18) = 69{,}29$

- $SUMAPRODUCTO(F9:F18;G9:G18) = 58{,}286$
- $SUMAPRODUCTO(F9:F18;H9:H18) = 153{,}6476$
- $SUMAPRODUCTO(G9:G18;H9:H18) = 26{,}472$

Resolución el sistema a través de los determinantes

En el modelo lineal general, el determinante del sistema está dado por la siguiente expresión:

$$\begin{matrix} T & \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{2t} & \cdots & \sum_{t=1}^{T} X_{kt} \\ \sum_{t=1}^{T} X_{1t} & \sum_{t=1}^{T} X_{1t}^2 & \sum_{t=1}^{T} X_{1t}X_{2t} & \cdots & \sum_{t=1}^{T} X_{1t}X_{kt} \\ \sum_{t=1}^{T} X_{2t} & \sum_{t=1}^{T} X_{1t}X_{2t} & \sum_{t=1}^{T} X_{2t}^2 & \cdots & \sum_{t=1}^{T} X_{2t}X_{kt} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{t=1}^{T} X_{kt} & \sum_{t=1}^{T} X_{1t}X_{kt} & \sum_{t=1}^{T} X_{2t}X_{kt} & \cdots & \sum_{t=1}^{T} X_{kt}^2 \end{matrix}$$

En el caso que se está resolviendo, se tiene un determinante de tercer orden –parte superior izquierda del determinante anterior– que se transcribe a continuación:

$$\begin{vmatrix} 10 & 10,74 & 24,9 \\ 10,74 & 11,6936 & 26,472 \\ 24,9 & 26,472 & 69,29 \end{vmatrix}$$

Este determinante es el que recibe el nombre de determinante del sistema.

Los determinantes correspondientes a cada una de las incógnitas surgen de reemplazar en el determinante del sistema la columna correspondiente a su ubicación por la correspondiente a los términos independientes que se detallan a continuación:

55,348
58,286
153,6476

Teniendo en cuenta lo expuesto, el determinante correspondiente a $\hat{a}_0$ es:

$$\begin{vmatrix} 55,348 & 10,74 & 24,9 \\ 58,286 & 11,6936 & 26,472 \\ 153,6476 & 26,472 & 69,29 \end{vmatrix}$$

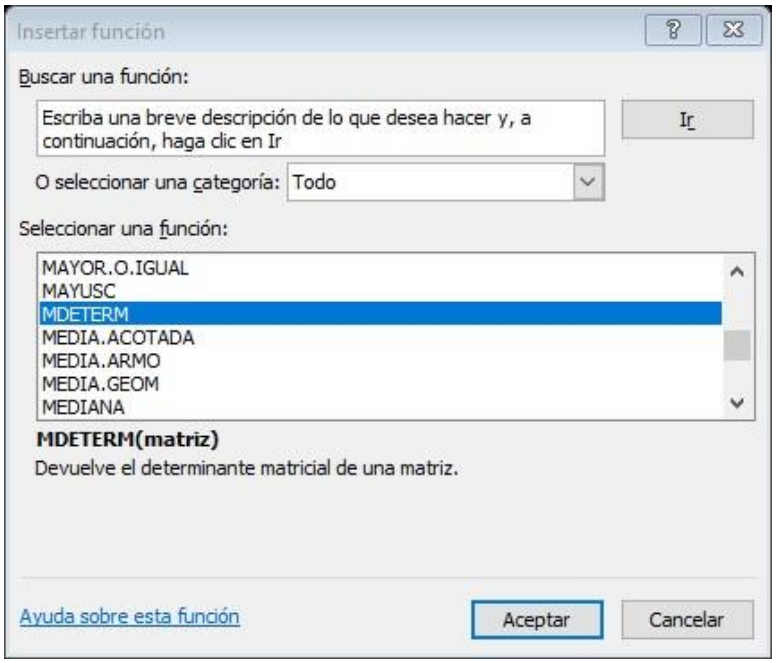
El determinante correspondiente a $\hat{a}_1$ es:

$$\begin{vmatrix} 10 & 55,348 & 24,9 \\ 10,74 & 58,286 & 26,472 \\ 24,9 & 153,6476 & 69,29 \end{vmatrix}$$

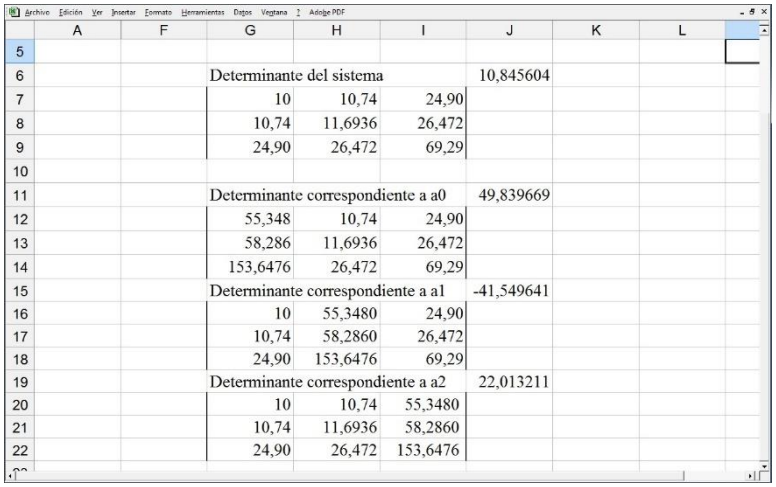
El determinante correspondiente a $\hat{a}_2$ es:

$$\begin{vmatrix} 10 & 10,74 & 55,348 \\ 10,74 & 11,6936 & 58,286 \\ 24,9 & 26,472 & 153,6476 \end{vmatrix}$$

La función de Excel que se aplica para la obtención de los valores correspondientes a los determinantes es *MDETERM*.



Para ello, es necesario entonces incluir como argumento de la mencionada función las celdas donde se encuentran los números correspondientes al determinante. Se muestra la planilla de Excel en la que han sido volcados los datos correspondientes a los determinantes y los resultados obtenidos:



	A	F	G	H	I	J	K	L
5								
6			Determinante del sistema			10,845604		
7			10	10,74	24,90			
8			10,74	11,6936	26,472			
9			24,90	26,472	69,29			
10								
11			Determinante correspondiente a a0			49,839669		
12			55,348	10,74	24,90			
13			58,286	11,6936	26,472			
14			153,6476	26,472	69,29			
15			Determinante correspondiente a a1			-41,549641		
16			10	55,3480	24,90			
17			10,74	58,2860	26,472			
18			24,90	153,6476	69,29			
19			Determinante correspondiente a a2			22,013211		
20			10	10,74	55,3480			
21			10,74	11,6936	58,2860			
22			24,90	26,472	153,6476			

Se tienen así los valores de los cuatro determinantes a partir de la utilización de la función indicada tal como se expone a continuación:

- $MDETERM(\mathrm{G7:I9})$ = 10,845604 ⇒ Determinante del sistema
- $MDETERM(\mathrm{G12:I14})$ = 49,8396689 ⇒ Determinante de $\hat{a}_0$
- $MDETERM(\mathrm{G16:I18})$ = −41,5496408 ⇒ Determinante de $\hat{a}_1$
- $MDETERM(\mathrm{G20:I22})$ = 22,0132106 ⇒ Determinante de $\hat{a}_2$

De este modo, cada estimador resulta del cociente entre el valor del determinante correspondiente a la respectiva incógnita y el determinante del sistema.

Los valores resultantes son:

$\hat{a}_0$ = 4,59537975

$\hat{a}_1$ = −3,83101216

$\hat{a}_2$ = 2,02968968

Resolución en forma matricial

Se definen los siguientes vectores y matrices:

- El vector Y de orden 10×1 está dado por los valores correspondientes a las observaciones de la variable a explicar:

$$Y = \begin{bmatrix} 5,30 \\ 8,10 \\ 4,51 \\ 4,82 \\ 9,77 \\ 6,49 \\ 2,02 \\ 4,68 \\ 5,52 \\ 4,15 \end{bmatrix}$$

- La matriz X de orden 10 × 3 está dada por la siguiente expresión:

$$X = \begin{bmatrix} 1 & 0,85 & 2,40 \\ 1 & 0,92 & 3,10 \\ 1 & 0,97 & 1,40 \\ 1 & 1,10 & 2,30 \\ 1 & 0,98 & 4,50 \\ 1 & 1,15 & 2,80 \\ 1 & 1,15 & 1,40 \\ 1 & 1,20 & 1,90 \\ 1 & 1,20 & 2,50 \\ 1 & 1,22 & 2,60 \end{bmatrix}$$

La segunda y tercera columnas corresponden a las observaciones efectuadas para cada una de las variables explicativas. Asimismo, es necesario completar una primera columna unitaria para que pueda resolverse el sistema.

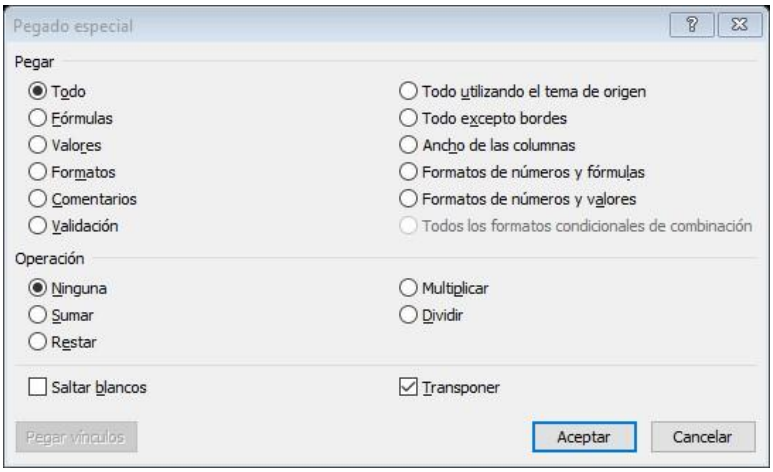
Para poder efectuar los cálculos matriciales, en primer término, debe calcularse la matriz traspuesta de X. La misma, como es sabido, resulta de trasponer filas por columnas. Esta trasposición puede obtenerse mediante la aplicación de Excel siguiendo los siguientes pasos:

1. Se copia la matriz a trasponer:

1	0,85	2,40
1	0,92	3,10
1	0,97	1,40
1	1,1	2,30
1	0,98	4,50
1	1,15	2,80
1	1,15	1,40
1	1,2	1,90
1	1,2	2,50
1	1,22	2,60

2. Se ubica el cursor en la celda donde se ubicará el primer elemento de la primera fila correspondiente a la matriz traspuesta.

3. Se selecciona la función pegado especial y se presiona trasponer.



La matriz traspuesta de la matriz X correspondiente al ejemplo bajo análisis es de orden 3x10:

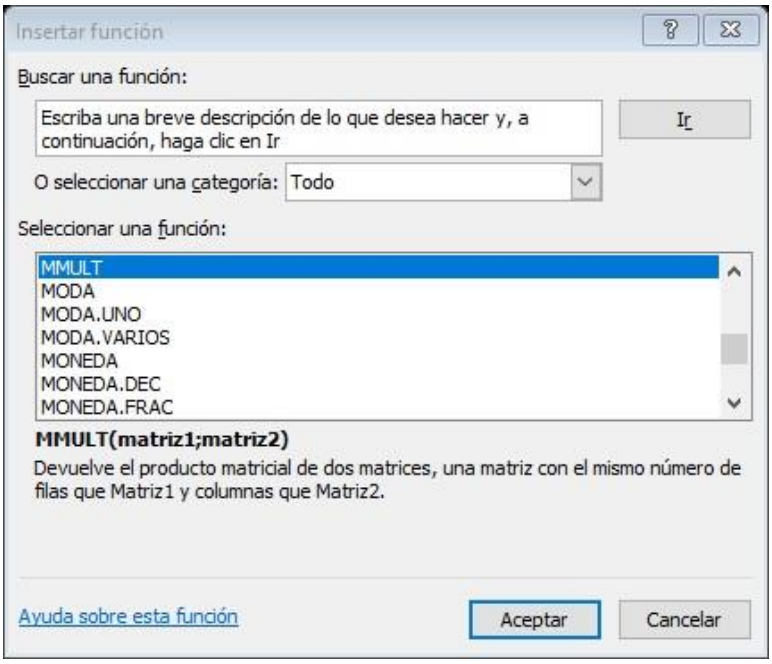
1	1	1	1	1	1	1	1	1	1
0,85	0,92	0,97	1,10	0,98	1,15	1,15	1,20	1,20	1,22
2,40	3,10	1,40	2,30	4,50	2,80	1,40	1,90	2,50	2,60

Los productos entre matrices necesarios para determinar el valor de los estimadores son X^tX y X^tY.

Para realizar el producto entre X^t y X utilizando la planilla de Excel deben seguirse los siguientes pasos:

- Ubicar el cursor en una celda cualquiera.
- Escribir la función +MMULT. (matriz X^t; matriz X)[10]

[10] NOTA: Se marcan las celdas correspondientes a las citadas matrices.



- Volver a la celda donde está escrita la fórmula y marcar arrastrando el cursor, el área donde se ubicará la matriz resultado[11].
- Presionar F2 y luego las teclas correspondientes a control, mayúscula y Enter en forma simultánea. Al hacerlo, aparecerá la matriz resultado.

De este modo, en el ejemplo desarrollado, teniendo en cuenta que X^tes de orden 3 ×10 y X de orden 10 × 3, se tiene que:

Matriz X^tX

$$\begin{bmatrix} 10,00 & 10,74 & 24,90 \\ 10,74 & 11,69 & 26,47 \\ 24,90 & 26,47 & 69,29 \end{bmatrix}$$

En este caso, el área donde se ubicará la matriz resultado es de 3 × 3.

[11] A tal fin, deben tenerse en cuenta las matrices que se multiplican. Es decir, si A es de orden $h \times r$ y B de orden $r \times s$, el producto AB es de orden $h \times s$. En el caso que se analiza, X^t es de orden $(k+1)\times T$ y X de orden $T \times (k+1)$; de donde, la matriz producto resultante es de orden $(k + 1) \times (k + 1)$.

Teniendo además en cuenta que la matriz Y es de orden 3 x 1, se tiene que:

Matriz X^tY

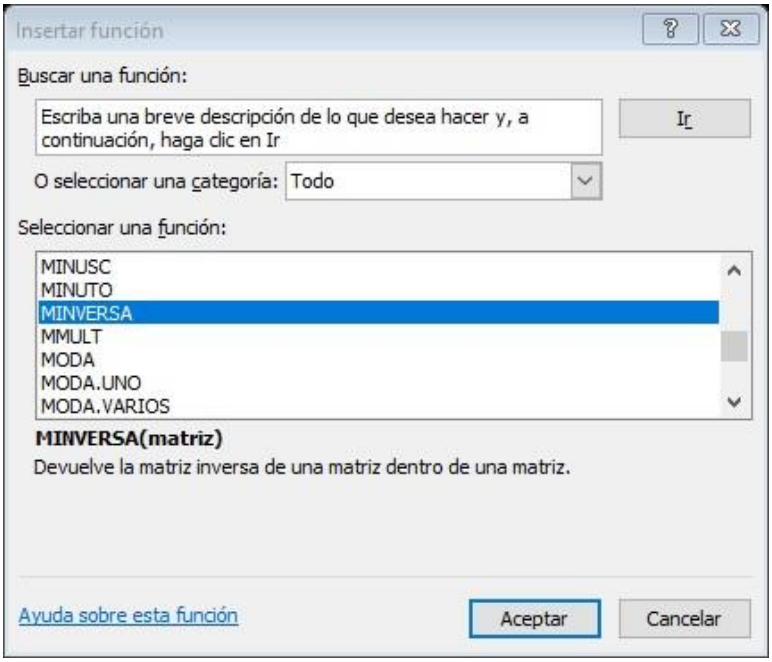
$$\begin{bmatrix} 55,35 \\ 58,29 \\ 153,65 \end{bmatrix}$$

En este caso, el área donde se ubicará la matriz resultado es de 3 × 1.

Una vez obtenida la matriz producto X^tX, debe obtenerse la matriz inversa $(X^tX)^{-1}$[12]

Los pasos a seguir para su obtención son los siguientes:

- Ubicar el cursor en una celda cualquiera.
- Escribir la función *+MINVERSA(matrizX^tX)*



- Volver a la celda donde está escrita la fórmula y marcar arrastrando el cursor, el área donde se ubicará la matriz resultado[13].

[12] NOTA: Debe recordarse que la matriz cuya inversa se calcula debe ser una matriz cuadrada.

[13] NOTA: A tal fin, deben tenerse en cuenta en este caso que el orden de la matriz inversa es el mismo que el de la matriz a invertir; es decir, $(k + 1) \times (k + 1)$ [12]Se trata del concepto matemático de derivada parcial.

- Presionar $F2$ y luego las teclas correspondientes a control, mayúscula y Enter en forma simultánea. Al hacerlo, aparecerá la matriz resultado.

  En el ejemplo desarrollado, la matriz inversa es de orden 3 × 3; con lo cual, el área donde se ubicará la matriz resultado es de 3 × 3:

$$\textbf{Matriz } (\boldsymbol{X^t X})^{-1}$$

$$\begin{bmatrix} 10,09 & -7,84 & -0,63 \\ -7,84 & 6,72 & 0,25 \\ -0,63 & 0,25 & 0,15 \end{bmatrix}$$

  Por último, deben obtenerse los valores de los estimadores mediante la aplicación de la siguiente fórmula:

$$\hat{a} = (X^t X)^{-1} X^t X$$

  Es decir, debe hacerse el producto entre $(X^t X)^{-1}$ y $X^t X$. Se vuelve a utilizar la función $MMULT$ ya explicada:

$$MMULT((X^t X)^{-1}; (X^t X))$$

  Donde $(X^t X)^{-1}$ es de orden (k + 1) × (k + 1) y $X^t X$ de orden (k + 1) × 1; con lo cual, la matriz resultado es de orden (k + 1) × 1.

  En el ejemplo desarrollado, los valores de los estimadores son:

$$\begin{bmatrix} 4,60 \\ -3,83 \\ 2,03 \end{bmatrix}$$

  Con lo cual la función estimada es:

$$\hat{Y}_t = 4{,}59537975 - 3{,}83101216 \cdot X_{1t} + 2{,}02968968 \cdot X_{2t} \qquad (36)$$

  Una variación en una unidad en el precio produce una variación en sentido contrario en las ventas de 3,83101216 unidades siempre y cuando el gasto en capacitación de personal se mantenga constante[14].

  Se determinan a continuación los valores correspondientes a los desvíos de los estimadores mínimos cuadráticos:

[14] NOTA: Se trata del concepto matemático de derivada parcial

- La cantidad de datos observados T es 10.
- La cantidad de variables explicativas k es 2.

Para poder calcular el desvío del residuo de la represión debe en primer término obtenerse $\sum_{t=1}^{T} \hat{u}_t^2$

Debe recordarse que los valores de Y_t son dato, los de $\hat{Y}_t$ se obtienen reemplazando los sucesivos valores de X_{1t} y X_{2t} en la función estimada (36) y los de $\hat{u}_t$ resultan de la diferencia entre ambos.

Se muestran a continuación los resultados:

$\mathbf{Y_t}$	$\mathbf{X_{1t}}$	$\mathbf{X_{2t}}$	$\mathbf{\hat{Y}_t}$	$\mathbf{\hat{u}_t}$	$\mathbf{\hat{u}_t^2}$
5,298	0,85	2,4	6,2165	−0,9185	0,8436
8,100	0,92	3,1	7,3694	0,7306	0,5338
4,506	0,97	1,4	3,7269	0,7791	0,6070
4,816	1,10	2,3	5,0560	−0,2400	0,0576
9,768	0,98	4,5	9,9816	−0,2136	0,0456
6,486	1,15	2,8	5,8795	0,6065	0,3678
2,022	1,15	1,4	3,0375	−1,0155	1,0312
4,676	1,20	1,9	3,8610	0,8150	0,6642
5,524	1,20	2,5	5,0790	0,4450	0,1980
4,152	1,22	2,6	5,2054	−1,0534	1,1097
					5,4586

De este modo, se tiene entonces que:

$$\sum_{t=1}^{T} \hat{u}_t^2 = 5{,}4586$$

Con lo cual;

$$\hat{\sigma}_u^2 = \frac{5{,}4586}{10-2-1} = 0{,}7798 \qquad (37)$$

A fin de obtener las varianzas estimadas de los estimadores, se transcribe la matriz $(X^t X)^{-1}$:

Matriz $(X^t X)^{-1}$

$$\begin{bmatrix} 10,09 & -7,84 & -0,63 \\ -7,84 & 6,72 & 0,25 \\ -0,63 & 0,25 & 0,15 \end{bmatrix}$$

- La varianza estimada de $\hat{a}_0$ se calcula como el producto entre la varianza estimada del residuo de la represión 0.7798, y el elemento w^{00} 10.09.
- La varianza estimada de $\hat{a}_1$ se calcula como el producto entre la varianza estimada del residuo de la represión 0.7798, y el elemento w^{11} 6.72.
- La varianza estimada de $\hat{a}_2$ se calcula como el producto entre la varianza estimada del residuo de la represión 0.7798, y el elemento w^{22}, 0.15.

El cuadro siguiente muestra los valores de las varianzas y de los desvíos estimados de los estimadores mínimo cuadráticos:[15]

Estimador	Varianza estimada	Desvío estimado
$\hat{a}_0$	7,8682	2,8050
$\hat{a}_1$	5,2403	2,2892
$\hat{a}_2$	0,1170	0,3420

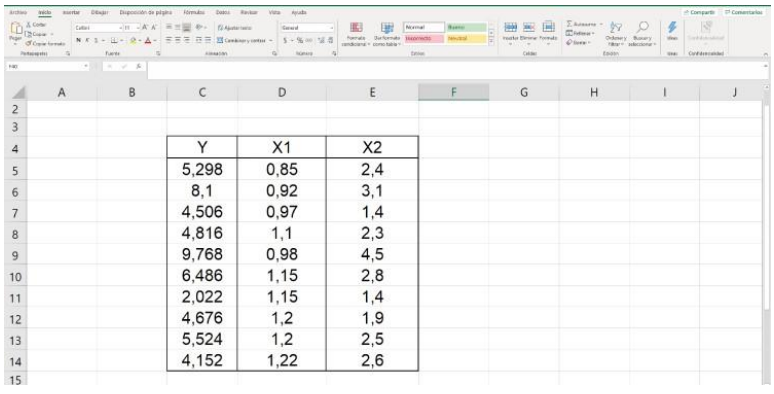
(38)

Resolución mediante la aplicación de la función: análisis de datos

La función análisis de datos se encuentra en la solapa de datos en la planilla de Excel.

A tal fin, debe seguirse el siguiente procedimiento:

1. Se vuelcan los datos en una planilla de Excel:

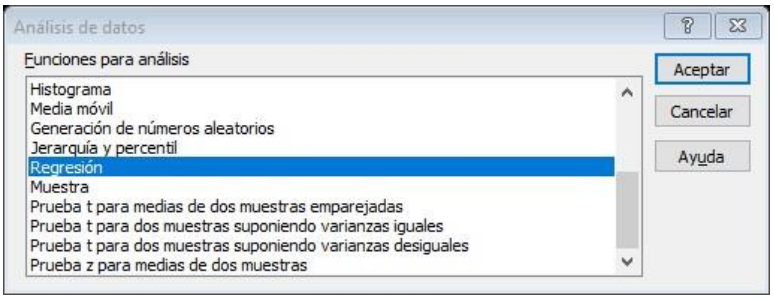


Y	X1	X2
5,298	0,85	2,4
8,1	0,92	3,1
4,506	0,97	1,4
4,816	1,1	2,3
9,768	0,98	4,5
6,486	1,15	2,8
2,022	1,15	1,4
4,676	1,2	1,9
5,524	1,2	2,5
4,152	1,22	2,6

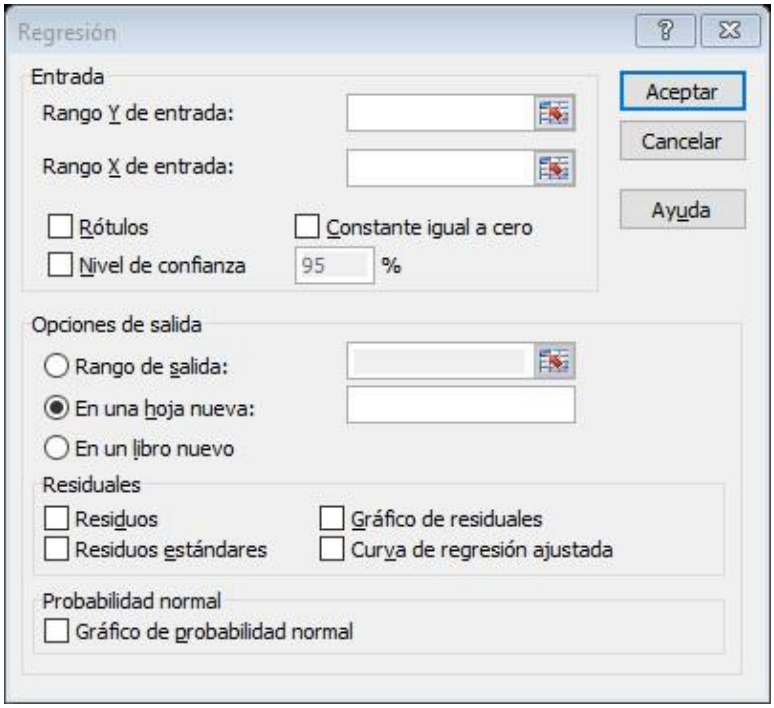
[15] NOTA: Los valores de los desvíos obtenidos pueden no coincidir exactamente con los que se obtienen a partir de la resolución mediante la aplicación de la función análisis de datos porque en este caso no se esta´ trabajando con todos los decimales.

2. Teclear en datos, de este modo a la derecha aparecerá análisis de datos.

3. Abrir “análisis de datos”. Allí se desplegará una lista en la que encontrarán regresión.



4. Abrir regresión. Aparece entonces la siguiente pantalla:



5. ¿Cómo han de introducirse los datos en este cuadro?

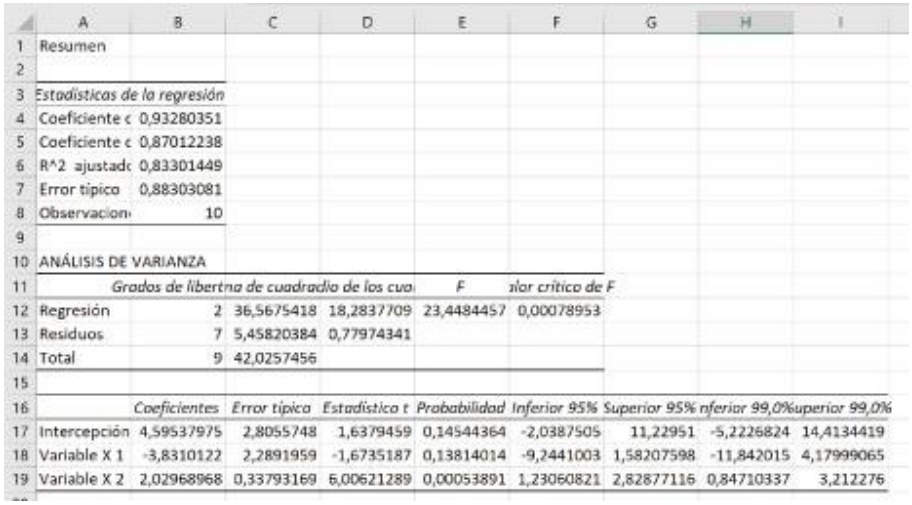
   - Rango Y de entrada: es el vector de valores correspondiente a la variable que se desea explicar.

- Rango X de entrada: es el vector de valores (o matriz si hubiere más de una variable explicativa) correspondientes a la/s variable/s X explicativas.
- Se recomienda sustituir el nivel de confianza por otro distinto al que aparece automáticamente, por ejemplo, 99%.
- En residuales marcar residuos.
- En este caso, como se van a considerar dos variables explicativas, se introducirá la matriz con los valores de X_{1t} y X_{2t}[16].

6. De acuerdo con ello, se obtendrá:

[16] NOTA: Se invita al lector a que efectúe las regresiones correspondientes considerando solo a X_1 como variable explicativa y solo a X_1 como variable explicativa.

7. Al teclear aceptar, se obtiene la siguiente salida:



	A	B	C	D	E	F	G	H	I
1	Resumen								
2									
3	*Estadísticas de la regresión*								
4	Coeficiente c	0,93280351							
5	Coeficiente c	0,87012238							
6	R^2 ajustad	0,83301449							
7	Error típico	0,88303081							
8	Observacion	10							
9									
10	ANÁLISIS DE VARIANZA								
11		*Grados de libertad*	*de cuadrados*	*de los cua*	*F*	*alor crítico de F*			
12	Regresión	2	36,5675418	18,2837709	23,4484457	0,00078953			
13	Residuos	7	5,45820384	0,77974341					
14	Total	9	42,0257456						
15									
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 99,0%*	*uperior 99,0%*
17	Intercepción	4,59537975	2,8055748	1,6379459	0,14544364	-2,0387505	11,22951	-5,2226824	14,4134419
18	Variable X 1	-3,8310122	2,2891959	-1,6735187	0,13814014	-9,2441003	1,58207598	-11,842015	4,17999065
19	Variable X 2	2,02968968	0,33793169	6,00621289	0,00053891	1,23060821	2,82877116	0,84710337	3,212276

Lectura de los resultados obtenidos

Estadísticas de la regresión:

- El valor del coeficiente de determinación R^2 es 0.8701, es decir, el 87,01% de las variaciones en las ventas están explicadas en forma conjunta por las variaciones en el precio y en las inversiones en publicidad.
- El R^2 ajustado o corregido es 0.833, debe recordarse que su valor es siempre inferior al R^2 que le da origen.
- El error típico de 0.883 es el desvío estimado del residuo de la regresión. Es la raíz cuadrada de la varianza obtenida en (36).

¿Cómo se leen los resultados del cuadro de análisis de varianza?

- La columna correspondiente a suma de cuadrados indica la descomposición de las variaciones.
- La variación de regresión – 36.5675 - es la denominada variación explicada; la de residuos – 5.4582 – es la variación no explicada y la suma de ambas – 42.0257 – es la variación total.
- La columna correspondiente a grados de libertad indica justamente que la variación de regresión tiene 2 (k) grados de libertad, la de residuos 7 (T-k-1) grados de libertad y la total 9 grados de libertad (T-1) grados de libertad.

- Si se realiza el cociente entre la columna de suma de cuadrados y la correspondiente a grados de libertad, se obtiene la columna de promedio de cuadrados. En este caso, puede observarse que el promedio de cuadrados correspondiente a los residuos es 0.7797 es decir, el valor correspondiente a la varianza estimada del residuo de la regresión obtenido en (36).
- Además, el cociente entre la variación explicada y la variación total da por resultado el valor del coeficiente de determinación.
- La columna F corresponde al valor del estadístico utilizado en el test de hipótesis para variables conjuntas y surge del cociente entre los valores sobrantes en la columna previa.

¿Cómo se leen los resultados del cuadro siguiente?

- La primera columna indica de qué estimador se trata. Intercepción se corresponde con $\hat{a}_0$, variable X_1 con $\hat{a}_1$ y variable X_2 con $\hat{a}_2$.
- La columna siguiente de coeficientes de da los valores de los estimadores que coinciden con los obtenidos en (35), la columna de error típico da el valor de los desvíos estimados de los estimadores que coinciden con los obtenidos en (37).
- La columna de estadístico t muestra los valores de los correspondientes estadísticos de prueba que se utilizan para la realización de los test de hipótesis para parámetros individuales.
- Las columnas correspondientes a inferior y superior dan los límites inferior y superior del intervalo de confianza para los parámetros individuales. En el primer intervalo se trabaja con un nivel de confianza del 95% mientras que en el segundo se trabaja con un nivel de confianza equivalente al 99%.
- El cuadro de análisis de residuales muestra en la columna de pronóstico el valor estimado de la variable, en este caso ventas, para los sucesivos valores de las variables explicativas mientras que la columna de residuos da los sucesivos valores del error.

Capítulo 7

Ejercicios

1. Se ha observado la demanda por manzanas en un supermercado, encontrándose los resultados de precio y cantidad exhibidos en el cuadro, correspondiente a 12 días consecutivos.

Cantidad (kilos)	**Precio (Cantidad de $ por kilo)**
55	100
70	90
90	80
100	70
90	70
105	70
80	70
110	65
125	60
115	60
130	55
130	50

Se pide:

a) Representar en un gráfico los pares de datos.

b) Encontrar la expresión lineal de la demanda de manzanas en función del precio.

c) Hallar la intensidad de la relación entre la variable dependiente y la independiente.

d) Predecir la demanda para un precio de $62 centavos y encontrar un intervalo de confianza para esa demanda, con un nivel de significación del 5%.

e) Establecer si la variable precio es significativa para explicar el modelo al mismo nivel de significación.

2. Indicar cuál de las variables X_{ii} utilizaría usted en un análisis de regresión y correlación simple para explicar la evolución de la variable Y_i. Contando con los datos abajo expuestos, justifique su respuesta explicando el porqué de su elección.

   - Y_j: Precio de la carne para consumo interno en el mercado de Liniers en el mes de septiembre;
   - X_{i1}: Precio del pollo en las bocas de expendio al público en el mes de septiembre; $r_{yx1} = -0{,}9$.
   - X_{i2}: Nivel del salario real en la Ciudad Autónoma de Buenos Aires en el mes de septiembre; $r_{yx2} = 0{,}95$.
   - X_{i3}: Precio promedio de los automóviles importados en el mes de septiembre; $r_{yx3} = 1$.
   - X_{i4} Nivel de lluvias en la Provincia de Buenos Aires en el mes de agosto, $r_{yx4} = -0{,}77$.

3. Se supone que Sudáfrica compite con Argentina en la colocación de carne en el Mercado Europeo. El coeficiente de correlación entre las ventas de carne de Argentina y de Sudáfrica fue, para un período de 6 años, igual a -0,55. ¿Cree usted que efectivamente existe alguna competencia entre los países?

4. Un estudio realizado sobre las ventas de una conocida marca de vino ha determinado que la publicidad llevada a cabo por la empresa elaboradora influye en las ventas del comercio A y B de la siguiente manera:

   Comercio A: $\hat{Y}_c = 200 + 2{,}5X_1$; $r_{yx} = 0{,}9$

   Comercio B: $\hat{Z}_c = 120 + 1{,}2X_1$; $r_{yx} = 0{,}75$

   Donde Y_i y Z_i son las ventas de vino del comercio A y B respectivamente, expresadas en miles de pesos y X_i es el monto invertido en publicidad por la empresa elaboradora de vino.

   a) ¿Qué comercio venderá más ante un aumento en las inversiones publicitarias? Fundamente su respuesta.

   b) ¿Para cuál de los dos comercios puede predecirse en forma más exacta las ventas de vino en base a las inversiones publicitarias? Fundamente su respuesta.

c) ¿En cuánto se pueden estimar las ventas del comercio B si se invierten $10.000 en publicidad?

d) Determinar qué porcentaje de la variación total de las ventas del comercio A se ve explicada por la inversión en publicidad.

5. A partir de un análisis de regresión y correlación con una muestra de tamaño 15, se obtienen los siguientes resultados:

   - Existe una correlación lineal inversa entre ambas variables.
   - La suma de los desvíos cuadráticos de la variable Y es 350.
   - La correspondiente a la variable X es 220.
   - r^2 es 0,94.
   - La suma de la variable X es 420 y la de la variable Y es 540.

   Se pide determinar la recta de regresión de Y en X.

6. Usted cuenta con la siguiente información en relación a dos rectas de regresión obtenidas a partir del mismo conjunto de datos.

   - Pendiente de Y en función de X: −0,6975
   - Coeficiente de determinación: 0,9123
   - Media de Y : 100
   - Media de X: 130

   Escriba la recta de regresión que permita explicar X en función de Y. Justifique.

7. Dados los siguientes datos:

Ingreso	**Ahorro**
12	Y1
14	Y2
18	Y3
20	Y4
26	Y5

Y sabiendo que:

- Suma de la variable ahorro es 35, la suma de los residuos de la regresión al cuadrado es 1,3; suma de los desvíos de la variable ahorro ajustada elevada al cuadrado es 86,7.
- Existe además una relación directa entre ambas variables.

Se pide estimar el ahorro si el ingreso adopta el valor de 22.

8. Dados los siguientes pares de rectas de regresión, discuta la consistencia de los resultados obtenidos:

   a) $\hat{Y}$= 2,34 + 0,96X $\hat{X}$ = 1,06 − 0,87Y

   b) $\hat{Y}$ = 1,22 − 0,78X $\hat{X}$ = 0,94 − 0,94Y r = 0,7332

9. Efectuada una regresión, se obtuvieron los siguientes datos:

   Y_{ajust} = 3,07 × 1,04^X r = −0,88

   Y_{ajust} = 2,04 + 2,2*X r = 0,78

   a) ¿Qué modelo seleccionaría y por qué?

   b) Interprete los resultados de los coeficientes de regresión del modelo seleccionado.

   c) ¿Qué porcentaje de la varianza del regresando no está explicado por la variable X en la forma seleccionada?

   d) Si Y representara el consumo de un bien determinado y X el precio de éste, ¿seguiría seleccionando ese modelo, ¿por qué?

   e) ¿Qué variable agregaría para mejorar la explicación del modelo y qué signo esperaría en el coeficiente de regresión?

   f) ¿Qué signo esperaría en el coeficiente de correlación múltiple?

10. El siguiente cuadro reúne los datos correspondientes a la venta mensuales de automóviles expresadas en miles de unidades $-Y-$, precio de las unidades en miles de dólares $-X_1-$ y gastos mensuales de mantenimiento expresados en miles de pesos $-X_2-$.

Y	**X_1**	**X_2**
6,288	0,94	2,34
9,115	0,85	3,15
5,127	1,03	1,44
5,222	1,21	2,32
10,088	0,96	4,54
6,327	1,18	3,81
1,922	1,18	1,46
4,874	1,22	2,12
5,745	1,22	2,65
4,121	1,24	3,21

A partir de esta información, se solicita:

a) Construir el modelo de regresión lineal que incluya a ambas variables como explicativas. Interprete los resultados obtenidos.

b) Determinar el valor del coeficiente de determinación, interpretando el resultado obtenido.

c) Pruebe la significatividad individual de cada una de las variables a un nivel de riesgo del 5%. Conclusiones.

d) Repita los puntos anteriores para los siguientes modelos:

1) Considerando sólo al precio como variable explicativa.
2) Considerando sólo al gasto de mantenimiento como variable explicativa.

e) Analice los resultados obtenidos a partir de los tres modelos planteados. Conclusiones.

11. Se desean estimar los gastos en alimentación de una familia –expresados en miles de pesos- en base a la información que proporcionan las variables: ingresos mensuales –expresados en miles de pesos-, X_1 y número de miembros de la familia X_2. Para ello se extrae una muestra aleatoria simple de 15 familias a partir de la cual se obtienen los siguientes resultados:

VARIABLE EXPLICATIVA	**VARIACIONES**	
	Regresión	Residual
Ingreso	12716,2442	1600,15582
Nro. miembros Flia		14087,4192
Ambas	13595,4215	

Modelo ingreso	***Coeficientes***	***Error típico***
Intercepción	16,18989784	4,67952953
Variable ingreso	1,343217934	0,13215284
Modelo número de miembros familia	***Coeficientes***	***Error típico***
Intercepción	67,1269231	30,2119966
Variable Nro. Miembros Flia.	-3,63461538	7,90684088
Modelo ambas variables	***Coeficientes***	***Error típico***
Intercepción	-16,04580427	9,03891033
Variable X 1	1,487270228	0,09971317
Variable X 2	7,691519426	2,0106868

Se pide:

a) Seleccione el mejor modelo para explicar el gasto, fundamentando su elección.

b) Para el modelo seleccionado, interprete el resultado de los coeficientes de regresión y del coeficiente de determinación.

c) Si la relación adoptada por el modelo fuera potencial, en base a los resultados obtenidos, a cuánto se espera que ascienda el gasto si el ingreso fuera de $15.000 y la cantidad de miembros de la familia tres.

d) Bajo el mismo supuesto interprete los coeficientes obtenidos.

e) Indique si las variables ingreso y número de miembros de la familia son significativas en forma conjunta para explicar al gasto en alimentación, a un nivel de riesgo del 5%.

f) Indique si la variable ingreso es significativa para explicar al gasto en alimentación a un nivel de riesgo del 5% en el modelo seleccionado.

g) Indique en qué porcentaje mejora la explicación de los gastos en alimentación la incorporación del número de miembros de la familia como variable explicativa. Conclusiones.

12. Una empresa desea estimar la demanda de un producto $D(t)$, a efectos de evaluar su producción para el futuro. A tal fin, recopilo datos mensuales desde enero 2015 a mayo 2018. Las variables explicativas que serán utilizadas en los distintos modelos son:

- $Y(t)$: ingreso familiar en el período considerado
- $P(t)$: precio del producto para el período considerado
- $S(t)$: precio de un producto sustituto para dicho período

Efectuadas las estimaciones correspondientes, se obtuvieron los siguientes modelos:

- $\underset{(0,121)}{\hat{D}(t) = 10 + 0{,}415Y(t)} \qquad \sum_1^{41} \hat{u}^2 = 315{,}40$
- $\underset{(0,156)}{\hat{D}(t) = 7 - 0{,}354P(t)} \qquad \sum_1^{41} \hat{d}^2 = 839{,}24$
- $\underset{(0,557)\quad(0,431)}{\hat{D}(t) = 1{,}20 + 1{,}35P(t) + 1{,}04S(t)} \qquad \sum_1^{41} \hat{d}^2 = 810{,}15$
- $\underset{(0,201)\quad(0,176)\quad(0,223)}{\hat{D}(t) = 15 + 0{,}332Y(t) - 0{,}414P(t) + 0{,}61S(t)} \qquad \sum_1^{41} \hat{u}^2 = 104{,}78$

Los valores entre paréntesis debajo de cada estimador representan los desvíos típicos del estimador correspondiente. Las letras minúsculas obrantes dentro de las sumas representan los desvíos de las variables.

Se sabe además que $\sum_1^{41} d^2$ = 10240. Se pide:

a) Interprete los coeficientes de regresión del modelo 3.

b) Efectúe la interpretación de los coeficientes de regresión del cuarto modelo si se tratara de un modelo potencial.

c) Establezca la composición de las variaciones en el modelo 3.

d) Indique qué porcentaje de las variaciones de la demanda no están explicadas por el modelo 2.

e) ¿Qué modelo seleccionaría para explicar la demanda del producto y por qué?

f) Pruebe la hipótesis que las variables incluidas en el modelo seleccionado no son significativas en forma conjunta para explicar la demanda del producto a un nivel de significación del 5%.

g) Para el modelo seleccionado, pruebe la hipótesis de que el precio del producto no es significativo para explicar la demanda del producto al mismo nivel de significación. ¿Es congruente el resultado con el obtenido previamente, ¿por qué?

h) ¿Qué porcentaje de las variaciones de la demanda del producto no explicada por el ingreso, se explica por el precio y el precio del artículo sustituto en forma conjunta?

13. Se desea realizar un estudio sobre el riesgo de ataques al corazón. A tal fin se ha seleccionado una muestra de 20 personas a las cuales se les midió el riesgo, se les tomó la presión, y se les consultó la edad y condición (fumador o no fumador). A partir de estos datos, se construyeron cinco modelos de regresión con las siguientes variables explicativas:

    - Edad
    - Presión
    - Edad y presión
    - Presión y condición
    - Todas (edad, presión y condición)

    El siguiente cuadro muestra las variaciones explicadas y no explicadas para los distintos modelos enunciados previamente:

Modelos	**Variación explicada**	**Variación no explicada**
Edad	1571,068447	
Presión		3460,633014
Edad y presión		763,2157829
Presión y cond.	2280,973696	
Todas	3793,648729	378,1512714

El siguiente cuadro muestra los coeficientes estimados para cada uno de los modelos:

Modelos	Ordenada	Coeficiente 1ra. variable	Coeficiente 2da. variable	Coeficiente 3ra. Variable
Edad	−32,35776699	0,873300971		
Presión	−1,334276411	0,187880954		
Edad y Presión	−106,185206	1,218226282	0,321514091	
Presión y Cond.	−2,272307692	0,135996714	18,02243812	
Todas	−88,36442241	1,005225073	0,269834013	9,83543997

El siguiente cuadro muestra el desvío de los estimadores para cada uno de los modelos:

Coef. Variables	Coeficiente 1ra. variable	Coeficiente 2da. variable	Coeficiente 3ra. Variable
Edad	0,264836537		
Presión	0,097687385		
Edad y Presión	0,157164014	0,05025551	
Presión y Cond.	0,075574238	4,797244013	
Todas	0,125650229	0,038645959	2,436687502

A partir del análisis de la información, se pide:

a) ¿Cuál es el mejor modelo para explicar el riesgo de ataques al corazón? Justifique teórica y numéricamente su elección. Interprete los resultados.

b) Determine qué porcentaje de las variaciones del riesgo de ataques al corazón no está explicados en forma conjunta por la edad y la presión.

c) Indique qué porcentaje de las variaciones no explicadas por la edad no son explicadas con la introducción de la presión y la condición.

d) En el modelo seleccionado indique a un nivel de riesgo del 1% si las variables seleccionadas son significativas individualmente para explicar el riesgo de ataque al corazón.

e) Construya un intervalo de confianza al 98% para la variable edad en el modelo donde ´esta es la única variable explicativa.

f) Indique si la variable presión es significativa en el modelo que tiene a la edad y presión como variables explicativas a un nivel de riesgo del 2.5%.

14. Usted cuenta con la siguiente información acerca de tres regresiones efectuadas sobre un conjunto de doce datos:

CONCEPTO	***VARIABLES***	***X1, X2***	***X1,X2,X3***	***X1,X2,X3,X4***
Coeficiente de correlación múltiple		0,733797013	0,75685658	0,76156383
Coeficiente de determinación R^2		0,538458056	0,57283188	0,57997947
R^2 ajustado		0,43589318	0,41264384	0,33996773
Error típico		10,9597578	11,1833276	11,8550357
Observaciones		12	12	12

Suma de cuadrados	**X1,X2**	**X1,X2,X3**	**X1,X2,X3,X4**
Regresión	1261,20338	1341,71547	
Residuos	1081,04662		983,793098

MODELO	***Coeficientes***	***Error típico***
Intercepción	178,7007246	18,0449327
Variable X 1	-2,543478261	3,95819938
Variable X 2	-5,188888889	1,63378423
Intercepción	207,0149572	39,8043636
Variable X 1	-3,925999572	4,39114583
Variable X 2	-4,843095143	1,7219195
Variable X 3	-1,353105964	1,68644545
Intercepción	205,3641424	42,4653742
Variable X 1	-5,298007045	6,12131855
Variable X 2	-4,821398618	1,82642591
Variable X 3	-1,438005407	1,8045831
Variable X 4	1,407668332	4,07856096

Se pide:

a) Seleccione el mejor modelo para explicar a la variable Y , indicando por qué lo selecciona e interpretando el resultado del coeficiente de determinación.

b) Explicite la forma que adoptara el modelo e interprete los coeficientes de regresión del mismo.

c) Realice la interpretación de los coeficientes si la forma del modelo seleccionado fuera potencial.

d) A un nivel de riesgo del 5% establezca si las variables son significativas individualmente para explicar a la variable Y.

e) A un nivel de riesgo del 1% establezca si las variables son significativas en forma conjunta para explicar a la variable Y. Justifique la selección de la zona de rechazo.

f)) Determine qué porcentaje de las variaciones no explicadas por las variables X_1 y X_2 no son explicadas con la introducción de las variables X_3 y X_4 .

15. A efectos de realizar un estudio del consumo (C) en función del ingreso (Y) y del ahorro (A) de períodos anteriores se ha realizado un análisis de regresión para un período de 20 meses a partir del cual se han obtenido los siguientes resultados:
    - $\sum_1^{20} C = 3097$
    - $\sum_1^{20} Y = 423$
    - $\sum_1^{20} A = 47{,}9$
    - $\sum_1^{20}(C-\overline{C})(Y-\overline{Y}) = 4563{,}45$
    - $\sum_1^{20}(C-\overline{C})(A-\overline{A}) = 11776{,}7$
    - $\sum_1^{20}(A-\bar{A})^2 = 7345,50$
    - $\sum_1^{20}(Y-\overline{Y})^2 = 4815{,}8$
    - $\sum_1^{20}(C-\overline{C})^2 = 30704{,}55$

    Se cuenta con la siguiente matriz centrada para dos variables explicativas:

$$(X^tX)^{-1} = \begin{bmatrix} 0{,}02082679 & -0{,}00786357 \\ -0{,}00786357 & 0{,}00317669 \end{bmatrix}$$

    Se solicita:

    a) Considerando una sola variable explicativa, cuál es el mejor modelo para explicar el consumo. Justifique su elección.

    b) Realice la estimación del modelo seleccionado e interprete los resultados obtenidos.

    c) Si el modelo resultante en función de las dos variables explicativas adoptara la siguiente expresión:

$$\hat{C} = 30{,}2515264 + 2{,}43517911\ Y + 1{,}52597986\ A$$

    Se pide:

    1) Con un nivel de riesgo del 5% indique si la variable ahorro es significativa para explicar al consumo.

2) Con un nivel de riesgo del indique 1% si las variables son significativas en forma conjunta.
3) Interprete el coeficiente de regresión correspondiente al ingreso.
4) Determine en cuánto mejora la explicación del modelo la introducción de la segunda variable explicativa al modelo elegido previamente en el punto anterior.

16. Usted cuenta con los siguientes resultados obtenidos luego de efectuar diversas regresiones sobre la variable Y para un conjunto de 14 observaciones:

Concepto	$Y\,X_1$	$Y\,X_3$	$Y\,X_1X_2$	$Y\,X_2X_3$	$Y\,X_1X_2X_3$
Coeficiente Determ.	0,486415	0,042768	0,713419	0,099873	0,72995839
a_0	253,46658	285,2934	288,5550	330,8008	235,6642
a_1	−3,8975640	No existe	−5,114134	No existe	−5,53982
a_2	No existe	No existe	−0,85426	−0,420670	−0,80906847
a_3	No existe	−0,638338	No existe	−0,938130	0,486715
Desvío a_1	1,156124	No existe	0,991716	No existe	1,146864
Desvío a_2	No existe	No existe	0,289403	0,50357	0,300247
Desvío a_3	No existe	0,871778	No existe	0,953108	0,621925
Variación explicada	1737,1999				
Var. no explicada		3418,6833			

Se pide:

a) Seleccione el mejor modelo para explicar a la variable Y. Justifique su elección.

b) Indique qué porcentaje de las variaciones de Y no están explicadas en forma conjunta por X_2 y X_3.

c) Escriba el modelo seleccionado e interprete el resultado de los coeficientes de regresión.

d) Indique si cada una de las variables del modelo seleccionado son significativas individualmente a un nivel de riesgo de 5%.

e) indique si las variables del modelo seleccionado son significativas en forma conjunta para explicar al modelo, al mismo nivel de riesgo. Conclusiones.

f) Indique si la variable X_3 es significativa para explicar a la variable Y en el modelo de tres variables explicativas a un nivel de riesgo del 5%. Compare sus resultados con los previamente obtenidos.

17. La demanda de un producto determinado puede explicarse por las variables precio, precio de un artículo sustituto e ingreso. A partir de la observación de 20 meses relacionada con el mencionado producto pudieron obtenerse los siguientes resultados:

    - Variación total: 1565
    - Variación explicada (por precio): 1321
    - Coeficiente de determinación (precio y precio del artículo sustituto): 81%
    - Variación no explicada (las tres variables expuestas): 205

    Se pide:

    a) Analice la coherencia de los datos.

    b) Seleccione el mejor modelo para explicar a la demanda.

    c) Indique si las variables correspondientes al modelo seleccionado son significativas en forma conjunta para explicar a la demanda a un nivel de riesgo del 2%. Justifique su respuesta.

18. La demanda de un producto determinado puede explicarse por las variables ingreso, precio y precio de un artículo sustituto. A partir de la observación de 15 meses de la demanda del mencionado, del precio del mismo, del precio del producto sustituto y del ingreso, pudieron obtenerse los siguientes resultados:

Variables explicativas	**Variación regresión**	**Variación residual**	**Coeficiente determinación**
Ingreso	552308,477		
Precio		517204,582	
Ingreso y precio			0,80263025
Precio y Precio Sustituto	229566,489	505433,511	
Ingreso, precio y Precio sustituto	603389,289		

Variables explicativas	Coef. Ord. Origen	Desvío Ord. Origen	Coef.	Desvío	Coef.	Desvío	Coef.	Desvío
Ingreso	$345,19370$	$432,53671$	$0,6947276$	$0,110818$				
Precio	$3934,847267$	$381,6748171$	$-51,24598071$	$21,90256335$				
Ingreso y precio	$1040,7017$	$562,46007$	$0,6185474$	$0,1114844$	$-23,103067$	$13,095622$		
Precio y P. Sust.	$4158,1382$	$576,74044$	$-54,202127$	$23,219402$	$-15,851063$	$29,984154$		
Ingreso, precio y P. Sust.	$1272,6942$	$600,80276$	$0,6199929$	$0,110918$	$-26,198188$	$13,351035$	$-16,948923$	$15,982050$

A partir de estos datos, se solicita:

a) ¿Qué modelo seleccionaría y por qué?

b) Escriba el modelo resultante e interprete los resultados del coeficiente de determinación y de los de regresión.

c) ¿Son individualmente significativas las variables del modelo seleccionado para explicar la demanda a un nivel de riesgo del 1%? Conclusiones.

d) ¿Son significativas las variables seleccionadas en forma conjunta para explicar la variación de la demanda a un nivel de riesgo del 1%? Conclusiones.

e) ¿Qué porcentaje de las variaciones de la demanda es explicado por las variaciones de las variables precio y precio del artículo sustituto en forma conjunta?

f) ¿Qué porcentaje de las variaciones no explicadas por el precio, es explicada con la introducción del precio del artículo sustituto?

Apéndice

El presente apéndice explica cómo puede resolverse un problema de regresión lineal mediante la aplicación de la herramienta Python. A tal fin se resolverá el mismo ejercicio que el obrante en el capítulo 6, explicando en forma detallada las instrucciones que deben darse al programa para lograr el objetivo.

En primer término, es importante señalar que en azul figuran las instrucciones y órdenes que deben ser introducidas por el usuario mientras que en rojo aparecen las respuestas que brinda el programa.

Se procede a continuación a mostrar la resolución del ejercicio mencionado mediante la aplicación directa de la herramienta regresión y también, en forma matricial.

1. Aplicación directa de la herramienta regresión

En la línea [44] se definen específicamente las librerías que deben ser importadas para resolver el problema de regresión. Las dos librerías de Python son pandas- denominada *pd*- y *sklearn*. La primera de ellas se utiliza para trabajar con bases de datos, en este caso, sirve para importar a Python los datos del ejercicio. Cada vez que se quiere hacer uso de un comando de esta librería, entonces hay que ingresar *pd*. La librería *sklearn* está compuesta por diversos módulos donde cada uno de ellos permite llegar, a través de comandos preestablecidos, a los valores de estimadores y coeficientes de regresión.

[44]:
```
#Importamos las librerias
import pandas as pd
from sklearn import datasets, linear_model, metrics
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```

En la siguiente línea [45] se importa la base de datos a la que se denomina *dataset*. Luego se la muestra en pantalla, como se ve la respuesta en color naranja. Es importante resaltar que en Python para utilizar una librería, la misma debe ser solicitada antes de que se apliquen cada uno de los comandos.

Si bien en este caso los datos se encuentran en una planilla de Excel, se destaca que puede también trabajarse con archivos del tipo TXT, SQL o cualquier otro formato de base de datos. Cabe resaltar que el archivo correspondiente a la base de datos debe estar en la misma ruta que el archivo del programa Python. Si no estuviera ubicado en la misma ruta será necesario entonces, detallar la ruta de acceso al archivo con el que se va a trabajar.

```
[45]: #Importamos el dataset
      dataset = pd.read_excel('regexcel.xlsx') #recordar que el dataset tiene que.
       ↪estar en el mismo directorio que el código, sino que hay indicar la.
       ↪ubicación.
      dataset
```

[45]:

	Y	X1	X2
0	5.298	0.85	2.4
1	8.100	0.92	3.1
2	4.506	0.97	1.4
3	4.816	1.10	2.3
4	9.768	0.98	4.5
5	6.486	1.15	2.8
6	2.022	1.15	1.4
7	4.676	1.20	1.9
8	5.524	1.20	2.5
9	4.152	1.22	2.6

En las siguientes líneas, desde [47] hasta [50], se genera a través del comando *fit* el modelo lineal de regresión. Este comando se aplica a los problemas de regresión – simple y múltiple. Así, mediante la utilización del comando *fit* en la línea [48], se le indica al programa cuales son las variables explicativas del *dataset* y cuál es la variable explicada.

```
[47]: model = linear_model.LinearRegression()
```

```
[48]: model.fit(dataset[["X1","X2"]],dataset.Y)
```

```
[48]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,.
       ↪normalize=False)
```

[49] y [50] azul solicitan cuales son los coeficientes de regresión a través de los comandos *coef* e *intercept*, respectivamente mientras que [49] y [50] rojo devuelve el vector de los coeficientes que acompañan a cada una de las variables y de la intersección.

```
[49] model.coef_
```

```
[49]: array([-3.83101216,  2.02968968])
```

```
[50] model.intercept_
```

```
[50]: 4.595379745010049
```

El comando que permite hallar el valor del coeficiente de determinación recibe el nombre de *score*, y simplemente hay que indicarle al programa, nuevamente, cuál es la variable que se desea explicar y cuales son las variables explicativas dentro del *dataset* ya mencionado. Así, en [10] azul se da la instrucción y en [10] rojo el programa devuelve el resultado:

```
[10] r2 = model.score(dataset[["X1","X2"]],dataset.Y) #coeficiente de determinacion
     r2
```

```
[10]: 0.870122379426416
```

La instrucción [11] permite obtener el coeficiente de determinación ajustado: *adjr*2. A tal fin, se utiliza directamente una de las fórmulas de cálculo del mencionado coeficiente. La cantidad de datos contenidos en el *dataset* se indica a través de la utilización del comando *len*. También debe indicarse la cantidad de variables explicativas – en este caso, se está trabajando con 2 variables -. [11] en rojo devuelve el resultado. Si bien puede sólo completarse la fórmula es conveniente asignarle un nombre a la expresión de referencia – en este caso *adjr*2 dado que, de este modo, cada vez que se necesite el resultado mencionado sólo basta con indicar dicho nombre.

```
[11]: adjr2 = 1-(1-r2)*(len(dataset)-1)/(len(dataset)-2-1)
      adjr2
```

```
[11]: 0.8330144878339634
```

2. Resolución en forma matricial

En este caso, es necesario importar la librería *numpy*, que es la que permite realizar operaciones con matrices. En la línea [15] azul se importa la librería *numpy* - denominada *np*-. Para poder trabajar con esta librería, la base de datos *dataset* debe estar expresada en forma de matrices.

De este modo en [15] rojo aparecen los valores de la matriz de datos. La primera columna de la mencionada matriz contiene los datos de la variable que se desea explicar mientras que las columnas siguientes muestran los valores correspondientes a las variables explicativas – en este caso, las dos columnas siguientes muestran los valores correspondientes a cada una de las variables mencionadas.

```
[14]: #Trabajarlo matricialmente

[15]: import numpy as np
      DatosMatriz = dataset.as_matrix()
      DatosMatriz
```

```
[15]: array([[5.298, 0.85 , 2.4  ],
             [8.1  , 0.92 , 3.1  ],
             [4.506, 0.97 , 1.4  ],
             [4.816, 1.1  , 2.3  ],
             [9.768, 0.98 , 4.5  ],
             [6.486, 1.15 , 2.8  ],
             [2.022, 1.15 , 1.4  ],
             [4.676, 1.2  , 1.9  ],
             [5.524, 1.2  , 2.5  ],
             [4.152, 1.22 , 2.6  ]])
```

La instrucción [16] solicita que se transponga la matriz de datos que se ha simbolizado como M. En [16] rojo el programa devuelve la matriz traspuesta.

```
[16]: M = np.transpose(DatosMatriz)
      M
```

```
[16]: array([[5.298, 8.1  , 4.506, 4.816, 9.768, 6.486, 2.022, 4.676, 5.524,
              4.152],
             [0.85 , 0.92 , 0.97 , 1.1  , 0.98 , 1.15 , 1.15 , 1.2  , 1.2  ,
              1.22 ],
             [2.4  , 3.1  , 1.4  , 2.3  , 4.5  , 2.8  , 1.4  , 1.9  , 2.5  ,
              2.6  ]])
```

La instrucción [17] permite identificar a cada una de las filas de la matriz M previamente expuesta. A tal fin, para extraer cada una de las filas de la matriz se escribe el nombre de la matriz, en este caso M, y luego entre corchetes la fila que se desea extraer. [17] rojo muestra los resultados que arroja el programa.

[18] azul es la instrucción que solicita los datos de la matriz original mientras que el programa devuelve los mismos en [18] rojo.

```
[17]: Y = M[0] #Variable a explicar
      X1 = M[1] #Primer variable explicativa
      X2 = M[2] #Segunda variable explicativa
      print(X1)
      print(X2)
      print(Y)
```

```
[18]: [0.85 0.92 0.97 1.1  0.98 1.15 1.15 1.2  1.2  1.22]
      [2.4 3.1 1.4 2.3 4.5 2.8 1.4 1.9 2.5 2.6]
      [5.298 8.1   4.506 4.816 9.768 6.486 2.022 4.676 5.524 4.152]
```

```
[18]: DatosMatriz
      [8.1  , 0.92 , 3.1  ],
      [4.506, 0.97 , 1.4  ],
      [4.816, 1.1  , 2.3  ],
      [9.768, 0.98 , 4.5  ],
      [6.486, 1.15 , 2.8  ],
      [2.022, 1.15 , 1.4  ],
      [4.676, 1.2  , 1.9  ],
      [5.524, 1.2  , 2.5  ],
      [4.152, 1.22 , 2.6  ]])
```

La instrucción [19] solicita eliminar de la matriz la primera columna correspondiente a los valores de la variable que se desea explicar Y. Esta matriz se denomina X_sinY .

Luego, la instrucción [20] solicita se agregue una primera columna de unos para construir la matriz X.

```
[19]: X_sinY =  np.delete(DatosMatriz, np.s_[0:1],axis=1) #elimino columa Y
      X_sinY

[19]: array([[0.85, 2.4 ],
             [0.92, 3.1 ],
             [0.97, 1.4 ],
             [1.1 , 2.3 ],
             [0.98, 4.5 ],
             [1.15, 2.8 ],
             [1.15, 1.4 ],
             [1.2 , 1.9 ],
             [1.2 , 2.5 ],
             [1.22, 2.6 ]])

[20]: X = np.c_[np.ones(10), X_sinY] #Agrego columna de 1
      X

[20]: array([[1.  , 0.85, 2.4 ],
             [1.  , 0.92, 3.1 ],
             [1.  , 0.97, 1.4 ],
             [1.  , 1.1 , 2.3 ],
             [1.  , 0.98, 4.5 ],
             [1.  , 1.15, 2.8 ],
             [1.  , 1.15, 1.4 ],
             [1.  , 1.2 , 1.9 ],
             [1.  , 1.2 , 2.5 ],
             [1.  , 1.22, 2.6 ]])
```

La instrucción [21] permite transponer la matriz anterior a través de la utilización del comando *transpose* del paquete *numpy*. Luego la instrucción [22] permite realizar el producto entre la matriz X - [20]y su transpuesta – [21] -. La matriz resultante se denomina X^tX.

```
[21]: Xt = np.transpose(X) #X traspuesta
      Xt

[21]: array([[1.  , 1.  , 1.  , 1.  , 1.  , 1.  , 1.  , 1.  , 1.  , 1.  ],
             [0.85, 0.92, 0.97, 1.1 , 0.98, 1.15, 1.15, 1.2 , 1.2 , 1.22],
             [2.4 , 3.1 , 1.4 , 2.3 , 4.5 , 2.8 , 1.4 , 1.9 , 2.5 , 2.6 ]])

[22]: XtX = np.dot(Xt,X) #X traspuesta por X
      XtX

[22]: array([[10.    , 10.74  , 24.9  ],
             [10.74  , 11.6936, 26.472],
             [24.9   , 26.472 , 69.29 ]])
```

La instrucción [23] con la ayuda del paquete *linalg* de *numpy*, solicita obtener la inversa de la matriz XtX, a la cual se denomina $InvXtX$.

```
[23]: InvXtX = np.linalg.inv(XtX)
      InvXtX

[23]: array([[10.09466693, -7.83928677, -0.63263973],
             [-7.83928677,  6.72069532,  0.24950201],
             [-0.63263973,  0.24950201,  0.14645565]])
```

A fin de poder construir el producto entre la matriz traspuesta de X – Xt – y el vector Y. La instrucción [24] solicita los datos correspondientes a este último vector. Luego, la instrucción [25] solicita que se realice el mencionado producto que se identifica como $XtPorY$.

```
[24]: Y = DatosMatriz[:,0]
      Y

[24]: array([5.298, 8.1  , 4.506, 4.816, 9.768, 6.486, 2.022, 4.676, 5.524,
             4.152])

[25]: XtPorY = np.dot(Xt,Y)
      XtPorY

[25]: array([ 55.348 ,  58.286 , 153.6476])
```

Finalmente, para hallar el valor de los estimadores, se realiza el producto entre la matriz inversa – [23] rojo – y el producto matricial obtenido en [25] rojo. La instrucción que solicita lo previamente expuesto es la [26].

```
[26]: Estimadores = np.dot(InvXtX,XtPorY)
      Estimadores

[26]: array([ 4.59537975, -3.83101216,  2.02968968])
```