Apuntes de Probabilidad y Estadística para Ingeniería y Administración

Ignacio Vélez Pareja
Decano
Facultad de Ingeniería Industrial
Politécnico Grancolombiano
Bogotá, Colombia
17 de octubre de 2003

Análisis de regresión

> *I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*
> Todavía no tengo datos. Es un error grave teorizar antes de tener la información. Sin darse cuenta, uno empieza a acomodar los hechos a las teorías en lugar de ajustar la teoría a los hechos.
> **Sherlock Holmes**

Con el análisis de regresión se busca encontrar modelos que a partir de las relaciones causales entre una variable dependiente (la que se supone que es el resultado de la influencia o comportamiento de otras variables) y una o más variables independientes, permitan predecir un resultado conociendo el valor estimado de una variable independiente.

Antes de realizar cualquier análisis se debe examinar si existe una relación lógica entre las variables independientes y la variable dependiente. Este esfuerzo es el más importante. Lo relacionado con los cálculos es muy fácil porque existen herramientas computacionales para hacerlos. Encontrar las posibles relaciones lógicas entre las variables es un trabajo de observación, inteligencia, experiencia e intuición.

Ajuste de una línea recta a datos observados

Examinemos por ejemplo dos variables: tasa de inflación y tasa de aumento del salario mínimo. ¿Cree usted que hay relación entre ellas? ¿La tasa de inflación dependerá del aumento del salario mínimo o viceversa? Para responder este tipo de preguntas se debe conocer cómo ocurren ambos fenómenos. Así mismo, si los fenómenos son actos de Dios o de la naturaleza o de muchísimas variables y circunstancias o son producto de decisiones tomadas por seres humanos de manera consciente y deliberada.

La inflación es el cambio porcentual que sufre un indicador de precios que se conoce como el Índice de Precios al Consumidor (IPC) y mide el cambio de precio de una canasta de bienes que consumen los hogares. Esto quiere decir que hay múltiples componentes en esa canasta de bienes y que la decisión en el cambio de precios de sus componentes no es producto de una decisión de una sola persona sino de miles de personas.

Por otro lado, el aumento en el salario mínimo es una decisión que toma un pequeño grupo que negocia ese valor o en el peor de los casos es una decisión de una persona (un ministro o un presidente) con base en el análisis de variables macroeconómicas tales como la inflación.

Observemos una serie de valores para cada una de estas variables. Esto se puede observar en la siguiente tabla.

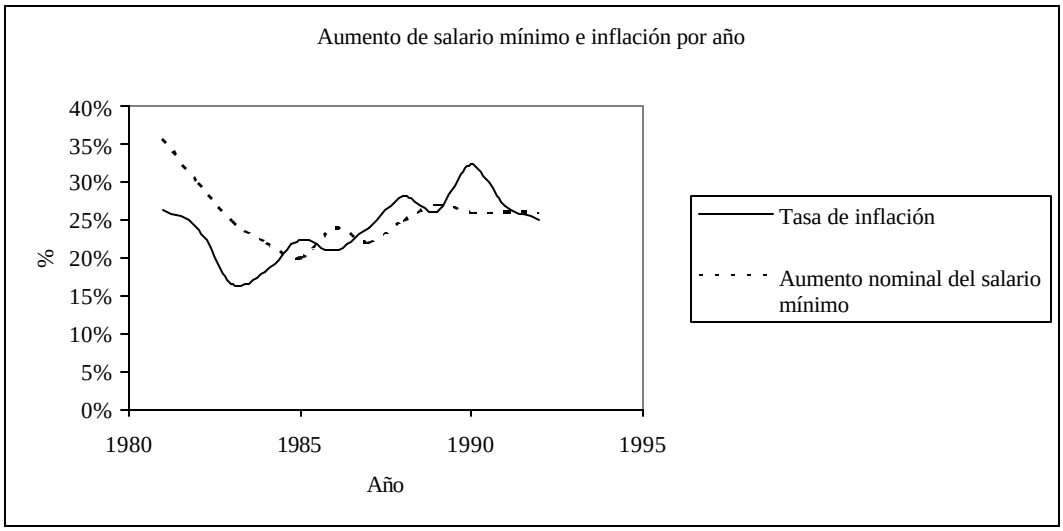
Tabla 1. Tasas de inflación y aumento del salario mínimo

Año	Tasa de inflación	Aumento nominal del salario mínimo
1981	26,35%	35,71%
1982	24,03%	30,00%
1983	16,64%	24,98%
1984	18,28%	22,00%
1985	22,45%	20,00%
1986	20,95%	24,00%
1987	24,02%	22,00%
1988	28,12%	25,00%
1989	26,12%	27,00%
1990	32,37%	26,00%
1991	26,82%	26,07%
1992	25,14%	26,04%
1993	22,61%	25,03%
1994	22,60%	21,09%
1995	19,47%	20,50%
1996	21,64%	19,50%
1997	17,68%	21,02%
1998	16,70%	18,50%
1999	9,23%	16,01%
2000	8,75%	10,00%
2001	7,65%	9,96%
2002	6,00% [1]	8,04%

[1] Estimada en enero de 2002.

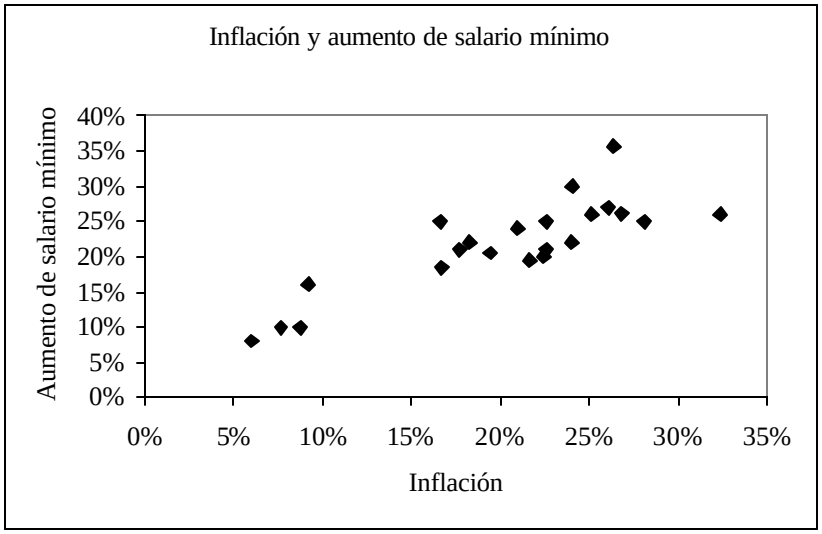
¿Se puede observar alguna relación entre las dos variables? En este caso en que analizamos dos variables esa posible relación se puede observar mejor por medio de una gráfica.

Figura 1. Tasa de inflación y aumento del salario mínimo por año



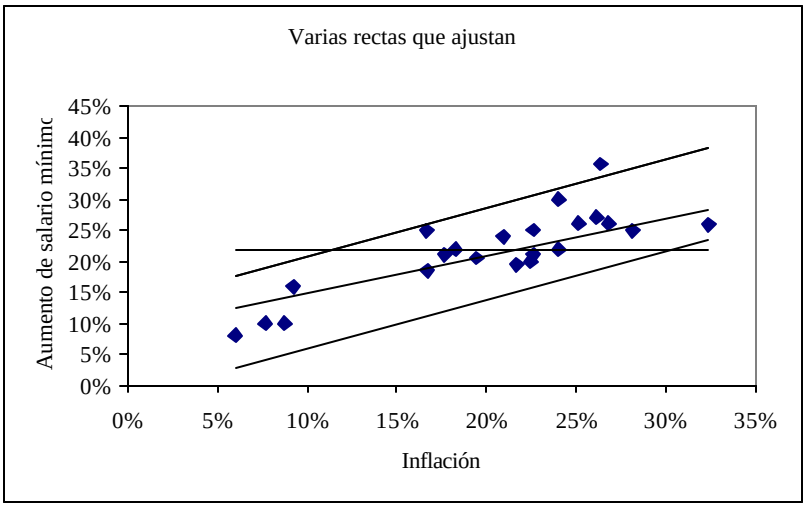
Más aun, si comparamos inflación contra aumento en salario mínimo, vemos de otra forma esa relación.

Figura 2. Tasa de inflación y aumento del salario mínimo



La pregunta que nos hacemos ahora es si esa influencia existe ¿podremos imaginarnos una relación matemática entre esos valores tal y como se muestran en la figura 2? Imaginemos que esa relación o tendencia se puede representar con una línea recta. Habrá muchas líneas rectas que "a ojo" nos parecen aceptables, por ejemplo, tal y como se muestra en la siguiente figura.

Figura 3. Varias rectas que ajustan los datos



El lector quedará más satisfecho con unas que con otras y hasta este momento la selección se haría por gusto. Tenemos que encontrar una forma "objetiva" con base en un criterio definido y preciso que nos permita encontrar cuál es la recta que mejor se ajusta a los datos. El lector con seguridad tendría muy claro que la recta inferior en esa gráfica no sería adecuada. Inclusive la superior le puede parecer inapropiada. La duda surge de las rectas intermedias (y de una cantidad infinita de posibilidades que habría con diferentes inclinaciones de las rectas).

Un criterio que se puede examinar con la intuición sería el de escoger una recta que fuera equidistante de alguna manera de todos los puntos. O que por ejemplo, la suma de las diferencias entre los puntos reales y la recta sea mínima o sea cero. La recta horizontal, que

es el promedio de los aumentos de salario mínimo cumple esta última condición. ¿El lector quedaría satisfecho con esa línea como la que señala la relación entre la inflación y el aumento de salario mínimo? Con seguridad no. El lector podrá verificar que la suma de las diferencias entre el promedio del aumento y cada aumento es cero.

Por último podemos pensar que la línea que refleje la relación entre las dos variables minimice la suma de los cuadrados de las diferencias (que en el párrafo anterior veíamos que se cancelaban entre sí). Esta línea se conoce como la recta de mínimos cuadrados. Los cuadrados de las diferencias serán siempre positivos porque una cifra negativa o positiva elevada al cuadrado será siempre positiva. Es fácil imaginar que la recta que está más arriba en la gráfica o la que está más abajo o la horizontal (que es el promedio) no cumplen con esta condición.

En cualquier caso nuestra recta se puede representar con la siguiente ecuación

$$Y_{est} = \mathbf{a} + \mathbf{b}X \qquad (1)$$

Donde Y_{est} es el valor de la variable dependiente, X el valor de la variable independiente observado, **a** es la pendiente de la línea y **b** es la constante que muestra el punto de corte con el eje de las coordenadas.

El modelo que represente el comportamiento de los datos será

$$Y_{obser} = \mathbf{a} + \mathbf{b}X + \varepsilon \qquad (2)$$

donde ε representa el error, o sea la diferencia entre el valor que toma la variable dependiente en la realidad y el valor que hemos pronosticado con nuestra recta.

Entonces lo que debemos minimizar es ε^2 y esto es igual a

$$(Y_{obser} - Y_{est})^2 = (Y_{obser} - bX - a)^2 \qquad (3)$$

En realidad lo que debemos encontrar es los valores de **a** y **b** que hacen que el valor de la anterior expresión sea mínimo. Esto se puede lograr hallando la derivada del cuadrado de la diferencia con respecto a **a** y a **b**.

$$\Sigma(Y_{obser} - \mathbf{b}X - \mathbf{a})^2$$

$$= \Sigma(Y_{obser}^2 + b^2X^2 + \mathbf{a}^2 - 2Y_{obser}\mathbf{b}X - 2Y_{obser}\mathbf{a} + 2\mathbf{ba}X) \quad (5)$$

Al derivar con respecto de **a** y haciendo el resultado igual a cero (para hallar el mínimo) se obtiene

$$\Sigma(2\mathbf{a} - 2Y_{obser} + 2\mathbf{b}X) = 0 \quad (6)$$

$$\Sigma(\mathbf{a} - Y_{obser} + \mathbf{b}X) = 0 \quad (7)$$

$$\Sigma\mathbf{a} - \Sigma Y_{obser} + \Sigma\mathbf{b}X = 0 \quad (8)$$

$$n\mathbf{a} - \Sigma Y_{obser} + \mathbf{b}\Sigma X = 0 \quad (9)$$

$$a = \frac{\sum Y_{obser} - b\sum X}{n} = \overline{Y} - b\overline{X} \quad (10)$$

De igual manera derivando con respecto a **b** y haciendo la derivada igual a cero se tiene,

$$\Sigma(2\mathbf{b}X^2 - 2Y_{obser}X + 2\mathbf{a}X) = 0 \quad (11)$$

dividiendo por 2,

$$\Sigma(\mathbf{b}X^2 - Y_{obser}X + \mathbf{a}X) = 0 \quad (12)$$

$$\Sigma\mathbf{b}X^2 - \Sigma Y_{obser}X + \Sigma\mathbf{a}X = 0 \quad (13)$$

$$\mathbf{b}\Sigma X^2 - \Sigma Y_{obser}X + \mathbf{a}\Sigma X = 0 \quad (14)$$

Despejando **a**

$$\mathbf{a}\Sigma X = \Sigma Y_{obser}X - \mathbf{b}\Sigma X^2 \quad (15)$$

$$\mathrm{a} = \frac{\sum Y_{obser} X \text{ - b} \sum \mathrm{X}^2}{\sum X} \tag{16}$$

Reemplazando **a** (16) en la derivada con respecto a **a** en (10), se tiene

$$a = \frac{\sum Y_{obser} - b \sum X}{n} = \overline{Y} - b\overline{X} \tag{17}$$

$$\frac{\sum Y_{obser} X \text{ - b} \sum \mathrm{X}^2}{\sum X} = \frac{\sum Y_{obser} - b \sum X}{n} \tag{18}$$

De esta expresión despejamos **b**

$$\frac{\sum Y_{obser} X - b \sum \mathrm{X}^2}{\sum X} = \frac{\sum Y_{obser} - b \sum X}{n} \tag{19}$$

$$\text{- b} \sum \mathrm{X}^2 + \sum Y_{obser} X = \frac{\sum X \sum Y_{obser}}{n} - \frac{b}{n}\left(\sum X\right)^2 \tag{20}$$

$$\text{- b} \sum \mathrm{X}^2 + \frac{b}{n}\left(\sum X\right)^2 = \frac{\sum X \sum Y_{obser}}{n} - \sum Y_{obser} X \tag{21}$$

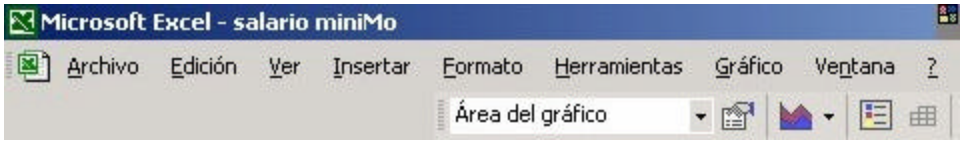
$$b = \frac{\frac{\sum X \sum Y_{obser}}{n} - \sum Y_{obser} X}{\sum \mathrm{X}^2 - \frac{\left(\sum X\right)^2}{n}} \tag{22}$$

$$b = \frac{\sum X \sum Y_{obser} - n \sum Y_{obser} X}{n \sum \mathrm{X}^2 - \left(\sum X\right)^2} \tag{23}$$

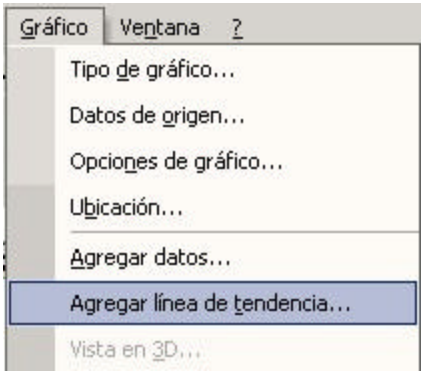
$$b = \frac{n \sum Y_{obser} X - \sum X \sum Y_{obser}}{\left(\sum X\right)^2 - n \sum \mathrm{X}^2} \tag{24}$$

Afortunadamente con la disponibilidad de hojas de cálculo estas fórmulas tan aparatosas no se requieren.

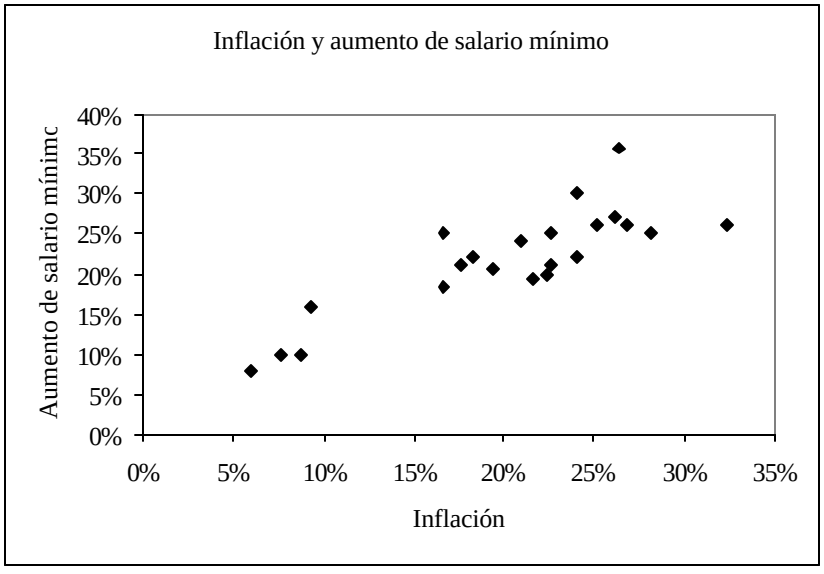
Excel nos permite calcular **a** y **b** de varias maneras. Aquí presentamos las más notables. La primera y más sencilla es desde la gráfica misma de los datos. Cuando se activa (se hace clic) la gráfica el Menú de texto cambia y aparece una nueva opción que se llama Gráfico, así



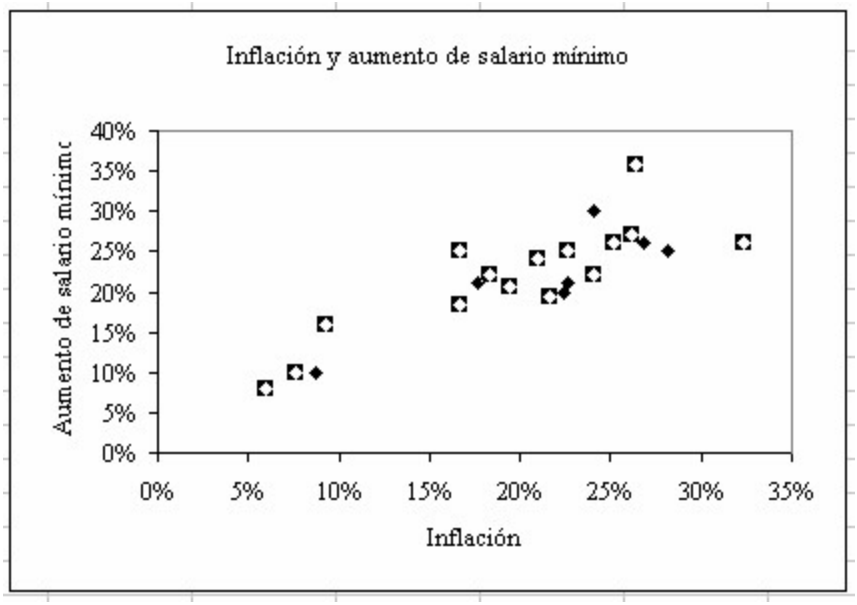
Cuando se activa esa nueva opción aparece el menú desplegado



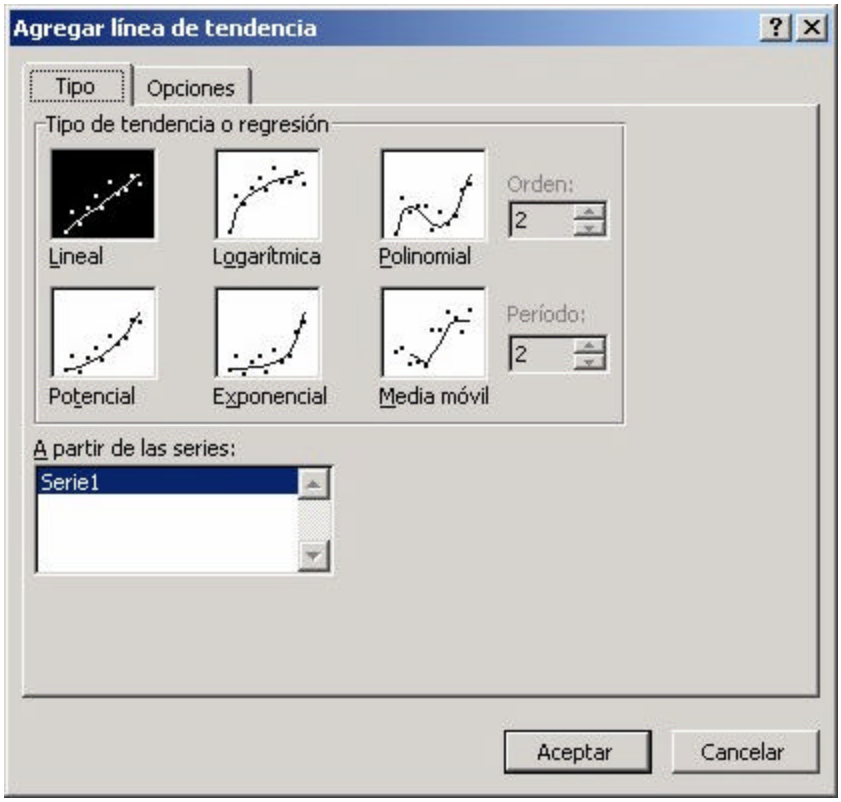
Nuestros datos aparecen como puntos en la gráfica así



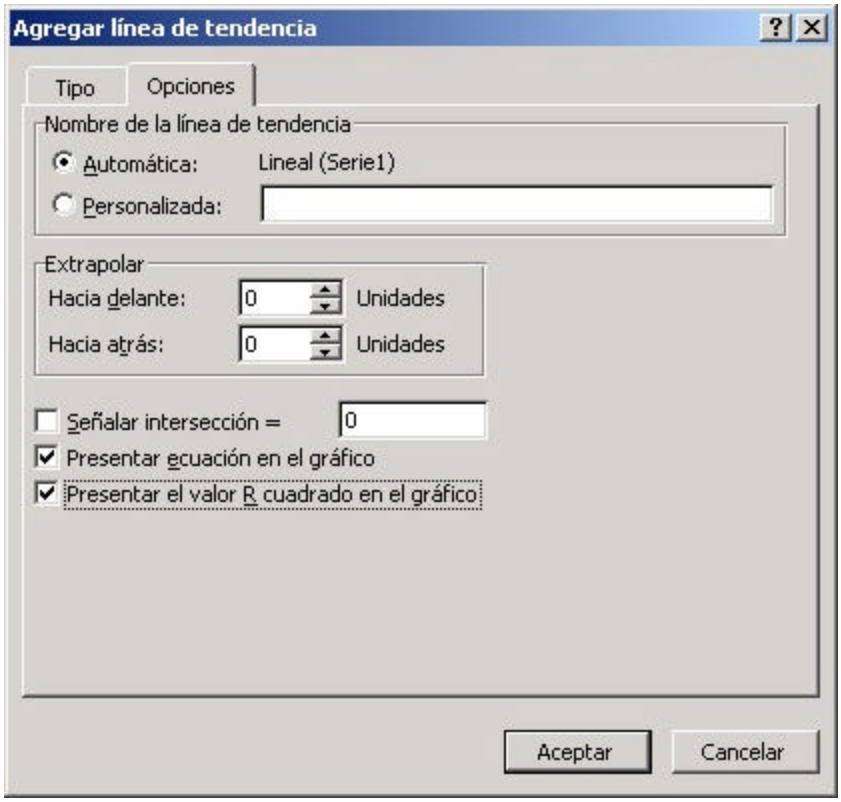
Si se activan los puntos haciendo clic sobre ellos entonces se puede solicitar que el programa añada una línea de tendencia.



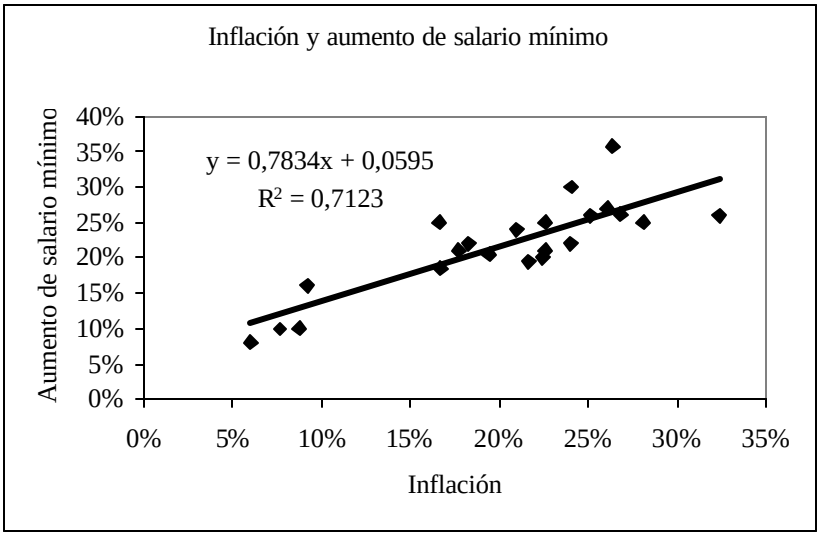
Al seleccionar la opción Agregar línea de tendencia se obtiene lo siguiente



Allí se puede seleccionar el Tipo de línea que se desea (depende del comportamiento de los datos) y en Opciones aparece lo siguiente (ya con Presentar ecuación y R^2 señalados por el usuario)



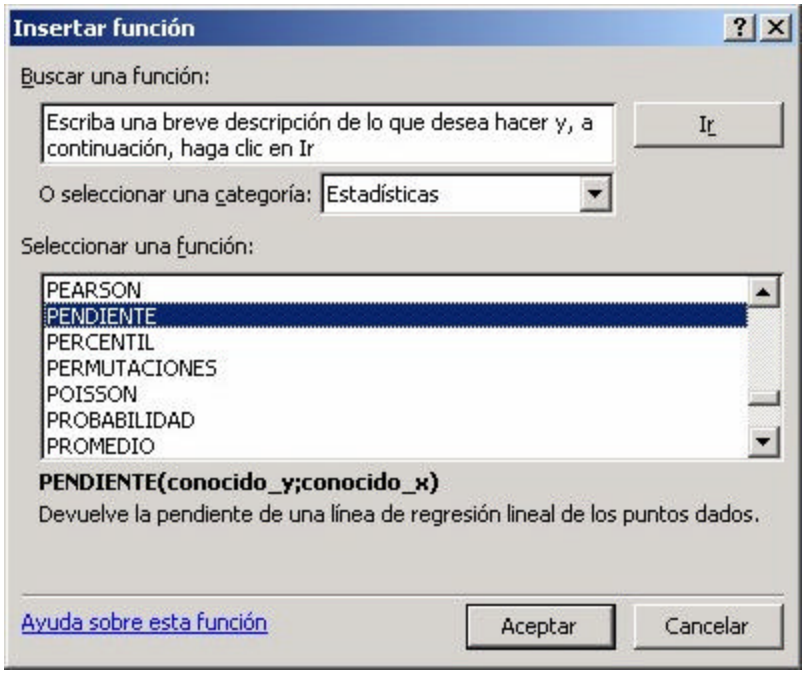
Cuando se oprime Aceptar se obtiene lo siguiente



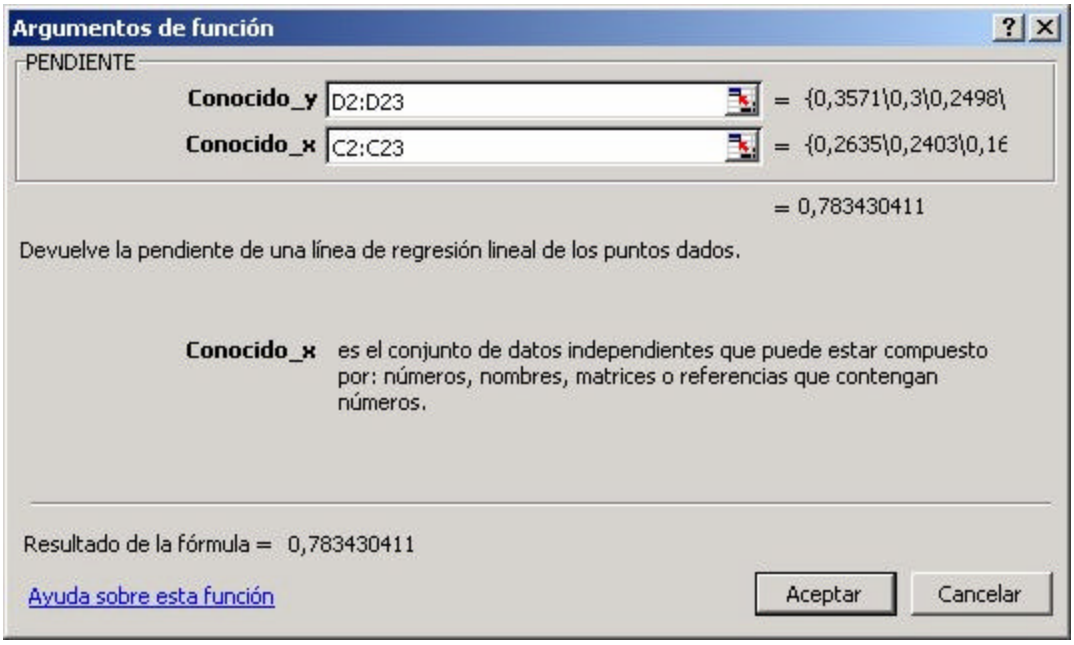
En este ejemplo a = 0,0595 y b = 0,7834. El lector puede verificar estos resultados utilizando la fórmula deducida arriba para cada coeficiente. La recta Y = 0,0595 + 0,7834 × (inflación) es la recta de mínimos cuadrados. De este modo, si se utiliza este modelo para

pronosticar el aumento de salario mínimo basados en la inflación, entonces se diría que para pronosticar el aumento del salario mínimo se toma el 78,34% de la inflación y se le añade 5,95%.

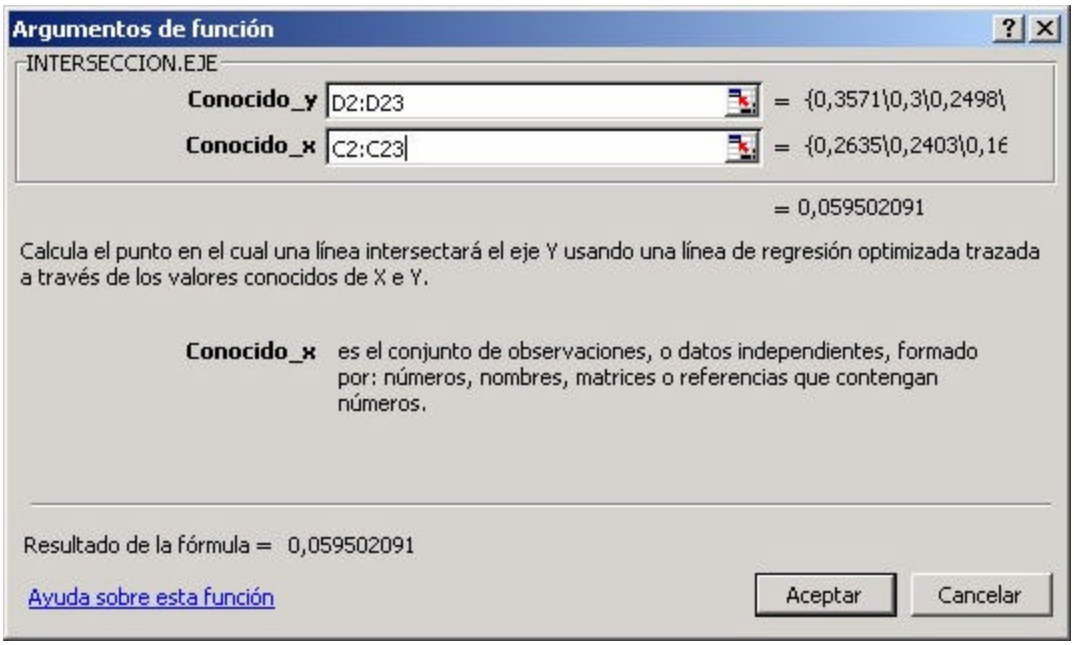
Hay otras formas de llegar a los mismos resultados. Por ejemplo, si se usan las funciones estadísticas se encuentra lo siguiente



La función Pendiente calcula el valor de **b** cuando se introducen los datos para las variables dependientes e independientes.



El resultado aparece debajo a la derecha de la caja para Conocido_X y es 0,783430411. Compare el resultado con el obtenido con la gráfica. La diferencia es el número de decimales. Con la función Intersección.eje se calcula de la misma manera, el valor de **a**.



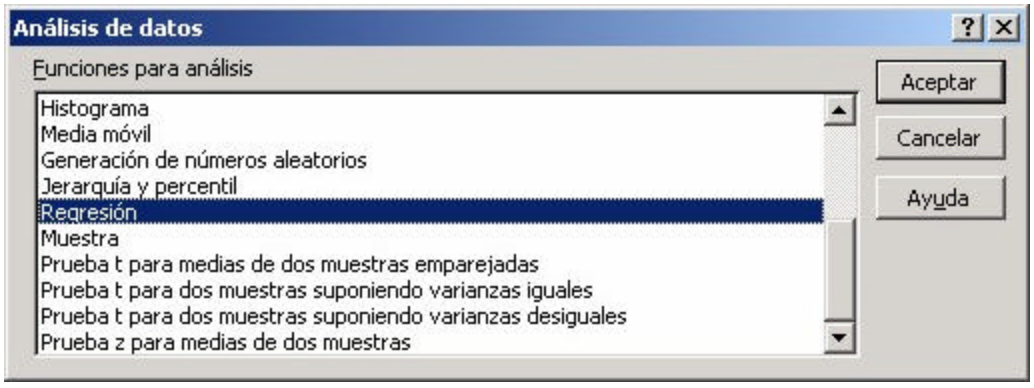
En este ejemplo, el valor de **a** es 0,059502091. Otra vez, la diferencia se debe al número de decimales.

El pronóstico basado en la regresión lineal se puede hacer también usando las fórmulas de Excel. En este caso se utiliza la función Tendencia. Esta función arroja los resultados de aplicar la ecuación de la recta de mínimos cuadrados a una serie de nuevos valores para la variable independiente (en el ejemplo, la inflación). Para este caso se deja al lector el ejercicio de usar esta función.

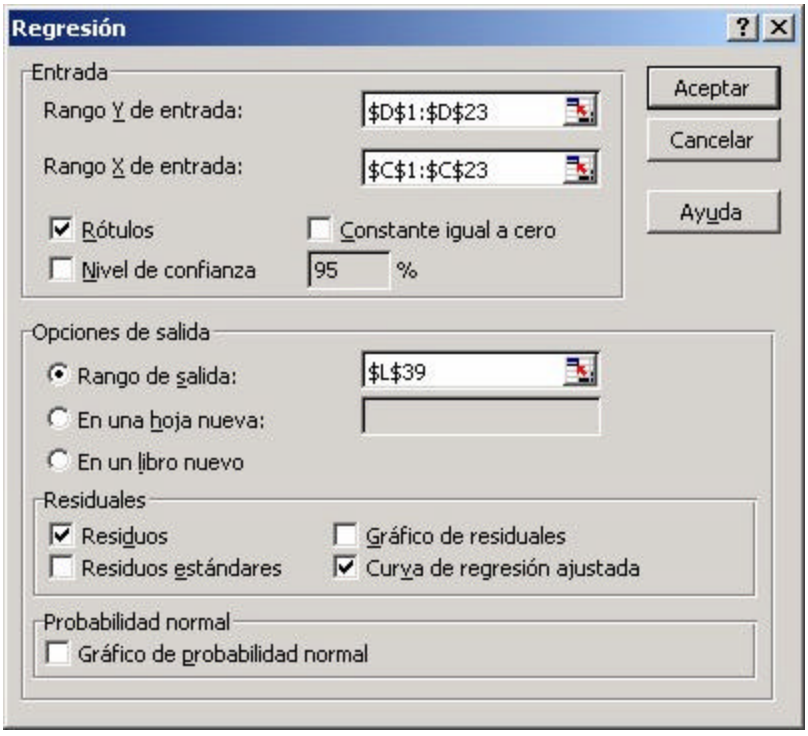
Análisis de varianza

Hay otra posibilidad que nos brinda Excel y además proporciona mucha más información acerca del modelo. Se trata de una herramienta para regresión que se encuentra en la opción Herramientas y allí en Análisis de datos.

Cuando se selecciona, aparece este cuadro de diálogo.



Al escoger Regresión aparece lo siguiente



Excel arroja los siguientes resultados

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,843963616
Coeficiente de determinación R^2	0,712274586
R^2 ajustado	0,697888315
Error típico	0,036018272
Observaciones	22

Análisis de varianza (ANOVA)

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	1	0,064231041	0,064231041	49,51071754	7,97608E-07	
Residuos	20	0,025946318	0,001297316			
Total	21	0,090177359				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	0,023728139	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

La tabla anterior se conoce como tabla de Análisis de varianza (o ANOVA por su nombre en inglés **An**alysis **o**f **Va**riance) y suministra información muy valiosa en relación con el modelo. Por el momento, el lector puede observar los coeficientes **a** y **b** obtenidos arriba. Intercepción, o sea a es 0,059502091 y la pendiente de la recta o coeficiente de Tasa de inflación, o sea b, es 0,783430411.

También arroja los siguientes valores conocidos como Análisis de los residuales.

Observación	*Pronóstico Aumento nominal del salario mínimo*	*Residuos*
1	26,59%	9,12%
2	24,78%	5,22%
3	18,99%	5,99%
4	20,27%	1,73%
5	23,54%	-3,54%
6	22,36%	1,64%
7	24,77%	-2,77%
8	27,98%	-2,98%
9	26,41%	0,59%
10	31,31%	-5,31%
11	26,96%	-0,89%
12	25,65%	0,39%
13	23,66%	1,37%
14	23,66%	-2,57%
15	21,20%	-0,70%
16	22,90%	-3,40%
17	19,80%	1,22%
18	19,03%	-0,53%
19	13,18%	2,83%
20	12,81%	-2,81%
21	11,94%	-1,98%
22	10,65%	-2,61%

Esta tabla indica el valor del aumento del salario mínimo si se hubiera comportado exactamente como indica el modelo. Así mismo, muestra los residuos, o sea, como vimos arriba, la diferencia entre el valor real que ocurrió y el valor calculado por el modelo.

También arroja la gráfica que hemos mostrado arriba con la línea de ajuste. No se reproduce aquí por razones obvias. (Puede producir otros informes y gráficas a solicitud del usuario).

La desventaja de esta opción radica en que los valores calculados en las tablas son números y no fórmulas. Es decir que si se hace un cambio en los datos es necesario repetir toda la operación. Por otro lado, la ventaja radica en que ofrece los resultados en una forma tabular bien organizada y usada comúnmente.

Varios de los datos que produce esta opción Análisis de datos también los produce la función Estimación.lineal. La tabla que se produce con esta función arroja los siguientes resultados (no se muestra el procedimiento de inclusión de los datos en la función):

Pendiente o coeficiente de la variable independiente (**b**) 0.783430411	Intercepción (**a**) 0.059502091
Error típico de **b** 0.111339897	Error típico de **a** 0.023728139
Coeficiente de determinación, R^2 0.712274586	Error típico 0.036018272
Valor de F 49.51071754	Grados de libertad 20
Suma de los cuadrados de la regresión 0.064231041	Suma de los cuadrados de los residuos 0.025946318

Al contrario de la opción Regresión de Análisis de datos, con esta función se obtienen los mismos datos básicos (como el lector habrá observado). Para obtener toda la información que aparece en las tablas de la opción Regresión ya mencionada, es necesario hacer algunas operaciones.

Coeficiente de correlación o de Pearson

Este indicador nos muestra qué tan relacionadas están dos variables. Está estrechamente ligado con la covarianza, ya estudiada. Este coeficiente de correlación se puede utilizar, por ejemplo, para determinar la relación entre dos variables, en nuestro ejemplo, entre la tasa de inflación y el aumento del salario mínimo.

En la tabla de arriba lo encontramos en

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,843963616**
Coeficiente de determinación R^2	0,712274586
R^2 ajustado	0,697888315
Error típico	0,036018272
Observaciones	22

La ecuación para el coeficiente de correlación es en general:

$$r = \frac{Cov(X, Y)}{s_x s_Y}$$

En nuestro ejemplo la expresión es

$$r = \frac{Cov(X, Y_{obser})}{s_x s_{Y_{obser}}}$$

donde $Cov(X,Y_{obser})$ es la covarianza entre las dos variables, y σ^2 es la varianza de las variables.

$$-1 \leq r \leq 1$$

y:

$$\text{Cov}(X, Y) = \frac{1}{n}\sum (X - \overline{X})(Y - \overline{Y})$$

El signo del coeficiente de correlación indica el sentido de la relación de la misma manera que la covarianza nos indica con su signo si la variación de las dos variables es en la misma dirección o en sentido contrario. Un valor negativo indica que si la variable independiente aumenta, la dependiente baja y viceversa. Mientras más cercano a 1 esté su valor absoluto, más relación podremos suponer que existe entre las variables.

Coeficiente de determinación, R^2 o medición de la bondad de ajuste

Podemos distinguir algunas de las diferencias entre los valores observados, los pronosticados con el modelo y los errores ya mencionados.

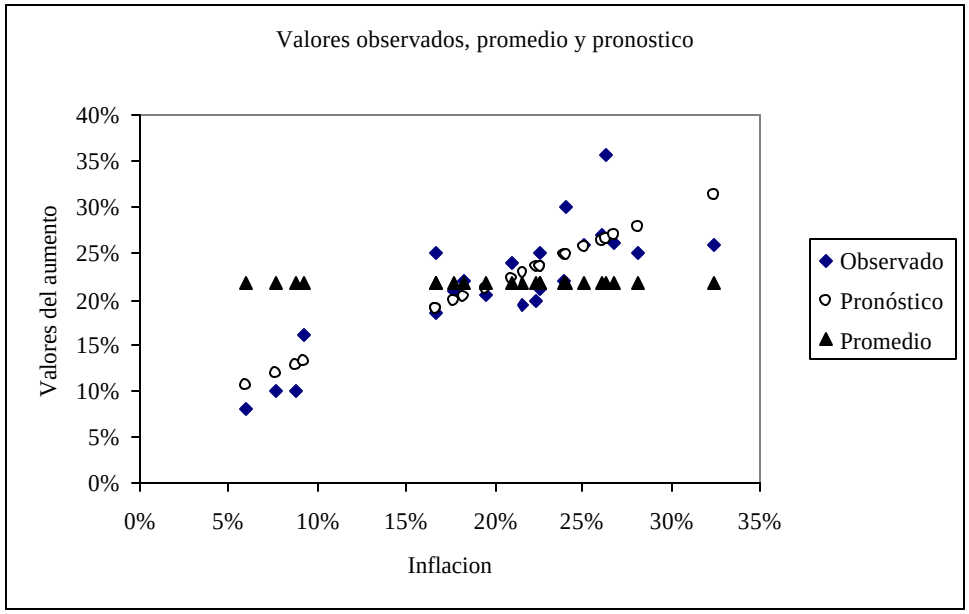
Vamos entonces a distinguir los siguientes

1. La suma total de los cuadrados STC, (en la tabla de Análisis de varianza que produce Excel se titula como Total, está ubicada bajo la columna Suma de cuadrados y vale 0,0901773586363636) es la diferencia entre el promedio de los valores observados y cada uno de esos valores elevada al cuadrado, o sea, $\text{STC} = \text{Total} = \sum (Y_{obser} - \overline{Y_{obser}})^2$. Este valor mide la variación total de la muestra que tenemos. O si se prefiere, qué tan dispersos están los valores Y_{obser} dentro de la muestra.
2. La suma de los cuadrados de la diferencia entre cada valor estimado por el modelo y el promedio de las Y_{obser}. Esta se denomina la suma explicada de los cuadrados SEC, (en la tabla de Análisis de varianza que produce Excel se titula como Regresión, está ubicada bajo la columna Suma de cuadrados y vale 0,064231041) y tiene sentido porque es la variación que se puede

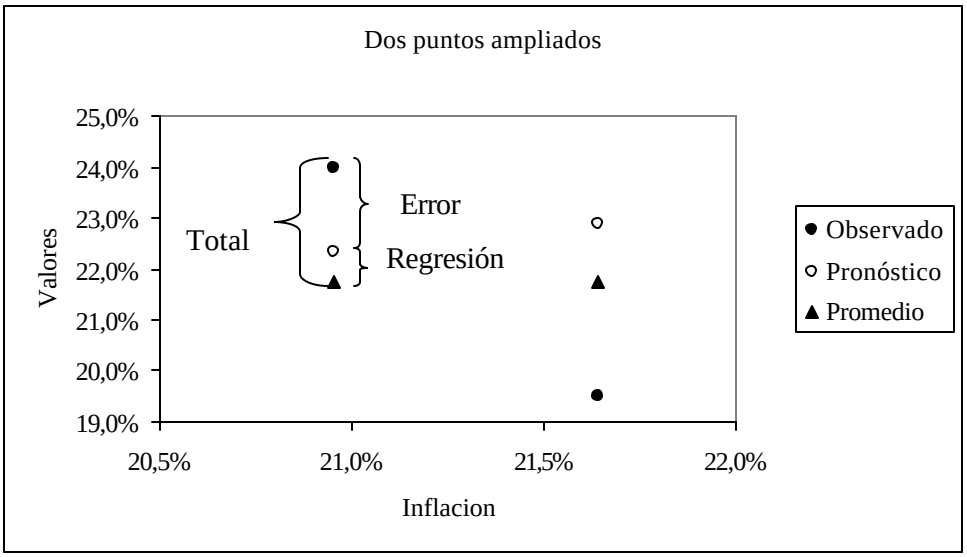
asociar a los datos basados en el modelo, es decir, $\text{SEC} = \text{Regresión} = \sum \left(Y_{est} - \overline{Y_{obser}}\right)^2$. Mide la dispersión de los valores estimados por el modelo alrededor de la media de los valores observados.

3. Una tercera es la suma del cuadrado de los residuos SCR, que es el cuadrado de la diferencia entre el valor observado y el valor calculado por el modelo SCR, (en la tabla de Análisis de varianza que produce Excel se titula como Residuos, está ubicada bajo la columna Suma de cuadrados y vale 0,025946318) es decir $\text{SCR} = \text{Residuales} = \sum \left(Y_{obser} - Y_{est}\right)^2$. Estos valores aparecen en la tabla de arriba que llamamos Análisis de los residuales.

Para entender la idea de las diferencias miremos la siguiente gráfica con valores observados, el promedio y el pronóstico



Si ampliamos esta gráfica podemos observar a qué se refiere cada una de las diferencias.



Los dos puntos ampliados corresponden a los siguientes

Inflación	Observado	Pronóstico	Promedio	Total	Regresión	Error
20.95%	24.00%	22.36%	21.75%	2.25%	0.62%	1.64%
21.64%	19.50%	22.90%	21.75%	-2.25%	1.16%	-3.40%

Visualmente y en la tabla anterior se puede comprobar que se cumple lo siguiente

Total = Regresión + Error (25)

Con los datos de la tabla Análisis de los residuales podemos comprobar la siguiente relación

STC = SEC + SCR (26)

Redondeando STC

0,090177359 = 0,064231041 + 0,025946318

Estos valores aparecen en la tabla Análisis de varianza.

	Grados de libertad	**Suma de cuadrados**	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	1	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	20	**0,025946318**	0,001297316			
Total	21	**0,090177359**				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	0,023728139	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

Esta relación es importante porque nos indica qué fracción de la variación total STC, se explica con el modelo y que parte no se puede explicar (los errores ε que habíamos definido arriba). Fácilmente podemos calcular la fracción correspondiente a cada uno de los componentes de la variación total. Ahora podemos responder la pregunta de qué tanto explica el modelo propuesto (SEC) la variación total y qué tanto no se puede explicar (SCR).

De la tabla anterior tenemos

	Suma de cuadrados
Regresión	0,064231041
Residuos	0,025946318
Total	0,090177359

Podemos expresar esos valores como un porcentaje del total, así

	Suma de cuadrados	Fracción	Valor porcentual
Regresión	0,064231041	0,7122746	71,23%
Residuos	0,025946318	0,2877254	28,77%
Total	0,090177359	1,0000000	100,00%

Observemos que el valor (redondeado) 0,7122746 lo encontramos en la tabla de arriba como Coeficiente de determinación R^2.

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,843963616**
Coeficiente de determinación R^2	**0,712274586**
R^2 ajustado	0,697888315
Error típico	0,036018272
Observaciones	22

Esto quiere decir que el R^2 es exactamente SEC/STC. Este valor estará siempre entre 0 y 1 simplemente porque SEC nunca podrá ser mayor que STC (SEC es un componente de STC).

Entonces se dice que R^2 es el porcentaje de variación de la variable dependiente que estaría explicado por la variable independiente en el modelo de regresión lineal. Si todos los puntos observados estuvieran en la línea de regresión, R^2 sería igual a 1. Esto quiere decir que hay un ajuste perfecto. Por lo tanto, un R^2 cercano a 1 indica buen ajuste y un R^2 cercano a cero indica un mal ajuste. Entonces R^2 mide la bondad del ajuste.

En nuestro ejemplo, R^2 es 0,712274586 lo cual significa que el 71,23% de la variación del aumento del salario mínimo se explica por la inflación.

Observe también que el Coeficiente de determinación R^2 es el cuadrado del Coeficiente de correlación múltiple.

Coeficiente de determinación, R^2 ajustado

Cuando definimos R^2 hicimos lo siguiente: partimos de la ecuación (26) y encontramos la proporción de SEC sobre el total. Es decir, dividimos (26) por STC

$$STC = SEC + SCR \qquad (26)$$

$$1 = \frac{SEC}{STC} + \frac{SCR}{STC} \qquad (27)$$

Al despejar SEC/STC encontramos

$$R^2 = \frac{SEC}{STC} = 1 - \frac{SCR}{STC} \qquad (28)$$

Esta ecuación (28) se puede escribir como

$$R^2 = \frac{SEC}{STC} = 1 - \frac{SCR/n}{STC/n} \qquad (29)$$

En (28) estamos definiendo R^2 como el complemento del error y en (29) hemos dividido ambos elementos del quebrado por n. Pero sabemos que por definición la varianza es la suma de los cuadrados de las diferencias con la media dividida por n, es decir que SCR/n y STC/n son la varianza de los residuos y la varianza total. Sin embargo, por razones que no están al alcance de estos apuntes, esas varianzas no son las verdaderas porque son lo que se llaman en estadística, estimadores sesgados. Para obtener la varianza no sesgada o insesgada, hay que dividir no por n, sino por el número de grados de libertad de cada elemento. Los grados de libertad se calculan para los residuos como (n - k - 1) donde n es el número de observaciones en la muestra y k es el número de variables independientes que para las cuales se desea estimar el coeficiente; en el caso de la STC los grados de libertad son (n - 1). Estos grados de libertad están en la tabla y son respectivamente 20 y 21. Entonces al usar los grados de libertad en (29) obtenemos el verdadero valor de R^2 es decir, el R^2 ajustado.

$$R^2_{2ajust} = 1 - \frac{SCR/\text{grados de libertad de SCR}}{STC/\text{grados de libertad de STC}} \qquad (30)$$

En nuestro ejemplo tenemos

	Grados de libertad	**Suma de cuadrados**	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	0,001297316			
Total	**21**	**0,090177359**				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	0,023728139	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

$$R^2_{2ajust} = 1 - \frac{0,025946318/20}{0,090177359/21} = 0,697888315$$

Este es el valor que aparece en la tabla de Análisis de Varianza.

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,843963616**
Coeficiente de determinación R^2	**0,712274586**
R^2 ajustado	**0,697888315**
Error típico	0,036018272
Observaciones	22

El error típico de los residuos (0,036018272) se obtiene como la raíz cuadrada de la suma de los cuadrados de los residuos y el número de grados de libertad de los mismos.

	Grados de libertad	**Suma de cuadrados**	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	0,023728139	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

$$\text{Error típico de los residuos} = \sqrt{\frac{\text{Suma de los cuadrados de los residuos(SCR)}}{\text{Grados de libertad}}}$$

$$= \sqrt{\text{Promedio de los cuadrados de los residuos}}$$

En nuestro ejemplo

$$\text{Error típico de los residuos} = \sqrt{\frac{0{,}025946318}{20}} = \sqrt{0{,}001297316} = 0{,}036018272$$

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,843963616**
Coeficiente de determinación R^2	**0,712274586**
R^2 ajustado	**0,697888315**
Error típico	**0,036018272**
Observaciones	22

La importancia de esta formulación es que mantiene un equilibrio entre el número de variables independientes y la bondad de ajuste del modelo. Al aumentar el número de variables independientes, SCR disminuye pero a la vez los grados de libertad aumentan y a menos que la disminución de SCR sea realmente significativa, ésta se verá compensada con el aumento de variables independientes. De alguna manera esto significa que el modelo de regresión R^2 ajustado) nos alerta sobre la introducción variables independientes que no representen una verdadera mejora en el modelo.

Pruebas de significancia (pruebas t) de los coeficientes del modelo de regresión

Hemos hallado los coeficientes estimados para la pendiente y el coeficiente de la variable independiente. La pregunta que nos hacemos ahora es si esos valores son o no ciertos. Más aun, lo que nos interesa saber es si son estadísticamente diferentes de cero. Una manera de determinarlo es planteando lo que se conoce como una prueba de hipótesis.

Se puede demostrar que las varianzas muestrales de **a** y **b** estimados son

$$Var(a) = \frac{\frac{SCR}{(n-2)n}\sum x_i^2}{\sum \left(x_i - \bar{x}\right)^2} \qquad (31)$$

y

$$Var(b) = \frac{\frac{SCR}{(n-2)^2}}{\sum \left(x_i - \bar{x}\right)^2} \qquad (32)$$

En nuestro ejemplo tenemos que $\sum \left(x_i - \bar{x}\right)^2$ es igual a 0,104651175 y $\sum x_i^2$ es igual a 0,99919074. Además, $\frac{SCR}{(n-2)}$ es lo que en nuestra tabla se llama Promedio de los cuadrados y en números es $\frac{0{,}025946318}{20} = 0{,}001297316$.

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,843963616**
Coeficiente de determinación R^2	**0,712274586**
R^2 ajustado	**0,697888315**
Error típico	**0,036018272**
Observaciones	**22**

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	0,023728139	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

De este modo valoramos (31) y (32)

$$Var(a) = \frac{\frac{0{,}001297316}{22}0{,}99919074}{0{,}104651175} = 0{,}000563025$$

La desviación estándar s_x o error típico en nuestra tabla, es la raíz de la varianza, entonces

$\sigma_a = 0{,}023728139$

Esta es la cifra que aparece enfrente del estimado de la intercepción en la tabla Análisis de varianza. Lo llamamos el error típico o estándar de la intercepción.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	**0,023728139**	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	0,111339897	7,036385261	7,97608E-07	0,551179564	1,015681259

Para valorar (32) tenemos

$$Var(b) = \frac{\frac{0{,}001297316}{20}}{0{,}104651175} = 0{,}012396573$$

La desviación estándar del coeficiente de la variable independiente es la raíz cuadrada de la varianza, entonces

$\sigma_b = 0{,}111339897$

Esta es la cifra que aparece enfrente del estimado del coeficiente de la variable independiente en la tabla Análisis de varianza. Lo llamamos el error típico o estándar del coeficiente de la variable independiente.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	**0,023728139**	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	0,783430411	**0,111339897**	7,036385261	7,97608E-07	0,551179564	1,015681259

Habíamos visto que el estadístico

$$t = \frac{\bar{x} - m}{s/\sqrt{n}}$$

tiene una distribución **t** de Student.

Nos interesa examinar la hipótesis que el estimado de **a** y de **b** es cero para saber si es o no significativo desde el punto de vista estadístico. Entonces usamos la anterior expresión y definimos μ igual a cero, el valor de x lo reemplazamos por el estimado del coeficiente o de la intercepción y en el denominador incluimos el error típico o estándar cada uno de ellos.

En nuestro ejemplo tendremos

$$t_b = \frac{0{,}783430411}{0{,}111339897} = 7{,}036385261$$

Este es el valor que aparece como Estadístico t para el coeficiente de la variable aleatoria.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	**Estadístico t**	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0,059502091	**0,023728139**	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	**0,783430411**	**0,111339897**	**7,036385261**	7,97608E-07	0,551179564	1,015681259

Mientras ese valor sea más grande será mejor, pero hay que hacer un cálculo con la función =DISTR.T(t;grados de libertad;colas) de Excel. Como el coeficiente puede ser negativo o positivo, le indicamos a la función que haga el cálculo con dos colas. Los grados de libertad son n – 2 (es decir 20) y el valor de t es el que acabamos de calcular. De modo que la función de Excel se valora como =DISTR.T(7,036385261;20;2). El resultado que arroja esta función es 7,97608E-07 (es decir 7,97608 dividido por 10 millones). Este resultado mide la probabilidad de que el valor obtenido para **t** ocurra por azar, dentro de una situación en que el verdadero valor de **b** sea cero y se le conoce como **valor p** o **p-value** en inglés. En nuestro ejemplo esta probabilidad es muy baja y muchísimo menor que los valores tradicionales para medir la significancia estadística de una variable. La conclusión de este análisis es que no podemos rechazar la hipótesis de que $b \neq 0$.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	**Estadístico t**	**Probabilidad**	Inferior 95%	Superior 95%
Intercepción	0,059502091	**0,023728139**	2,507659433	0,020888923	0,010006084	0,108998099
Tasa de inflación	**0,783430411**	**0,111339897**	**7,036385261**	**7,97608E-07**	0,551179564	1,015681259

Procedemos de manera similar para **a**. El estadístico **t** para **a**, $\mathbf{t_a}$, es en nuestro ejemplo

$$t_a = \frac{0,059502091}{0,023728139} = 2,507659433$$

Al hacer la prueba con =DISTR.T(2,507659433;20;2) obtenemos una probabilidad de 0,020888923 es decir, algo más de 2%.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	F	Valor crítico de F	
Regresión	**1**	**0,064231041**	0,064231041	49,51071754	7,97608E-07	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	**Estadístico t**	**Probabilidad**	Inferior 95%	Superior 95%
Intercepción	**0,059502091**	**0,023728139**	**2,507659433**	**0,020888923**	0,010006084	0,108998099
Tasa de inflación	**0,783430411**	**0,111339897**	**7,036385261**	**7,97608E-07**	0,551179564	1,015681259

Aquí la conclusión de este análisis es la misma que para **b**: que no podemos rechazar la hipótesis de que **a** ≠0. Si estamos dispuestos a aceptar el riesgo tradicional de 5% decimos que el coeficiente a es significativo desde el punto de vista estadístico al 5% porque la suma del doble de la probabilidad (dos colas) es menor que 5%.

En este sentido, entonces, decimos que los coeficientes estimados hallados por la regresión son estadísticamente significativos. Esto significa que nuestro modelo

Aumento de salario mínimo = 0,059502091 + 0,783430411(Tasa de inflación) es aceptable desde el punto de vista de R^2 y R^2 ajustado y desde el punto de vista de la significancia estadística de los coeficientes. Sin embargo, los valores de R^2 y R^2 ajustado nos parecen altos. Pero esto es un juicio subjetivo. Debemos hallar la forma de examinar esta apreciación de manera más contundente y sin el elemento subjetivo de parecernos altos o bajos. Para eso utilizaremos la distribución F.

Pruebas de significancia conjunta del grupo de variables (prueba F)

Habíamos estudiado que nos interesaba saber qué tanto de la variación total de los datos se explicaba por medio de la regresión. A partir de este análisis se calculó el coeficiente de determinación R^2 y R^2 ajustado. Si construimos un estadístico F como a continuación

$$F = \frac{\text{Promedio de la suma explicada de los cuadrados}}{\text{Promedio de la suma no explicada}} = \frac{SEC/k_1}{SCR/k_2}$$

donde k_1 y k_2 son los grados de libertad de cada uno, tendríamos para nuestro ejemplo

$$F = \frac{0,064231041/1}{0,025946318/20} = \frac{0,064231041}{0,001297316} = 49,51071754$$

Este es el valor que aparece en la tabla de Análisis de varianza como F. Si usamos la función de Excel =DISTR.F(49,51071754;1;20) encontramos el valor 7,97608E-07 que mide la probabilidad de que ese valor ocurra por azar. Si nuestro nivel de significancia

estadística es de, por ejemplo, 5%, esta prueba es aceptable ya que es mucho menor que 5%.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	**F**	**Valor crítico de F**	
Regresión	**1**	**0,064231041**	**0,064231041**	**49,51071754**	**7,97608E-07**	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	**Estadístico t**	**Probabilidad**	Inferior 95%	Superior 95%
Intercepción	**0,059502091**	**0,023728139**	**2,507659433**	**0,020888923**	0,010006084	0,108998099
Tasa de inflación	**0,783430411**	**0,111339897**	**7,036385261**	**7,97608E-07**	0,551179564	1,015681259

En el caso de una sola variable independiente, este valor es el mismo que se obtiene para la significancia estadística de **b**, el coeficiente de la variable independiente. En nuestro ejemplo diremos que el modelo es adecuado desde el punto de vista estadístico. Esta prueba F adquiere mayor sentido cuando trabajamos con regresión lineal múltiple.

Intervalos de confianza

Hemos establecido un valor puntual estimado de **a** y de **b**. Nos interesa establecer un rango de valores posibles entre los cuales se puede encontrar los valores verdaderos de esos estimados **a** y **b**.

Para establecer un intervalo de confianza definimos un nivel de confianza. El valor típico o usual es el de 95%. Como hemos establecido un estadístico t de dos colas, entonces nuestros intervalos de confianza serán

a ± c(error típico de **a**)

y

b ± c(error típico de **b**)

donde c es el percentil correspondiente a 97,5% de la distribución **t** con (n– número de variables independientes – 1) grados de libertad.

En nuestro ejemplo usamos la función =DISTR.T.INV(Probabilidad;grados de libertad) de Excel para hallar c.

c =DISTR.T.INV(0,05;20) = 2,08596248

De manera que nuestro intervalo de confianza al 95% será

a ± c(error típico de **a**)

0,059502091 ± 2,08596248×0,023728139 = (0,010006084 , 0,108998099)

y

b ± c(error típico de **b**)

0,783430411 ± 2,08596248×0,111339897 = (0,551179564 , 1,015681259)

Estos son los valores que encontramos en nuestra tabla de ANOVA.

	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	**F**	**Valor crítico de F**	
Regresión	**1**	**0,064231041**	**0,064231041**	**49,51071754**	**7,97608E-07**	
Residuos	**20**	**0,025946318**	**0,001297316**			
Total	**21**	**0,090177359**				
	Coeficientes	**Error típico**	**Estadístico t**	**Probabilidad**	**Inferior 95%**	**Superior 95%**
Intercepción	**0,059502091**	**0,023728139**	**2,507659433**	**0,020888923**	**0,010006084**	**0,108998099**
Tasa de inflación	**0,783430411**	**0,111339897**	**7,036385261**	**7,97608E-07**	**0,551179564**	**1,015681259**

Regresión lineal múltiple

Cuando tenemos más de una variable la tabla de ANOVA es básicamente la misma. Las diferencias radican en que se añaden más líneas inferiores, una para cada variable adicional y ya el valor crítico de F obviamente difiere de la probabilidad de t.

Supongamos ahora que nos consideramos que la variable tiempo desempeña un papel importante en nuestro análisis del aumento del salario mínimo. Nuestro modelo sería

Y_{obser} = **a** + **b**(inflación) + **c**(año) + ε

Nuestro modelo para la estimación será

Y_{est} = **a** + **b**(inflación) + **c**(año)

Usando, como lo hicimos para el caso de una variable independiente, la opción Análisis de datos obtenemos las tablas ANOVA siguientes:

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,910928644
Coeficiente de determinación R^2	0,829790994
R^2 ajustado	0,811874257
Error típico	0,028422562
Observaciones	22

Análisis de varianza

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	2	0,07482836	0,03741418	46,31373293	4,94691E-08	
Residuos	19	0,015348999	0,000807842			
Total	21	0,090177359				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	9,170898574	2,515719334	3,645437888	0,001720922	3,905435859	14,43636129
Año	-0,00454765	0,001255603	-3,621886091	0,001815722	-0,00717566	-0,00191964
Tasa de inflación	0,511899903	0,115497985	4,432111105	0,000286055	0,270159766	0,753640039

Como ya sabemos "leer" la tabla ANOVA encontramos lo siguiente:

El modelo "explica" más el comportamiento del salario mínimo puesto que R2 y R2 ajustado aumentan, así

Estadísticas de la regresión	Una variable	Dos variables
Coeficiente de correlación múltiple	0,843963616	0,910928644
Coeficiente de determinación R^2	0,712274586	0,829790994
R^2 ajustado	0,697888315	0,811874257
Error típico	0,036018272	0,028422562

Observaciones	22	22

Se debe observar que el error típico ha disminuido, mientras los coeficientes R^2 que miden la explicación de la variable dependiente han aumentado.

	F	Valor crítico de F
Una variable	49,51071754	7,97608E-07
Dos variables	46,31373293	4,94691E-08

Observemos que las probabilidades críticas para F han disminuido.

Las probabilidades asociadas a los estadísticos t y los estadísticos t de los coeficientes son

	Dos variables		Una variable	
	Estadístico t	Probabilidad	Estadístico t	Probabilidad
Intercepción	3,645437888	0,001720922	2,507659433	0,020888923
Año	-3,621886091	0,001815722		
Tasa de inflación	4,432111105	0,000286055	7,036385261	7,97608E-07

En este caso las probabilidades de los estadísticos t no son estrictamente comparables precisamente porque hay más variables que han "asumido" parte de la explicación.

Los grados de libertad también nos han cambiado porque ahora son dos variables independientes, así

	Una variable	Dos variables
Regresión	1	2
Residuos	20	19
Total	21	21

Esto significa que al calcular los valores de F y de t debemos tener en cuenta que para F los grados de libertad son 2 para el numerador y 19 para denominador. En el caso de una variable teníamos 1 para el numerador y 20 para el denominador.

Al calcular los valores t debemos utilizar 19 grados de libertad para dos variables mientras que en el caso de una variable utilizamos 20.

Lo importante de observar en este caso de dos variables es que el modelo es adecuado (con base en los R^2 y F) y explica más y la nueva variable contribuye a la explicación del comportamiento de la variable dependiente.

Relaciones espurias

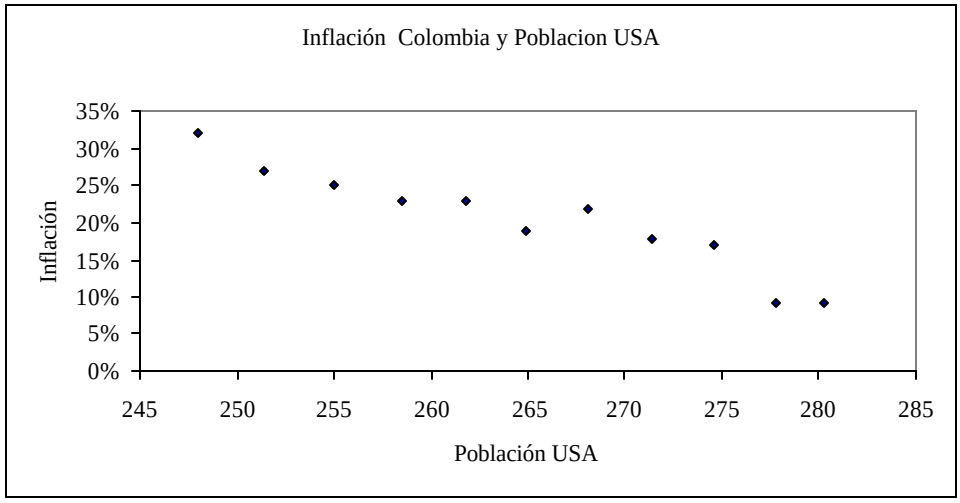
Al inicio de esta nota preveníamos al lector sobre el cuidado que se debe tener de establecer relaciones lógicas entre variables. El mayor esfuerzo que se debe dedicar al análisis de datos es éste. Como ya se vio hay programas como Excel y muchos otros especializados para hacer lo que algunos llaman el trabajo "sucio" de los cálculos.

A manera de ilustración vamos a hacer un análisis de regresión entre dos variable que no tienen ninguna relación entre sí. ¿El lector creería la aseveración que mientras más crece la población de los Estados Unidos la inflación en Colombia baja? Con toda seguridad tildarían de loco a quien hiciera esta afirmación.

Examinemos algunos datos al respecto en la siguiente tabla:

	Población USA en millones	Inflación Colombia
1990	247,98	32%
1991	251,37	27%
1992	254,93	25%
1993	258,45	23%
1994	261,71	23%
1995	264,93	19%
1996	268,11	22%
1997	271,39	18%
1998	274,63	17%
1999	277,84	9%
2000	280,22	9%

La gráfica que ilustra este comportamiento es la siguiente



Tanto la tabla como la gráfica nos indican una relación estrecha entre las variables. Más aun, si hacemos un análisis de varianza como el ilustrado en esta nota encontramos lo siguiente:

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,94986527
Coeficiente de determinación R^2	0,90224403
R^2 ajustado	0,89138226
Error típico	0,0231768
Observaciones	11

Análisis de varianza

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	1	0,04462007	0,04462007	83,0659923	7,7014E-06	
Residuos	9	0,00483448	0,00053716			
Total	10	0,04945455				
	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	1,8387209	0,17953859	10,241369	2,9329E-06	1,43257609	2,24486571
Población USA en millones	-0,00617742	0,00067779	-9,11405466	7,7014E-06	-0,00771069	-0,00464415

Si nos atenemos a las cifras que resultan del análisis de varianza deberíamos concluir que a medida que la población en Estados Unidos aumenta, la inflación en Colombia disminuye. Podemos ver que los R^2 son altos, que los coeficientes **a** y **b** son estadísticamente significativos y que la prueba F nos da más que satisfactoria. ¿Significa esto que sí hay una relación entre la variable independiente (población de los Estados Unidos) y la variable dependiente (inflación en Colombia)? De ninguna manera. Antes de hacer un análisis de regresión es necesario encontrar relaciones causales o razonables o lógicas entre las variables. No cabe la menor duda que en este ejemplo la variable independiente no tiene nada que ver con la variable dependiente a pesar de que los indicadores estadísticos son muy buenos. A esto se le llama relación espuria.

A manera de conclusión

En esta nota pedagógica hemos explicado paso a paso los diversos procedimientos para hacer un análisis de regresión. Así mismo, hemos explicado en detalle cómo leer una tabla ANOVA. Se espera que el lector pueda, con esta guía elaborar modelos de regresión simple y multivariada y comprender el significado de esos modelos desde el punto de vista estadístico.

Hay que advertir que este campo de la estadística pertenece a lo que se conoce como econometría y el tema es muchísimo más complejo de lo que se ha presentado en estos apuntes. Hay pruebas (cuya información provee en gran medida Excel) que se deben realizar para verificar que los supuestos básicos (ver Apéndice) del análisis de regresión se cumplen.

El mensaje que deja esta nota es el siguiente: los recursos de cómputo hacen más fácil la tarea sucia de calcular indicadores, tablas, etc.; esto deja tiempo para dedicar la

inteligencia al diseño de modelos apropiados y para encontrar relaciones causales o lógicas entre las variables.

Referencias

Bowker, Albert H. y Gerald J. Lieberman, *Engineering Statistics*, Prentice-Hall, 1959.

Draper, N. R. y H.Smith, *Applied Regression Analysis*, Wiley, 1966.

Klein, Lawrence R. *Introducción a la econometría*, Aguilar, 1966.

Wonnacott, Ronald J. y Thomas H. Wonnacot, *Econometrics*, 2nd ed., Wiley, 1979.

Wonnacott, Thomas H., Ronald J. Wonnacot, *Introductory Statistics for Business and Economics*, 2ª ed., John Wiley, 1977.

Wooldridge, Jeffrey M., *Introducción a la econometría*, Thompson, 2001. (Traducción de la edición de 2000).

Apéndice

Supuestos que se deben cumplir al hacer análisis de regresión múltiple

1. Existe linealidad en los parámetros. El modelo se puede representar como un modelo lineal, como por ejemplo,

$$Y = a + b_1X_1 + b_2X_2, +\ldots+ b_nX_n + e$$

2. Muestra aleatoria. Se supone que se cuenta con una muestra aleatoria de un universo para este modelo lineal.
3. La media condicional es 0. Esto significa que el valor esperado (promedio) de los errores es cero.
4. Colinearidad imperfecta. Ninguna de las variables independientes es constante y no hay relaciones lineales entre ellas.
5. Homocedasticidad. La varianza del error es la misma para todas las combinaciones de las variables independientes.
6. Normalidad. El error de la población o universo, ε, es independiente de las variables independientes y tiene una distribución normal.