12.
El error de ajuste como herramienta dentro de la metodología en los estudios culturales

FRANCISCO SÁNCHEZ,[1] JULIÁN HERNÁNDEZ,[2]
CARLOS GÓMEZ[3] Y OLGA NELLY ESTRADA[4]

Introducción

La ciencia es ante todo un método de conocimiento basado en una serie de pasos. Este conocimiento tiene sus propias características, que sirven para diferenciarla de otros métodos: el objetivo es la inferencia, obtener inferencias válidas a partir de información empírica; el contenido de la ciencia, el centro, es el método, se trata de una serie de reglas y pasos y no de un objeto de estudio; el método y técnica deben ser explícitos y reconstruibles, y; los resultados de su aplicación son inciertos.

1 Profesor –investigador, Doctor en Sociología por el ICSyH de la BUAP, adscrito a la Licenciatura en Ciencias Políticas de la Facultad de Derecho y Ciencias Sociales de la misma institución.

2 Profesor-Investigador, Doctor en Sociología del ICS y H de la BUAP, adscrito a la licenciatura en Ciencias Políticas de la Facultad de Derecho y Ciencias Sociales de la BUAP.

3 Doctorado de Tercer Ciclo en Derecho Público por la Universidad de París XI; Docente de tiempo completo en la Facultad de Ciencias Políticas y Relaciones Internacionales de la Universidad Autónoma de Nuevo León; Reconocido con el Perfil deseable PRODEP.

4 Doctora en Humanidades y las Artes. Profesora- Investigadora de la UANL. Miembro del Sistema Nacional de Investigadores Nivel 1.

En este trabajo me detengo en la cuarta característica: los conocimientos obtenidos, los resultados, con el método científico son inciertos, no se tiene la certeza, en otras palabras, si se imagina la certeza como el área bajo una curva normal, entonces un porcentaje de ella es confiable, pero uno es de error. Otros métodos de conocimiento aseguran que sus resultados son totalmente confiables, la ciencia no.

Se explora una de las formas de asumir el error. El error predictivo paramétrico, se presenta en particular en dos análisis paramétricos, en la correlación y la regresión. Cuando se han obtenido las betas de regresión y se hacen interpolaciones en los valores de la variable explicada se pueden establecer restas, elevar al cuadrado, sumarlas y establecer la bondad de ajuste entre la curva proyectada y la curva real, eso es el tamaño del error.

Muchos estudios culturales llevan a cabo encuestas, los resultados de los cuestionarios son vaciados en distintos paquetes computacionales y algunos llevan a cabo análisis con el refinamiento de técnicas de análisis paramétricas como la correlación o la regresión, que aquí se toman en consideración.

Palabras clave: cultura, método, estadística paramétrica, correlación, regresión, cálculo del error.

La condición de normalidad de los datos

En el análisis de regresión se encuentra un tipo de error, el de ajuste. La regresión es una de las formas de análisis de la estadística paramétrica, por lo cual se aplica bajo sus supuestos, a saber:

a) La distribución de la población es normal en la variable dependiente.
b) La medición se elabora por intervalos o razón en la variable dependiente.
c) Si dos o más poblaciones forman parte del estudio, entonces tienen una varianza homogénea.[5]

5 Wiersma. "Research methods in education: an introduction." Allyn And Bacon, Boston, Massachusets, 1986. P. 344.

La normalidad de una distribución se encuentra en dos condiciones: la de asimetría y la de curtosis

Para abordarlas, la referencia son los momentos de la forma de la distribución. Si se considera una distribución de datos, se establecen las desviaciones respecto de la media, con la siguiente fórmula:

$$x = (x_i - \overline{x})$$

Establecido el valor de x, el momento de primer orden respecto de la media de la distribución está dado por la siguiente:

$$m_1 = \frac{\sum x}{N}$$

Este primer momento siempre tiene un valor de 0, lo cual expresa una característica que define a la media, a saber, que se encuentra ubicada en la parte media de la distribución, algunas observaciones se encuentran a su derecha y otras a su izquierda, en otras palabras, el área del histograma que queda a la derecha es la misma que la que aparece a su izquierda, de modo que, al efectuar las restas unas resultan positivas y otras negativas y, la suma de las restas siempre da 0.

El momento de segundo orden es:

$$m_2 = \frac{\sum x^2}{N}$$

Es la fórmula de la varianza. Para evitar el resultado constante de 0 del momento de primer orden, se elevan las restas al cuadrado antes de sumarlas y dividirlas entre el número de observaciones. De esa manera, la varianza es el promedio de las desviaciones de las observaciones respecto de la media, elevadas al cuadrado. Recuérdese que, cuando es necesario, se puede retornar a las unidades de medición originales obteniendo la raíz cuadrada de la varianza, operación cuyo resultado llamamos Desviación Estándar.

Los siguientes momentos son los de tercer y cuarto orden, los cuales, al igual que en el caso de segundo orden, elevan las restas a una potencia, tercera o cuarta, el efecto, cuando los resultados de las restas son mayores a la unidad, es que se hacen más grandes las diferencias, asimismo cuando el resultado de la resta es positivo, siempre será positivo el resultado de la elevación independientemente de la potencia; cuando las restas son negativas, lo siguen siendo, después de la elevación en primera potencia y en tercera; en tanto que en segunda y cuarta se vuelve positivo el resultado. Las fórmulas, ya definido que $x = (x_i - \overline{x})$, son:

$$m_3 = \frac{\sum x^3}{N}$$

$$m_4 = \frac{\sum x^4}{N}$$

A partir de los momentos es que se construyen las fórmulas de cálculo. La de asimetría es la siguiente:

$$B_1 = \frac{m_3}{\sqrt{m_2^{\ 3}}}$$

En tanto que la fórmula para calcular la curtosis es:

$$B_2 = \frac{m_4}{m_2^2}$$

Si la curva está sesgada a la derecha, B_1 tendrá un valor positivo, mientras que si el sesgo es negativo, B_1 ofrecerá un valor negativo. Por ser una magnitud relativa, B_1 expresa la cantidad relativa de asimetría y puede ser utilizada para comparar distribuciones que contienen diferentes unidades de medición.

En cuanto a B_2, se utiliza como coeficiente de curtosis o medida del grado de apuntamiento de una distribución. Los valores pequeños de B_2 representan una curva platicúrtica (más baja que la curva normal), mientras que valores altos de B_2 indican una distribución leptocúrtica o apuntada. La curva normal tiene un valor de B_2 igual a 3 y se denomina mesocúrtica.

La medición debe ser por intervalos o razón

Para establecer con mayor claridad lo que son tales niveles de medición, es necesario revisar otros niveles de medición.

a) Nominal. Son categorías sin orden ni jerarquía, sólo marcan diferencias en una o más características.

Un ejemplo es la variable sexo, la cual toma una de dos categorías, masculino o femenino.

Pueden utilizarse como símbolos de clasificación números, 1 = masculino, 2 = femenino, pero no se pueden aplicar a ello las operaciones aritméticas.

b) Ordinal. También hay categorías, pero están ordenadas de mayor a menor, entonces las etiquetas o símbolos indican ese orden.

Por ejemplo: 1. Presidencia de gobierno; 2. Consejería de cultura, deportes políticas sociales y vivienda; 3. Viceconsejería de políticas sociales e inmigración; 4. Dirección general de políticas sociales.[6]

No se pueden aplicar las operaciones aritméticas.

c) Intervalos. En este caso las categorías están ordenadas, hay una unidad de medida y el intervalo es constante.

Como ejemplo una escala en la que se menciona una frase y el encuestado contesta si está: 1. Muy en desacuerdo; 2. En desacuerdo; 3. Indiferente; 4. De acuerdo; 5. Muy en desacuerdo.

Se pueden aplicar análisis de datos con estadística descriptiva e incluso inferencial.

6 Gobierno de Canarias. http://www.gobcan.es/organizacion/departamentos.jsf (Consulta: 22 de Junio de 2012).

d) Razón. Tiene las mismas características que la medición por intervalos, es decir, categorías ordenas, unidad de medida e intervalo constante, pero además el cero es real, entonces hay un punto en la escala en el que no está presente la propiedad.

Por ejemplo, el número de hijos por pareja. Hay parejas con 0 hijos, un hijo, dos hijos, etc.

En este caso se aplican técnicas estadísticas descriptivas e inferenciales.

Entonces, los dos últimos niveles de medición son los que corresponden a la utilización de la estadística paramétrica, entre otras técnicas la regresión.

La condición de varianzas homogéneas

La última condición es que si son dos o más poblaciones las implicadas en el estudio, entonces las varianzas deben ser homogéneas. Para ello debe recurrirse a lo ya explicado en el cálculo de la asimetría y la curtosis. Se trata del momento de segundo orden, es decir, el promedio de las desviaciones de las observaciones respecto de la media, elevadas al cuadrado. Pues bien, las varianzas deben ser parecidas en las poblaciones.

Regresión, correlación y error de ajuste

Ya habiendo establecido los supuestos del análisis paramétrico, debe adentrarse en la regresión.

La regresión lineal establece la relación entre series de datos, estableciendo la relación de cambio y, adicionalmente, un punto de intercepto.

Los posibles valores fluctúan desde –n hasta +n.

El formato para obtener los valores es:

$$y_i = \beta_0 + \beta_1 \ x$$

La pendiente de la línea de regresión se obtiene con la siguiente fórmula:

$$\beta_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

El intercepto con Y de la línea de regresión se obtiene con la siguiente fórmula:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Donde la barra superior en y y x indica que se trata de los promedios.[7]

Entonces, los elementos necesarios son los siguientes:

$$\sum xy = \qquad \sum x =$$

$$\sum y = \qquad \sum x^2 =$$

$$n = \qquad \bar{y} =$$

$$\bar{x} =$$

Más recientemente se incluyen en la literatura especializada otras formulaciones para el cálculo de la β_0 y la β_1, los dos siguientes pares son muestra de ello y constituyen el segundo y tercero de los presentados en este escrito.[8]

7 Wayne, Daniel. *Estadística con aplicaciones a las ciencias sociales y a la educación.* Mc Graw-Hill, México, 1988. (Ver el capítulo 9. Análisis de regresión lineal y de correlación simple).

8 R. Carter E. A. "Undergraduate econometrics." John Willey and Sons Inc. United States of America, 2000. P. 55; Correa, Guillermo. "Contribuciones al análisis multivariante no lineal." Memoria para optar por el grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca, Universidad de Salamanca, 2008.

Segundo:

$$\beta_1 = \frac{\sum (x_t - \overline{x})\ y_t}{\sum (x_t - \overline{x})^2}$$

$$\beta_0 = \overline{y} - \beta_2 \overline{x}$$

Tercero:

$$\beta_1 = \frac{T \sum x_t y_t - \sum x_t \sum y_t}{T \sum x_t^2 - \left(\sum x_t\right)^2}$$

$$\beta_0 = \overline{y} - \beta_2 \overline{x}$$

Para ejemplo, se retoman los siguientes datos, los cuales corresponden a pacientes que sufren un infarto y que les es inyectada una sustancia en distintas dosis, que aumenta desde 0.5 mm. cúbicos, hasta 3.0, dicho aumento va en media unidad creciendo paulatinamente. Además se da el dato de los segundos que tarda en reaccionar el paciente.[9]

En el cuadro siguiente se encuentran realizadas las operaciones necesarias para obtener los valores de los parámetros de regresión:

Sustituyendo y despejando:

$$\beta_1 = 18.8571 .$$

$$\beta_0 = \overline{y} - \beta_2 \overline{x} = 42.7778 - 18.8571 * 1.75 = 9.7779$$

$$\beta_0 = 9.7779$$

Entonces, la función de regresión muestral es la siguiente:

$$y_i = \beta_0 + \beta_1\ x$$

$$y_i = 9.7779 + 18.8571\,x$$

9 Wayne, Daniel. *Estadística con aplicaciones a las ciencias sociales y a la educación*. Mc Graw-Hill, México, 1988. pp.

$$\beta_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{1595 - \frac{31.5 * 770}{18}}{68.25 - \frac{(31.5)^2}{18}} = \frac{1595 - 1347.5}{68.25 - 55.125} = \frac{247.5}{13.125} = 18.8571$$

Con el segundo par de fórmulas se requieren los siguientes cálculos:

Sustituyendo y despejando:

$$\beta_1 = \frac{\sum (x_t - \bar{x})\, y_t}{\sum (x_t - \bar{x})^2} = \frac{247.5}{13.125} = 18.8571$$

$$\beta_1 = 18.8571$$

$$\beta_0 = \bar{y} - \beta_2 \bar{x} = 42.78 - 18.8571 * 1.75 = 9.78$$

$$\beta_0 = 9.78$$

Con el tercer par de fórmulas (véase el cuadro 4)

Sustituyendo y despejando:

$$\beta_1 = \frac{T \sum x_t y_t - \sum x_t \sum y_t}{T \sum x_t^2 - \left(\sum x_t\right)^2} = \frac{18 * 1595 - 31.5 * 770}{18 * 68.25 - (31.5)^2} = \frac{28710 - 24255}{1228.5 - 992.25} = \frac{4455}{236.25} = 18.8571$$

$$\beta_1 = 8\ .8571$$

$$\beta_0 = \bar{y} - \beta_2 \bar{x} = 42.78 - 18.8571 * 1.75 = 9.7801$$

Ahora se calcularán, con la función de regresión muestral, los valores para la variable explicada, en este caso Y, que es el tiempo de reacción:

Para explicar la obtención del error estándar, primero se revisa la definición:

Cuadro 1
Solución inyectada y tiempo de reacción en pacientes con paro cardiaco

	Xi	**Yi**
1	0.5	12
2	0.5	22
3	0.5	30
4	1	18
5	1	32
6	1	36
7	1.5	30
8	1.5	34
9	1.5	46
10	2	40
11	2	44
12	2	50
13	2.5	44
14	2.5	60
15	2.5	64
16	3	64
17	3	68
18	3	76

Fuente: Waynem Daniel. Estadística con aplicaciones a las ciencias sociales y a la educación. McGraw-Hill, México, 1988. p 317

> "Error estándar: … definido como la desviación promedio de un estimado de los valores reales de la población."[10]

Como desviación se entiende una resta; si es promedio entonces se divide entre el número de observaciones. Otra definición, de García Ferrando, más explícita, nos dice:

> … podemos seguir el criterio de la varianza, que consistirá simplemente en restar de cada valor real de Y el resultante de la ecuación, se eleva al cuadrado la diferencia, se suman todos los casos

10 Hernández Sampieri, Roberto, E. A. *Metodología de la investigación*. Mc Graw-Hill, México, D. F., 1991. P. 230.

Cuadro 2
Obtención de la función de regresión (Método A)

	Xi	Yi	Xi2	Yi2	XiYi
1	1	12	0	144	6
2	1	22	0	484	11
3	1	30	0	900	15
4	1	18	1	324	18
5	1	32	1	1,024	32
6	1	36	1	1,296	36
7	2	30	2	900	45
8	2	34	2	1,156	51
9	2	46	2	2,116	69
10	2	40	4	1,600	80
11	2	44	4	1,936	88
12	2	50	5	2,500	100
13	3	44	6	1,936	110
14	3	60	6	3,600	150
15	3	64	6	4,096	160
16	3	64	9	4,096	192
17	3	68	9	4,624	204
18	3	76	9	5,776	228
	31.50	770	68.25	38,508	1,595
	1.75	42.78			

Fuente: Elaboración propia.

y se divide por N. Es decir, mediante la estimación de la varianza $S^2_{yx} = \frac{\sum (Y - Y')^2}{N}$, en donde Y´ representa el valor de Y calculado mediante la aplicación de la ecuación de predicción.[11]

Entonces, mediante la fórmula de regresión muestral, con los parámetros obtenidos, se calculan los valores de Y para cada valor dado de X. Por ejemplo, para el primer valor, X = 0.5 el valor de Y calculada:

11 García Ferrando, Manuel. "Socioestadística." Alianza Universidad Textos, Madrid, 1997. P. 268.

Cuadro 3
Obtención de regresión (Método 2)

	Xi	**Yi**	**Xt - Xmed**	**(Xt-Xmed) Yt**	**(Xt-Xmed) (Xt-Xmed)**
1	1	12	-1	-15.00	1.56
2	1	22	-1	27.50	1.56
3	1	30	-1	-37.50	1.56
4	1	18	-1	-13.50	0.56
5	1	32	-1	-24.00	0.56
6	1	36	-1	-27.00	0.56
7	2	30	0	-7.50	0.06
8	2	34	0	-8.50	0.06
9	2	46	0	11.50	0.06
10	2	40	0	10.00	0.06
11	2	44	0	11.00	0.06
12	2	50	0	12.50	0.06
13	3	44	1	33.00	0.56
14	3	60	1	45.00	0.56
15	3	64	1	48.00	0.56
16	3	64	1	80.00	1.56
17	3	68	1	85.00	1.56
18	3	76	2	95.00	1.56
	31.50	770	-	247.50	13.125
	1.75	42.78			

Fuente: elaboración propia.

$$y_i = \beta_0 + \beta_1 \ x$$

$$y_i = 9.7779 + 18.8571\, x$$

$$y_i = 9.7779 + 18.8571(.5)$$

$$y_i = 19.20$$

Obtenidos todos los valores, se procede a elaborar las restas en otra columna, Y – Y´; posteriormente, en una columna adyacente,

Cuadro 4
Obtención de la función de Regresión (Tercer Método)

	Xt	Yt	Xt2	XtYt
1	1	12	0	6
2	1	22	0	11
3	1	30	0	15
4	1	18	1	18
5	1	32	1	32
6	1	36	1	36
7	2	30	2	45
8	2	34	2	51
9	2	46	2	69
10	2	40	4	80
11	2	44	4	88
12	2	50	4	100
13	2	44	6	110
14	3	60	6	150
15	3	64	6	160
16	3	64	9	192
17	3	68	9	204
18	3	76	9	228
	31.50	770	68.25	1,595
	1.75	42.78		

Fuente: elaboración del autor.

se elevan al cuadrado esas restas, ya que, tal como sucede con el cálculo de la varianza, si se suman inmediatamente las restas al calce de su columna, el resultado sería invariablemente de cero, para evitar eso se elevan al cuadrado las restas y se suman. Para finalizar este procedimiento se divide, según lo señalado por García Ferrando, entre el número de observaciones. El resultado es de 50.11, que es una medida del ajuste entre la Y real y la Y proyectada con la función de regresión muestral.

Cuadro 5
Cálculo del error estándar

	Xi	Yi	Y'	Y-Y'	(Y-Y')(Y-Y')
1	0.5	12	19.20	-7.20	51.87
2	0.5	22	19.20	2.80	7.83
3	0.5	30	19.20	10.80	116.60
4	1	18	28.63	-10.63	112.93
5	1	32	28.63	3.37	11.38
6	1	36	28.63	7.37	54.36
7	1.5	30	38.05	-8.05	64.83
8	1.5	34	38.05	-4.05	16.42
9	1.5	46	38.05	7.95	63.17
10	2	40	47.48	-7.48	55.91
11	2	44	47.48	-3.48	12.09
12	2	50	47.48	2.52	6.37
13	2.5	44	56.90	-12.90	166.46
14	2.5	60	56.90	3.10	9.60
15	2.5	64	55.90	7.10	50.38
16	3	64	66.33	-2.33	5.41
17	3	68	66.33	1.67	2.80
18	3	76	66.33	9.67	93.57
	31.5	770			901.97
					50.11
					7.08

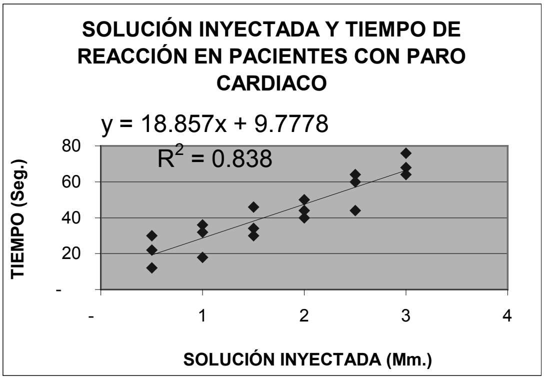
Fuente: elaboración del autor.

Para una mejor apreciación se puede calcular y representar gráficamente la línea de la función de regresión muestral, la función y el error estándar con Excel (véase cuadro y gráfica página siguiente).

Para obtener lo que se presenta en el cuadro y gráfico anterior, se obtiene, a partir de la información de las dos variables, dispuesta en un par de columnas, un gráfico en Excel de tipo "Estándar", escogiendo el de "XY (Dispersión)" y el primer "Subtipo", se puede dar nombre al gráfico y establecer las variables en los ejes; se finaliza el gráfico. Ahora se trabaja sobre el gráfico: en uno de los puntos de dispersión se da "click" derecho y "Agregar línea de tendencia"

CUADRO 6 Y GRÁFICO 1

x	y
0.5	12
0.5	22
0.5	30
1	18
1	32
1	36
1.5	30
1.5	34
1.5	46
2	40
2	44
2	50
2.5	44
2.5	60
2.5	64
3	64
3	68
3	76



r	r*r
0.915	0.838

Fuente: elaboración del autor

y en la pestaña "Tipo" se escoge "Lineal"; en la pestaña "Opciones" se escoge "Presentar ecuación en el gráfico" y "Presentar el valor R cuadrado en el gráfico".[12]

Los valores de los parámetros se obtienen más rápidamente que mediante una calculadora de bolsillo, pero coinciden en su cuantía. Dichos valores se pegan en Excel dando como valor x el correspondiente a la celda, se arrastra extendiendo el pegado tanto como sea necesario y se procede a hacer las restas y la elevación al cuadrado de las mismas, para luego sumarlas y proceder a obtener la varianza.

En otro texto no solamente se llega al cálculo de la varianza, sino que se procede a obtener la desviación estándar. Puede leerse en Gujarati lo siguiente.

"La raíz cuadrada positiva de δ^2 :

12 Una mayor explicación se presenta al final del texto, en la parte 5.

> $\delta = \sqrt{\frac{\sum e_i^2}{N-2}}$ se conoce como el error estándar de la estimación. Básicamente, es la desviación estándar de los valores de Y con respecto a la línea de regresión estimada y que se utiliza con frecuencia como una medida que resume la "bondad del ajuste" de la línea de regresión estimada…"[13]

A continuación se hace el cálculo de la desviación estándar bajo la definición anterior:

$$\delta = \sqrt{\frac{\sum e_i^2}{N-2}} = \sqrt{\frac{901.97}{18-2}}$$

$$\delta = \sqrt{56.37} = 7.508$$

Como puede observarse, la medida de García llega hasta el cálculo de la varianza, y en Gujarati se prolonga hasta la desviación estándar, asimismo hay otra diferencia en que la división García la hace respecto de N y en Gujarati entre N-2. En este cálculo al término N-2 se le conoce como el "número de grados de libertad":

> El término grados de libertad muestra el número total de observaciones en la muestra (= N) menos el número de restricciones lineales impuesto en ellas.[14]

Por ejemplo, antes de que la suma de los errores se realice, debieron haberse calculado las Betas de la regresión lineal, estos dos estimadores ponen dos restricciones sobre el cálculo de la desviación estándar, entonces hay N-2 y no N observaciones independientes para el cálculo.

Por lo que se ve es más preciso el resultado si se sigue el método de Gujarati, en él se consideran los grados de libertad y se llega, no sólo a la varianza, sino hasta la desviación estándar, como esta desviación se calcula respecto de los errores, es decir, de las des-

13 Gujarati, Damodar. "*Econometría.*" Mc Graw Hill, México, 1993. P. 64.

14 *Íbidem.*

viaciones de las Y´ respecto a Y, entonces se denomina, además de desviación estándar de los errores, por contracción, error estándar. Es, entonces, el error estándar, otra forma en la que la ciencia asume las posibilidades de error, al aplicar una herramienta de orden paramétrico.

Por otra parte, a un procedimiento como el que se ha elaborado, en el que se calcula el error estándar de las observaciones, se le denomina intrapolación. Ésta consiste en calcular valores, mediante las Betas de regresión, para los puntos en los que hay medición de la variable explicada, de la coincidencia o falta de ella se obtienen las restas o suma de los errores, pero las Betas o parámetros también se utilizan con fines de pronóstico. Como ejemplo tómense los siguientes datos, correspondientes al ingreso familiar y los gastos en alimentos de 40 universitarios. Se presentan los cálculos para las Betas:[15]

Sustituyendo y despejando:

$$\beta_1 = \frac{\sum (x_t - \overline{x})\, y_t}{\sum (x_t - \overline{x})^2} = \frac{196597.54}{1532463.02} = 0.128$$

$$\beta_1 = 0.128$$

$$\beta_0 = \overline{y} - \beta_2 \overline{x} = 130.31 - 0.128 * 698 = 40.768$$

$$\beta_0 = 40.768$$

Elaborar un pronóstico o extrapolar, sería por ejemplo, aplicar los parámetros a un valor de la variable independiente que no se encuentra en el rango de valores reales. El mínimo es de $ 258.30 y el máximo de $ 1, 154.60, entonces se aplicará a un valor de $ 1, 300.00:

$$y_i = \beta_0 + \beta_1\ x$$

15 R. Carter E. A. *Undergraduate econometrics.* John Willey and Sons Inc. United States of America, 2000. P. 50.

$$y_i = 40.768 + 0.128\,x$$

$$y_i = 40.768 + 0.128\,(1300)$$

$$y_i = 207.56$$

Pero la pregunta es qué tanto se apega esa predicción a la realidad, es decir, puede ser que la predicción pueda ser errónea, ya que hay valores en la variable explicatoria menores a $ 1, 300.00 que tienen gastos mayores a $ 207.56, podría ser que la relación no sea estrictamente lineal, y entonces tenemos lo que puede llamarse una forma derivada de error dentro de lo que implica el proceso de regresión.

La correlación puede ser útil para comprender esto. La fórmula de la correlación de Pearson, cuyo indicador mide el grado de asociación entre variables, es la siguiente:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\ \sqrt{n\sum y^2 - (\sum y)^2}}$$

Por ejemplo, la correlación para los ingresos y gastos de familias de universitarios es de 0.563

Puede darse otro uso a esa medición, si se eleva al cuadrado:

> … obsérvese que el r2 definido previamente también se puede calcular como el cuadrado del coeficiente de la correlación existente entre el valor real de yt y el valor estimado de Yi.[16]

Se trata de una medida del ajuste entre los valores reales y los obtenidos con la función de regresión muestral:

> La cantidad r2 definida … se conoce como coeficiente de determinación (muestral), y es la más ampliamente utilizada medida de la

16 Gujarati, Damodar. *Econometría.* Mc Graw Hill, México, 1993. P. 71.

> bondad del ajuste de una línea de regresión. Claramente, r2 mide la proporción o porcentaje de la variación total en Y explicada por el modelo de regresión.[17]

En búsqueda de un mejor ajuste puede variarse, usando el lenguaje de Excel, el "tipo" del formato de la línea de tendencia. Se presentan a continuación los gráficos con los valores de las funciones, lineal, logarítmica, polinomial, potencial y exponencial, y los R^2:

El modelo potencial es el que presenta el mayor R2, que es de 0.4014. La forma general de los modelos con distintas funciones se representan en la siguiente gráfica:

Como puede observarse, un pronóstico "exitoso" depende de diversos aspectos, como el modelo teórico que se utilice, es decir las variables explicatorias; tener un R^2 alto; un error estándar bajo; pero también es importante el "tipo" de modelo que se utilice, ya que la sola alta correlación o bajo error estándar no aseguran el pronóstico "exitoso".

Manual de interpolación y extrapolación con Excel

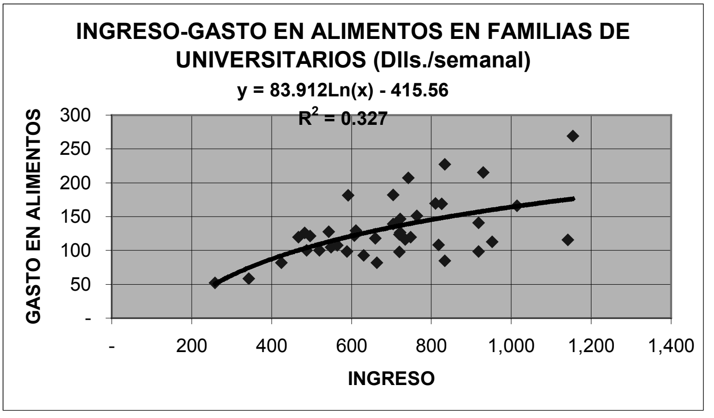
a) Datos
   1. Introducir los datos a excel.

b) Gráfico de líneas
   1. Obtener un gráfico de "líneas".
   2. "Clic", imponer orden de pestaña de "selección de serie", obteniendo así el "asistente para gráficos".
   3. Tipo Estándar, de dispersión.
   4. Finalizar.

c) Sobre el gráfico

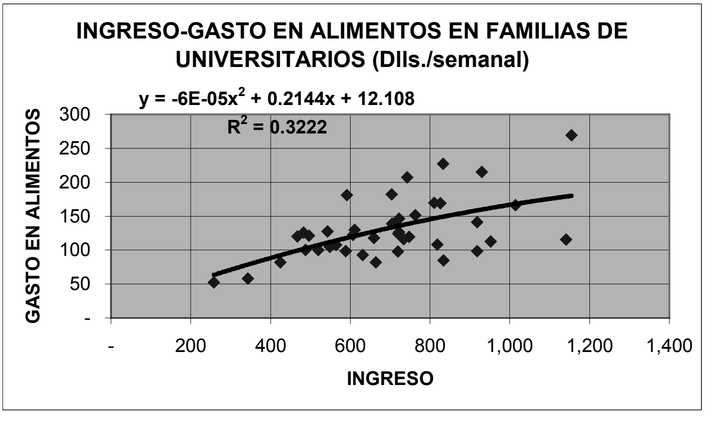
17 *Íbidem.* P. 70.

GRÁFICA 3. MODELO LOGARÍTMICO



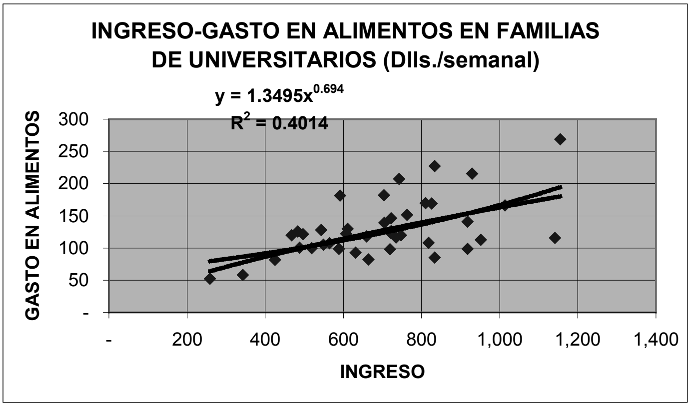
Fuente: elaboración del autor.

GRÁFICA 4. MODELO POLINOMIAL



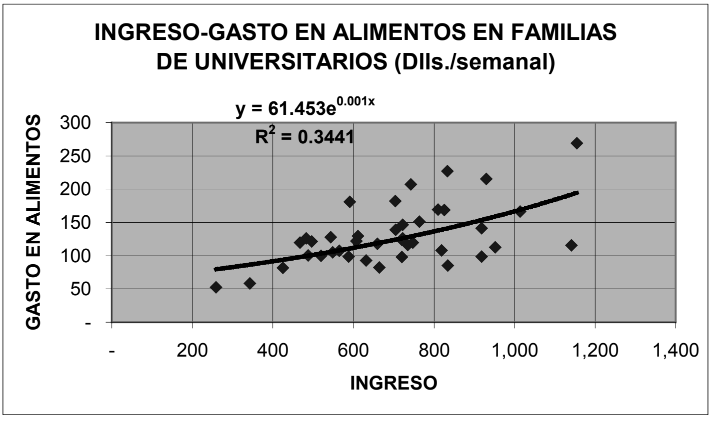
Fuente: elaboración del autor.

GRÁFICA 5. MODELO POTENCIAL



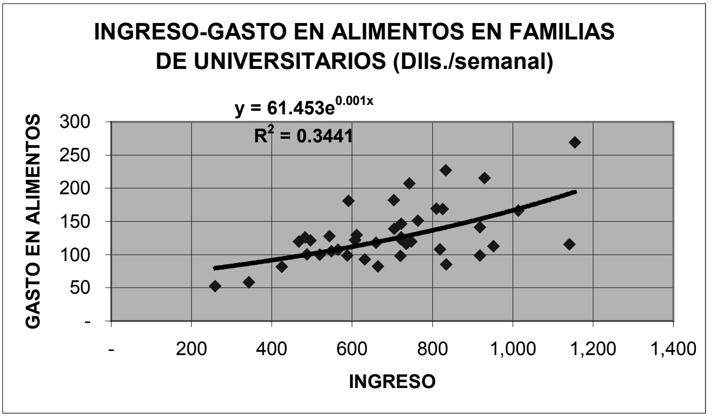
Fuente: elaboración del autor.

GRÁFICA 6. MODELO EXPONENCIAL



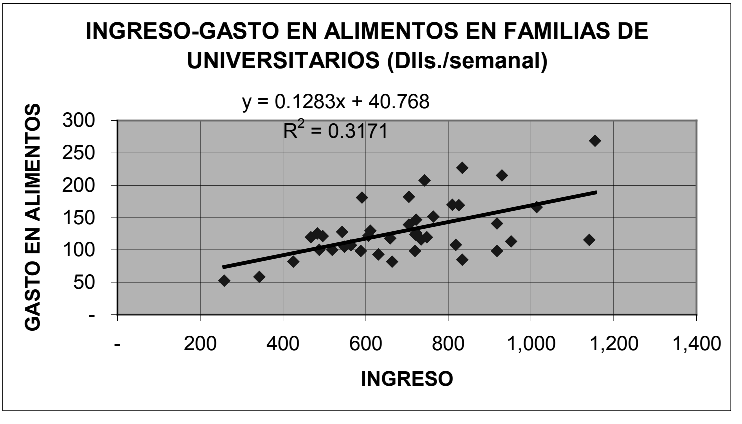
Fuente: elaboración del autor.

GRÁFICA 6. MODELO EXPONENCIAL



Fuente: elaboración del autor.

GRÁFICA 2. MODELO LINEAL



Fuente: elaboración del autor.

1. Hacer "clic derecho" sobre la línea, cuidando de no hacerlo sobre la línea de fila, hacer "clic" en línea de tendencia.
2. Aparecen varios tipos de gráfico: lineal, logarítmica, polinomial, potencial, exponencial y media móvil.
3. Escoger pestaña "opciones".
4. Extrapolar, darle períodos atrás y adelante.
5. En "señalar intersección" se refiere al punto de intercepto, y se refiere a forzar el valor a en $y = b_0 + b_1x$, esto podría variar la línea de tendencia porque ella buscará pasar por entre los datos, ya que es la línea de mínimos cuadrados ordinarios.
6. Dar "palomita" en "Presentar ecuación en el gráfico", lo mismo en "presentar el valor R cuadrado en el gráfico". Dar "aceptar".
7. Dar "clic derecho" en el área del gráfico.
8. Escoger "ubicación", y escoger entre "en hoja nueva" o "como objeto" y dar "clic" en la escogida.
9. Se obtiene $y = b_1x + b_0$ y no $y = b_0 + b_1x$. La fórmula y el R cuadrado pueden agrandarse o arrastrarse seleccionándolos.

e) Para mejorar la regresión

1. "Clic derecho" en la curva de M. C. O. y seleccionar "formato de línea de tendencia", aparece un cuadro anterior, escoger en él la opción "tipo" y se escoge "logarítmica". Los valores de M. C. O. y de R cuadrado variarán.
2. De la misma forma que en 1, se siguen tomando las otras opciones de "formato de línea de tendencia", que se escoge con "clic derecho" en la línea de tendencia y se escoge "tipo" para escoger el gráfico.
3. Con 2 se está buscando el R cuadrado más elevado. Recuérdese que el R cuadrado indica qué tanto se ajusta la tendencia a los datos reales, es decir, qué tanto la línea de datos calculados se ajusta a la línea de tendencia de los datos reales, ambos aparecen sobre el gráfico.

Considérense los siguientes casos: 1) cuando se utiliza la media móvil aparece "período", el cual contiene un máximo y un mínimo, es necesario darle orden. No presenta R cuadrado ni ecuación, se presenta según los períodos que se hayan escogido. Todos los demás tipos sí presentan fórmula y R cuadrado, pero si se ha utilizado anteriormente la media móvil se borra la orden de presentarlos, por lo que es necesario volver a dar la orden de darlos; 2) En la polinomial al aumentar los "orden" aumentan las b y se eleva el R cuadrado, el mínimo de "orden" que acepta es 2.

NOTA 1.

A pesar de encontrar R cuadrado = 1 o muy cercano a 1 hay que evaluar lógicamente.

NOTA 2.

Si no tuviera PC haría como si fuera lineal y calcularía a y b mediante fórmulas, considerando los valores de x. Si no tengo datos para x y para y, tomo los dados por default.

f) Interpolación y extrapolación

1. Ya obtenida la línea de tendencia, seleccionando un R cuadrado y una ecuación, escogiendo o tratando de escoger una ecuación con R cuadrado alto, con las salvedades mencionadas en las notas anteriores. Con la ecuación se pueden obtener los datos intermedios o exteriores al período de observaciones, con una calculadora de bolsillo.
2. Pero si se desea hacerlo con excel: a) Seleccionar la ecuación y pegarla en la hoja de cálculo donde están los datos originales, se exponen los valores para x (se deben hacer explícitos por default) y se elabora la y estimada adecuando la función y rellenando hacia abajo.
3. Se puede obtener el error: e = y – y estimada. Hay que evaluar por qué en algunos puntos el error es alto, puede deberse a que haya puntos de inflexión en los datos observados.

4. Según se desee interpolar o extrapolar deben darse los espacios necesarios y, los datos para x (considerando los dados por default), y se pegará la función rellenando hacia abajo.

Bibliografía

CAMPBELL, Donald y Stanley, Julian. *Diseños experimentales y cuasiexperimentales en la investigación social.* Amorrortu editores, Buenos Aires, 1995.

CORREA, Guillermo. *Contribuciones al análisis multivariante no lineal.* Memoria para optar por el grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca, Universidad de Salamanca, 2008.

DUHNE, Carlos. *Técnicas estadísticas y administrativas.* LIMUSA, México, 1984.

GARCÍA Ferrando, Manuel. *Socioestadística.* Alianza Universidad Textos, Madrid, 1997.

GOBIERNO DE CANARIAS. http://www.gobcan.es/organizacion/departamentos.jsf (Consulta: 22 de Junio de 2012).

GUJARATI, Damodar. *Econometría.* Mc Graw Hill, México, 1993.

HERNÁNDEZ Sampieri, Roberto, E. A. "Metodología de la investigación." Mc Graw-Hill, México, D. F., 1991.Koosis, Donald. *Elementos de inferencia estadística.* LIMUSA, México, 1980.

R. CARTER E. A. *Undergraduate econometrics.* John Willey and Sons Inc. United States of America, 2000.

WAYNE, Daniel. *Estadística con aplicaciones a las ciencias sociales y a la educación.* Mc Graw-Hill, México, 1988.

WIERSMA. *Research methods in education: an introduction.* Allyn And Bacon, Boston, Massachusets, 1986.