Aplicación de la regresión lineal en un problema de pobreza

Application of linear regression on the problem of poverty

Diego Fernando Cardona Madariaga[*]
Javier Leonardo González Rodríguez[**]
Miller Rivera Lozano[***]
Edwin Hernán Cárdenas Vallejo[****]
Universidad del Rosario

Recibido: 25-07-13 /Aceptado: 01-11-13

Resumen

Este artículo pretende mostrar al profesional de cualquier área, las bondades de la estadística inferencial en lo referente al análisis de regresión lineal simple. Para ello se recurre a la observación de algunas situaciones de la administración y la ingeniería y en particular, al desarrollo de un caso aplicado a la economía colombiana.

Palabras clave: Regresión lineal, estadística inferencial.

Abstract

This article attempts to demonstrate to professionals in any field the importance of inferential statistics with regard to simple linear regression analysis. To accomplish this we turn to the observation of a few situations of administration and engineering and in particular, the development of a case applied to the Colombian economy.

Key words: lineal regression analysis, inferential statistics.

[*] Matemático, Ingeniero Civil, MSc y PhD en Ciencias Administrativas; profesor titular de la Escuela de Administración de la Universidad del Rosario con funciones de director del Doctorado en Ciencias de la Dirección. Correo electrónico: diego.cardona@urosario.edu.co.

[**] Médico, Especialista en Salud Pública y Ph. D. en Economía y Gestión de la Salud; profesor principal de la Escuela de Administración de la Universidad del Rosario con funciones de director de la Maestría en Administración de la Salud y Especializaciones de Gerencia en la Salud. Correo electrónico: javier.gonzalez@urosario.edu.co

[***] Ingeniero de Sistemas, Especialista en Auditoría de Sistemas e Ingeniería de Software, M. Sc. en Administración; con funciones de coordinador del Laboratorio de Modelamiento y Simulación en la Escuela de Administración de la Universidad del Rosario. Correo electrónico: miller.rivera@urosario.edu.co

[****] Ingeniero Electrónico, Especialista en Educación Matemática, aspirante a Magister en Educación, con funciones de docencia en la Secretaría de Educación Distrital de Bogotá y Corporación Unificada Nacional de Educación Superior. Correo electrónico: edwin_cardenas@cun.edu.co

Introducción

En la mayoría de las investigaciones –sin importar el campo del conocimiento en las que se desarrollen– en las cuales se realicen mediciones, observaciones o experimentos de donde se obtengan datos de diferentes variables; es fundamental determinar algún tipo de relación de dependencia entre las variables con el fin de hacer predicciones o pronósticos de eventos futuros de acuerdo con el comportamiento de ellas. Por ejemplo, existen un gran número de estudios en administración donde se demuestra la relación de dependencia entre los gastos en publicidad y el volumen de ventas de cierto producto; también, en economía, se ha demostrado la relación entre la demanda u oferta de cierto producto con respecto al número de artículos que se han colocado en el mercado; y así mismo, la relación entre la variación en el precio de ese producto y la cantidad de unidades producidas.

En medicina, se han efectuado estudios de la reducción del peso de una persona en términos del número de semanas que ha seguido una dieta específica; o la cantidad de medicamento absorbido por el organismo en función del tiempo.

En otro caso, los ingenieros civiles saben que el concreto de alta calidad tiene unos componentes específicos en las concentraciones adecuadas y que la resistencia del material disminuye conforme ese concreto se mezcle con otros elementos; pero el concreto más puro aumenta muchísimo los costos de la obra, entonces al ingeniero le interesará hacer un análisis y encontrar la relación entre porcentaje de pureza del concreto y su resistencia y también la relación entre porcentaje de pureza del concreto y costo de la obra para determinar el nivel óptimo de resistencia que exige la obra sin exceder el presupuesto y sin bajar la calidad.

Algunos profesionales, sin importar su especialidad, confían en su intuición para juzgar como se relacionan dos variables. Debido a ello, hacen pronósticos a tientas e incluso temerarios; sin embargo, si dichos profesionales tienen la posibilidad de tomar datos y utilizar un procedimiento estadístico de análisis para determinar cómo lo conocido se relaciona con el evento futuro, podrían ayudar considerablemente en el mejoramiento de los procesos que administran o en la solución eficaz de los problemas que se les presentan.

El procedimiento estadístico que se utiliza para este fin se conoce como **análisis de regresión** que permite establecer la relación funcional o ecuación matemática que relaciona las variables, así como la fuerza de esa relación.

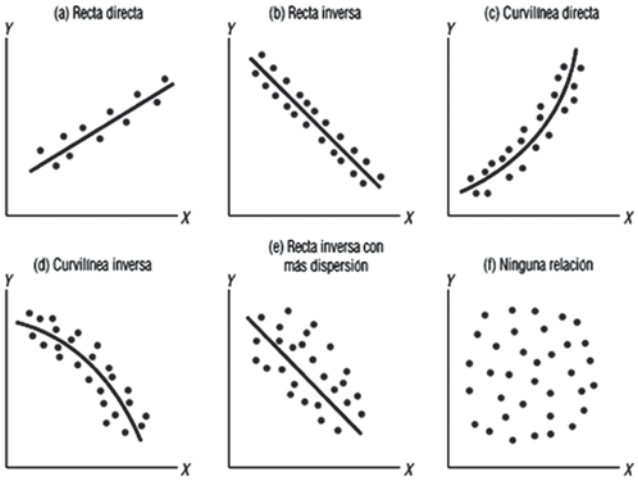
El término regresión fue utilizado por primera vez como un concepto estadístico en 1877 por sir Francis Galton, quien llevó a cabo un estudio que mostró que la estatura de los niños nacidos de padres altos tiende a retroceder o "regresar" hacia la estatura media de la población. Designó la palabra regresión como el nombre del proceso general de predecir una variable (la estatura de los niños) a partir de otra (la estatura del padre o de la madre). Más tarde, los estadísticos acuñaron el término regresión múltiple para describir el proceso mediante el cual se utilizan varias variables para predecir otra. (Devore, 2005)

En la terminología de la regresión, la variable que se va a predecir se llama dependiente, a explicar, o endógena. La o las variables que se usan para predecir el valor de la variable dependiente se llaman independientes, explicativas o exógenas.

En general, existen cuatro posibles formas en que las variables se pueden relacionar, a saber: Relación lineal directa, relación lineal inversa, relación no lineal directa y relación no lineal inversa (Figura 1), cuya estructura formal y funcional, permite dilucidar con objetividad las actividades orientadas a decidir qué ecuación se debe emplear, cuál ha

de ser la ecuación que mejor se ajusta a los datos y cómo debe validarse la significancia estadística de los pronósticos realizados.

Figura 1. Tipos de relación entre dos variables.



En este artículo se describirá el análisis de regresión donde intervienen una variable dependiente y una independiente, y en la cual la relación entre ellas se aproxima por medio de una línea recta. A esto se le llama regresión lineal simple.

Este análisis se aplicará a una situación particular en el campo de la economía.

Una aplicación del modelo de regresión lineal

Con el fin de estudiar este modelo, se emplearán los datos tomados de una muestra real, extraídos de un comunicado de prensa que revela el porcentaje de pobreza, pobreza extrema y el coeficiente de Gini (indicador de la desigualdad económica en una población) en los años 2010 y 2011 de las trece principales ciudades de Colombia.

Muchos autores que han hecho estudios sobre modelos de regresión, entre los que se pueden citar a: Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2001), Devore, J. L. (2005), Evans, M., & Rosenthal, J. S. (2005), Freund, J. E., & Simon, G. A. (1994), Levin, R. I., & Rubin, D. S. (2004) y Miller, I. (2000); coinciden en que siempre que se analizan datos observados o recopilados para

Tabla 1. Datos de pobreza en Colombia en 2010 y 2011.

Dominio	Nueva Metodología					
	Pobreza		Pobreza Extrema		Gini	
	2010	2011	2010	2011	2010	2011
Pasto	43,2	40,6	11,7	8,8	52,3	52,2
Montería	39,7	37,5	6,7	6,5	52,5	53,0
Barranquilla	39,5	34,7	7,4	5,3	49,7	47,2
Cúcuta	39,3	33,9	8,4	5,7	47,9	47,1
Cartagena	34,2	33,4	6,2	4,7	48,9	48,8
Cali	26,1	25,1	6,4	5,2	52,9	50,4
Villavicencio	25,4	23,0	4,8	4,0	46,7	46,7
Ibagué	26,6	22,0	4,3	2,7	49,5	44,9
Pereira	26,8	21,6	3,8	2,2	45,6	45,1
Manizales	23,8	19,2	4,7	2,3	49,5	47,1
Medellín	22,0	19,2	5,6	4,0	53,8	50,7
Bogotá	15,5	13,1	2,6	2,0	52,6	52,2
Bucaramanga	10,9	10,7	1,2	1,1	45,0	44,9

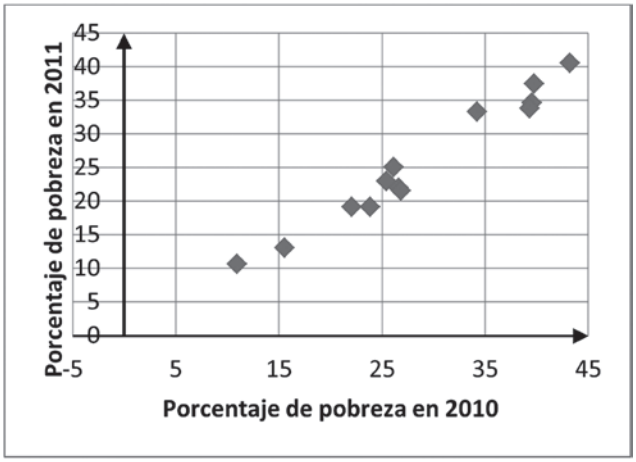
Fuente: DANE (2012)

llegar a una función o ecuación matemática que describa la relación entre las variables por medio de una regresión, se deben enfrentar tres problemas:

1. Decidir qué clase de curva muestran los puntos y por tanto qué clase de ecuación se debe usar.
2. Encontrar la ecuación particular que mejor se ajuste a los datos.
3. Demostrar que la ecuación particular encontrada cumple con ciertos aspectos referentes a los méritos de ésta para hacer pronósticos

Para decidir qué clase de función podría ajustarse a la curva, de acuerdo con las posibilidades de la figura 1, debe hacerse una gráfica de dispersión de los datos observados. Si en dicha gráfica se aprecia que los puntos se distribuyen alrededor de una recta, se procede a realizar un análisis de regresión lineal.

Figura 2. Gráfica de dispersión de los datos de pobreza en Colombia.



La gráfica de dispersión nos sugiere que existe una relación lineal entre la variable independiente porcentaje de pobreza en 2010 y la variable dependiente porcentaje de pobreza en 2011 (Figura 2)

El modelo de regresión lineal simple es:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (\mathbf{1})$$

De acuerdo con Anderson, et al. (2001), en este modelo, y es una función lineal de x (la parte $\beta_0 + \beta_1 x$) más ϵ que representa el término de error y explica la variabilidad en y que no se puede explicar con la relación lineal. A este término se le asocian los siguientes supuestos

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

Los valores de los parámetros β_0 y β_1 no se conocen y deben estimarse a partir de los datos de la muestra . Estos coeficientes que se calculan de la muestra son conocidos como regresores (b_0 y b_1). La ecuación estimada de regresión es

$$\hat{y} = b_0 + b_1 x \qquad (\mathbf{2})$$

Para calcular los regresores se emplea el método de los mínimos cuadrados el cual es un procedimiento que se remonta al inicio del siglo XIX por el trabajo del matemático francés Adrien Legendre

Criterio de los mínimos cuadrados

Este método emplea los datos de la muestra para determinar las características de la recta que hacen mínima la suma de los cuadrados de las desviaciones:

$$min \sum (y_i - \hat{y}_i)^2$$

Siendo:

y_i = valor observado de la variable dependiente para la i-ésima observación.
$\hat{y}_i$ = valor estimado de la variable dependiente para la i-ésima observación.

$$\sum(y_i - \hat{y}_i)^2 = \sum[y_i - (b_0 + b_1 x_i)]^2 \quad (3)$$

Minimizar el miembro derecho de la ecuación (3) implica calcular las derivadas parciales de la expresión con respecto a los coeficientes de regresión b_0 y b_1 e igualar a cero las dos derivadas. Al finalizar este procedimiento se llega a las siguientes ecuaciones, conocidas como ecuaciones normales. (Walpole & Myers, 1999)

Ecuaciones normales

$$\sum y_i = nb_0 + b_1 \sum x_i$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum {x_i}^2$$

Donde n es el número de observaciones.

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

Con el fin de evitar desarrollos algebraicos y aritméticos engorrosos se utilizará la hoja de cálculo EXCEL®. Al ingresar los datos de la tabla 1 y empleando la herramienta ANÁLISIS DE DATOS en la opción REGRESIÓN, se obtiene la siguiente información:

Tabla 2. Coeficientes de regresión e intervalos de confianza.

	Coef	*Err típ*	*Estad t*	*Probab*	*Inf 95%*	*Sup 95%*
Intercep	-1,31	1,530	-0,859	0,408	-4,682	2,053
Var X 1	0,941	0,051	18,6	0,000	0,830	1,053

En esta tabla los valores de intercepción y variable X1 hacen referencia a los coeficientes b_0 y b_1 respectivamente. Remplazando estos datos en la ecuación 2, se tiene que la ecuación de regresión para las variables pobreza en 2010 y pobreza en 2011 es:

$$\hat{y} = -1{,}3148 + 0{,}94126x$$

El coeficiente b_1 también corresponde a la pendiente de la recta. En general, este coeficiente expresa la razón de cambio entre la variable dependiente con respecto a un cambio unitario en la variable independiente x.

En el ejemplo, la pendiente de la recta es positiva, lo que implica que en las ciudades donde se observó mayor pobreza en 2010, también se observó mayor pobreza en 2011. Pero como la pendiente es un número entre cero y uno, significa que el incremento en el porcentaje de pobreza en 2011 entre una ciudad y otra es menor que en el 2010.

Con respecto a los casos enunciados en la introducción, se puede decir que si la relación entre los gastos en publicidad y el volumen de ventas del producto es lineal; el regresor estimado b_1 indicaría la cantidad en que se incrementan las ventas por cada unidad monetaria en que se incremente el gasto en publicidad.

Así mismo, si la relación entre la cantidad de masa que la persona pierde y el número de semanas que sigue la dieta es lineal; el coeficiente b_1 indicaría el peso perdido por semana y con ello el nutricionista podría predecir la cantidad de masa que perdería el paciente en un cierto número de semanas o determinar la cantidad de semanas necesarias para que el paciente pierda el peso deseado.

Análisis de regresión

Con el fin de determinar la pertinencia de la ecuación de regresión hallada, es necesario hacer un análisis de la bondad de ajuste de la recta, demostrar si la relación es estadísticamente

significativa y validar los supuestos acerca del término de error.

Para ello se deben calcular los siguientes estadísticos:

El coeficiente de determinación:

Es una medida de la bondad de ajuste para una ecuación de regresión.

Para la i-ésima observación de la muestra, la desviación entre el valor observado de la variable dependiente y_i y el valor estimado de la variable dependiente $\hat{y}_i$, se llama i-ésimo residual. Representa el error que se comete al usar $\hat{y}_i$ para estimar y_i.

La suma de los cuadrados de esos residuales es lo que se minimiza en el método de mínimos cuadrados. También se le conoce como la suma de los cuadrados debidos al error (SSE)

$$SSE = \sum(y_i - \hat{y}_i)^2$$

El valor de SSE es una medida del error que se comete al usar la ecuación de regresión para calcular los valores de la variable dependiente en la muestra.

Otro valor de importancia es la medida del error incurrido al usar para estimar y_i, llamado suma total de cuadrados (SST):

$$SST = \sum(y_i - \bar{y})^2$$

Para saber cuánto se desvían los valores de $\hat{y}_i$ medidos en la línea de regresión, de los valores de $\bar{y}$, se calcula otra suma de cuadrados. A esa suma se le llama suma de cuadrados debida a la regresión, y se representa por SSR.

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

Existe una relación entre las tres sumas:

$$SST = SSR + SSE \quad (4)$$

Ahora bien, es posible entender cómo se pueden emplear las tres sumas de cuadrados para suministrar una medida de la bondad de ajuste para la ecuación de regresión.

Esa ecuación tendría un ajuste perfecto si cada valor observado de la variable independiente estuviera sobre la línea de regresión. En este caso cada diferencia $y_i - \hat{y}_i$ sería cero, por tanto SSE=0. De la ecuación (4) se tendría que SST=SSR y por consiguiente la relación SSR/SST sería igual a 1 como el máximo ajuste. De manera análoga, los ajustes menos perfectos darán como resultado mayores valores de SSE. En consecuencia, de (4) se deduce que el máximo valor de SSE se tiene cuando SSR es cero.

La relación SSR/SST, se denomina coeficiente de determinación y se representa por r^2.

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

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

Para el caso que se está analizando, el programa EXCEL® entrega también la siguiente información

Tabla 3. Estadísticos de la regresión.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,984467
Coeficiente de determinación R^2	0,969175
R^2 ajustado	0,9663728
Error típico	1,73579
Observaciones	13

Análisis de varianza		
	Grados de libertad	*Suma de cuadrados*
Regresión	1	1042,046578
Residuos	11	33,14265228
Total	12	1075,189231
Promedio de los cuadrados	F	Valor crítico de F
1042,046578	345,854	1,16437E-09
3,012968389		

La tabla 3 muestra el valor de SSE, SSR Y SST en la columna que indica la suma de cuadrados, de alli se obtiene el coeficiente r^2 que aparece (0,969175). Esto revela que la ecuación de regresión explica en un 96,92% los valores observados de la pobreza en 2011 según los valores de pobreza en 2010.

En la mayoría de situaciones prácticas no es común obtener coeficientes de determinación tan altos, pero existen valores aceptables que varían de acuerdo con la rama del conocimiento sobre el que se verse el estudio o investigación.

Coeficiente de correlación:

Es la segunda medida que se usa para describir qué tan bien explica una variable a la otra. El coeficiente de correlación de la muestra se denota por r y es la raíz cuadrada del coeficiente de determinación:

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

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

La tabla 3 muestra un coeficiente de correlación muy alto (r = 0,9845), lo que implica una relación de dependencia lineal muy fuerte entre los valores de pobreza de 2010 y 2011 en la principales ciudades de Colombia.

Es importante resaltar que **el coeficiente de correlación solo mide la fuerza de asociación en una relación lineal**, el coeficiente de determinación se puede usar en relaciones no lineales y en relaciones con dos o más variables independientes. En este sentido, el coeficiente de determinación tiene mayor aplicabilidad. (Walpole & Myers, 1999)

Los coeficientes de determinación y correlación no son suficientes para llegar a la conclusión acerca de si la relación es estadísticamente significativa. Esa conclusión se debe basar en consideraciones donde intervenga el tamaño de la muestra y las propiedades de las distribuciones muestrales adecuadas de los estimadores de los mínimos cuadrados.

Pruebas de significancia

La ecuación de regresión lineal simple indica que el valor medio esperado de y es una función lineal de x:

$$E(y) = \beta_0 + \beta_1 x$$

Si β_1=0, entonces E(y)=β_0. En este caso el valor medio de y no depende del valor de x y se concluye que no existe relación lineal entre las variables. En forma análoga, si el valor de β_1 no es igual a cero, se concluye que las dos variables se relacionan. Así, para probar si hay alguna relación importante de regresión debemos efectuar una prueba de hipótesis para determinar si el valor de β es cero. Existen dos pruebas que se usan con más frecuencia y para ellas se necesita un estimado de la varianza del error en el modelo de regresión.

Estimado de σ^2

La varianza de ϵ, también representa la varianza de los valores de y respecto a la línea de regresión. Así, la suma de los residuales al cuadrado, SSE, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión. Cada suma de cuadrados tiene asociado un número que

llamamos grados de libertad. Se ha demostrado que SSE tiene n–2 grados de libertad, porque se deben estimar dos parámetros β_0 y β_1.

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

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

Desviación estándar de la estimación

El error típico o desviación estándar del estimado se calcula como la raíz cuadrada de la varianza del estimado.

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

La tabla 3 muestra un error típico de 1,73579

Las pruebas de significancia que se efectúan son: la prueba t y la prueba F.

La explicación detallada acerca de la obtención de estos estadísticos de prueba t y F se encuentra en Rivera, M., & Cárdenas V., E. (2013)

La tabla 3 muestra el valor del estadístico de prueba F y la tabla 2 el valor del estadístico de prueba t para b_1. En ambos casos el valor de probabilidad o valor crítico es practicamente cero lo cual implica que la relación es estadísticamente significativa.

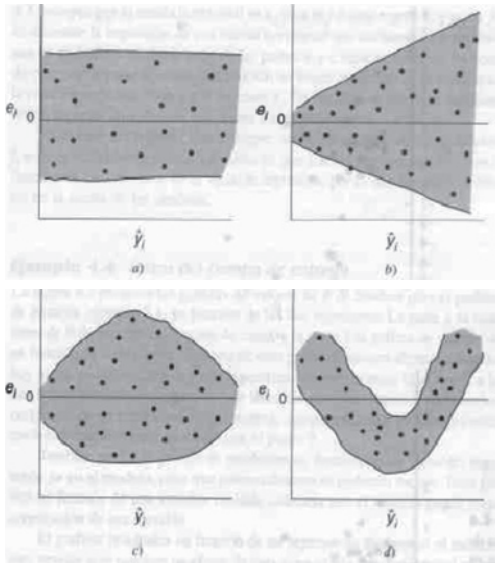
Análisis de Residuales

Este análisis permite validar los supuestos del modelo con respecto al error y se basa en el examen de varias gráficas a saber:

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

En una gráfica de residuales se puede presentar alguno de estos patrones.

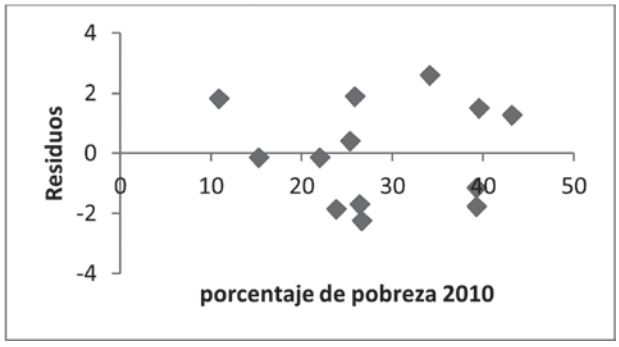
Figura 3. Patrones de una gráfica de residuales.



Si la relación de dependencia hallada cumple la hipótesis de que la varianza de es igual para todos los valores de x y si el modelo de regresión lineal es una representación adecuada de la relación entre las variables; entonces, la gráfica debe mostrar un patrón muy similar a una franja horizontal de puntos (figura 3a).

Para el caso que se está analizando, la herramienta de regresión de EXCEL® da la posibilidad de mostrar este tipo de gráficas y hacer un análisis de residuales

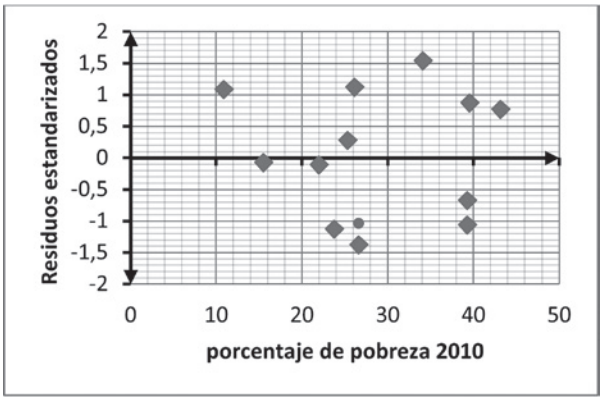
Figura 4. Gráfica de residuales.



Esta gráfica no muestra un patrón que haga dudar sobre la hipótesis de que la varianza es constante.

La gráfica de residuos estandarizados muestra el mismo patrón que la anterior y se usa para observar la existencia de valores atípicos o influyentes. Si el error se distribuye en forma normal, los residuos deben estar en el rango de dos desviaciones estándar.

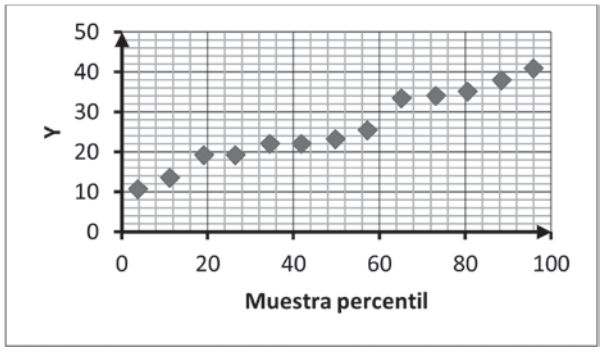
Figura 5. Gráfica de residuales estandarizados.



En la gráfica se observa que todos los residuales se encuentran en el intervalo de dos desviaciones estándar.

La gráfica de probabilidad normal, también se usa para validar el supuesto de que el error tiene una distribución normal. Esta gráfica debe mostrar una recta.

Figura 6. Gráfico de probabilidad normal.



Uso de la ecuación de regresión para estimar y predecir

Si el análisis de la ecuación de regresión obtenida con los datos demuestra que existe una relación estadísticamente significativa entre las variables, y si el ajuste que proporciona la ecuación es bueno, esa ecuación podría usarse para estimaciones y predicciones.

Estimación de intervalo

Al hacer una estimación puntual de y un valor de dado un valor de x, no se tiene idea alguna de la precisión asociada con el valor estimado.

Con ese fin, se determinan estimaciones de intervalo. El primer tipo de estimado es el de *intervalo de confianza*, que es un estimado del valor medio de y para determinado valor de x. El segundo tipo es el *estimado de intervalo de predicción*, que se usa cuando deseamos un estimado de intervalo de valor individual de y que corresponda a determinado valor de x. Con la estimación puntual se obtiene el mismo valor, sea que estemos estimando el valor medio de y o prediciendo un valor individual de y , pero con los estimados de intervalo se obtienen valores distintos. (Freund & Simon, 1994)

Estimado del intervalo de confianza del valor medio de y

Al estimar el porcentaje promedio de pobreza en 2011 de todas las ciudades que en 2010 mostraron un índice de pobreza de 25,3%. El estimado de $E(y_p)$, el valor medio desconocido, es:

$$\hat{y}_p = -1{,}3148 + 0{,}94126(25{,}3) = 22{,}5$$

Donde $\hat{y}_p$ es el estimado del valor particular de y.

Dado que no se puede esperar que $\hat{y}_p$ sea exactamente igual a $E(yp)$. Entonces es necesario

considerar la varianza de los estimados basados en la ecuación de regresión. La fórmula para estimar la desviación estándar de $\hat{y}_p$ dado un valor particular de x , x_p, es:

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \qquad (5)$$

La ecuación general para un estimado del intervalo de confianza de $E(y_p)$ dado un valor particular de x es:

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{\hat{y}_p}$$

En donde el coeficiente de confianza es $1-\alpha$ y $t_{\alpha/2}$ se basa en una distribución t con n–2 grados de libertad.

Para determinar un estimado de intervalo de confianza de 95% para el porcentaje promedio de pobreza en 2011 de todas las ciudades que en 2010 mostraron un índice de pobreza de 25,3%, necesitamos el valor de t para $\alpha/2=0.025$ y n–2= 11 grados de libertad. Así, con $\hat{y}_p$ =22,5 $t_{0.025}$=2,201 y $s_{\hat{y}_p}$ =0,5111, tenemos:

$$22{,}5 \pm 2{,}201 \cdot 0{,}5111$$

$$22{,}5 \pm 1{,}125$$

Entonces, con una confianza del 95% se puede decir que el porcentaje promedio de pobreza en 2011 de todas las ciudades que en 2010 mostraron un índice de pobreza de 25,3% está entre 21,375% y 23,625%.

Obsérvese que la desviación estándar estimada de x_p expresada en la ecuación (5) es mínima cuando $x_p=\bar{x}$. Esto implica que se puede hacer el mejor estimado, o el más preciso, del valor medio de siempre que se use el valor medio de x.

En la estimación y la inferencia, un error común es suponer que la línea de regresión, así el ajuste sea muy bueno (valor de r^2 muy alto), puede aplicarse en cualquier intervalo de valores. Aun cuando una relación se cumpla para el intervalo de puntos de la muestra, puede existir una relación completamente distinta para un intervalo diferente. Por ejemplo, la relación gastos en publicidad y volumen de ventas puede ser lineal para cierto intervalo de gasto o inversión publicitaria pero en la medida que el público conozca del producto y se sature de publicidad esa relación ya no será lineal, pues las ventas no se incrementarán en la misma medida que incrementa la inversión en publicidad.

Una ecuación de estimación es válida para el mismo rango dentro del cual se tomó la muestra inicialmente (Levin & Rubin, 2004). Sin embargo, si el investigador tiene la certeza de que el comportamiento entre las variables será el mismo en otros intervalos fuera del rango de la muestra, entonces puede usar la ecuación para hacer predicciones.

Estimado del intervalo de predicción para un valor particular de y

Para este análisis, se supone que en vez de estimar el valor medio del porcentaje de pobreza, deseamos estimar el porcentaje de pobreza en 2011 para la ciudad de Armenia con un índice de pobreza de 25,3% en 2010.

El estimado para ese valor particular por medio de la ecuación de regresión es:

$$\hat{y}_p = -1{,}3148 + 0{,}94126(25{,}3) = 22{,}5$$

Que es el mismo valor que el estimado puntual para el porcentaje promedio.

Para determinar un estimado del intervalo de predicción debemos determinar primero la varianza asociada al empleo de $\hat{y}_p$ como estimado de un valor individual de y. Esta varianza está formada por la suma de dos componentes:

1. La varianza de los valores individuales de y respecto del promedio cuyo estimado es s^2
2. La varianza asociada al uso de $\hat{y}_p$ para estimar $E(y_p)$ cuyo estimado es $s_{\hat{y}_p}$.

Así, el estimado de la varianza de un valor individual es:

$$s_{ind}^2 = s^2 + s_{\hat{y}_p}$$

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

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

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

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{ind}$$

En donde el coeficiente de confianza es 1–α y $t_{\alpha/2}$ se basa en una distribución t con n–2 grados de libertad.

Para determinar un estimado de intervalo de predicción de 95% para el porcentaje de pobreza en 2011 de la ciudad de Armenia que en 2010 mostró un índice de pobreza de 25,3%, se necesita el valor de t para α/2=0.025 y n–2= 11 grados de libertad. Así, con $\hat{y}_p$=22,5, $t_{0.025}$=2,201 y s_{ind} = 1,8095, se tiene:

$$22{,}5 \pm 2{,}201 \cdot 1{,}8095$$

$$22{,}5 \pm 3{,}9827$$

Entonces, con una confianza del 95% se puede decir que el porcentaje de pobreza en 2011 de la ciudad de Armenia que en 2010 tenía un porcentaje de pobreza de 25,3% está entre 18,52% y 26,48%.

De acuerdo con lo anterior, el intervalo de predicción es mayor que el intervalo de confianza.

Estimación de los parámetros del modelo de regresión lineal

Uno de los conceptos fundamentales sobre el que se ha basado este análisis, es que la ecuación de regresión lineal obtenida a partir de los datos de la muestra es un estimado de los parámetros del modelo para la población. Por lo tanto, es posible determinar intervalos de confianza para los coeficientes de la ecuación de regresión:

$$\beta_0 = b_0 \pm t_{\alpha/2} \cdot s\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

$$\beta_1 = b_1 \pm t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}}$$

La tabla 2 muestra los valores inferior y superior para el intervalo de confianza del 95% de los parámetros del modelo, de tal forma que:

$$\beta_1 = -1{,}3148 \pm 3{,}3675$$

$$\beta_0 = 0{,}94126 \pm 0{,}1114$$

Con esta información, el investigador encuentra que la tasa de incremento de la pobreza en 2011 está entre 0,83% y 1,053% por cada 1% de incremento de la pobreza en 2010 entre una ciudad y otra. Además con las predicciones y estimaciones hechas puede hacer conclusiones sobre la situación de pobreza esperada en las principales ciudades del país y si este investigador tiene un cargo de responsabilidad e influencia en el gobierno colombiano, puede generar propuestas para disminuir significativamente esos índices.

Finalmente, es importante mencionar que se puede cometer otro error al utilizar el análisis de regresión, y es suponer que un cambio en una variable es "ocasionado" por un cambio en la otra variable. Los análisis de regresión y correlación no pueden, de ninguna manera, determinar la causa y el efecto. Si se dice que existe una relación lineal

entre el número de canas y de arrugas que van apareciendo en una persona, no se puede decir que una ocasiona la otra pues es muy posible que existan otras variables asociadas que sean la causa; en este caso la edad de la persona, por ejemplo.

La validez de una conclusión de tipo causa y efecto requiere de una justificación teórica, o del buen juicio por parte del analista. (Anderson, Sweeney, & Williams, 2001)

Conclusiones

El análisis de regresión lineal simple, como parte de la inferencia estadística, es fundamental para determinar relaciones de dependencia lineal entre variables y establecer su validez con el fin de hacer estimaciones y predicciones dentro de un intervalo de confianza deseado.

Obtener una ecuación de regresión que describe el comportamiento lineal entre dos variables permite pronosticar valores futuros de la variable bajo análisis con cierto grado de certeza, lo cual constituye una herramienta poderosa pues le da al profesional la posibilidad de hacer ajustes en los procesos, tomar decisiones o establecer políticas. Por ejemplo, si un profesional en ciencias políticas o administración pública utiliza el estudio sobre índices de pobreza realizado con los datos de las trece principales ciudades del país y concluye que los valores observados y estimados están por debajo de la media en América Latina o que están por debajo de la meta nacional; podría establecer un programa que disminuya en forma eficaz esos índices de pobreza.

Así mismo, si un administrador o economista realiza el análisis sobre la relación de dependencia entre el gasto en publicidad y el volumen de ventas de un producto podría determinar la inversión óptima en publicidad para ese producto y obtener el máximo de ventas o predecir la cantidad de unidades vendidas de acuerdo con un valor invertido en publicidad.

A pesar de lo importante que resulta ser para cualquier profesional el conocimiento y uso del análisis de regresión, es una herramienta muy poco aprovechada como lo demuestran un gran número de trabajos de grado a nivel de posgrado y trabajos de investigación en los cuales el desarrollo estadístico solo se limita a la parte descriptiva y no a la inferencial.

Referencias

Anderson, D., Sweeney, D. & Williams, T. (2001). *Estadística para administración y economía* (7a ed., Vol. II). México: Thomson.

DANE. (17 de mayo de 2012). "Pobreza en Colombia". Comunicado de prensa, 6.

Devore, J. L. (2005). *Probabilidad y estadística para ingeniería y ciencias*. (6a ed.). México: Thomson Learning.

Evans, M. & Rosenthal, J. (2005). *Probabilidad y estadística. La ciencia de la incertidumbre*. Barcelona: Reverté.

Freund, J. & Simon, G. (1994). *Estadística elemental*. (8a ed.). México: Prentice Hall.

Levin, R. & Rubin, D. (2004). *Estadística para administración y economía*. México: Pearson Educación.

Lopera, C. (2002). Análisis de residuales. Universidad Nacional de Colombia. Disponible en: http://www.docentes.unal.edu.co/cmlopera/docs/Estad2/2_RLM/2.(Complemento)Análisis de Residuales y Otros en RLM.pdf

Mendoza, H., Vargas, J., López, L. & Bautista, G. (2002). Métodos de regresión. Bogotá: Universidad Nacional de Colombia. Disponible en: http://www.virtual.unal.edu.co/cursos/ciencias/2007315/

Miller, I. (2000). *Estadística matemática con aplicaciones*. (6a ed.). México: Pearson Educación.

Muñoz, R. (2006). Comprobación de los supuestos del modelo de regresión lineal. Cali: Universidad Autónoma de Occidente. Disponible en: http://augusta.uao.edu.co/moodle/file.php/284/18_supuestos_de_la_regresion_lineal.pdf

Pacheco, P. (2012). Validación de supuestos. Bogotá. Universidad Nacional de Colombia. Disponible en: http://www.virtual.unal.edu.co/cursos/ciencias/dis_exp/und_3/pdf/validaciondesupuestosunidad 3b[1].pdf

Vilar, J. (2006). Identificación de valores atípicos y observaciones influyentes. La Coruña: Universidad de La Coruña. Disponible en: http://www.udc.es/dep/mate/estadistica2/sec4_6.html

Walpole, R. & Myers, R. (1999). *Probabilidad y estadística para ingenieros*. (6a ed.). México: Prentice Hall.