UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2014, Vol. 7, No. 1, pp. 89–98

Transformaciones logarítmicas en regresión simple

Logarithmic transformations in simple regression analysis

Jorge Ortiz Pinilla[a]
jorgeortiz@usantotomas.edu.co

Diana Gil[b]
dianagil@usantotomas.edu.co

Resumen

En este artículo se investiga los efectos de las transformaciones logarítmicas en un análisis de regresión simple. En la práctica, es muy común que los parámetros de los modelos conocidos como *exponencial* y *potencial* se estimen de manera habitual mediante una transformación logarítmica, que los reduce a modelos lineales y se "regresa" al modelo original aplicando la función exponencial a la estimación del intercepto. En este trabajo se encuentra que este procedimiento no genera estimadores de mínimos cuadrados para el modelo inicial e introduce variaciones en la forma como se conciben las relaciones entre las variables. La popularidad de las herramientas de análisis hace que el riesgo de utilizar modelos que no correspondan a los datos pase desapercibido.

***Palabras clave*:** modelo exponencial, modelo potencial, mínimos cuadrados, regresión no lineal, modelos de regresión.

Abstract

In this paper the effect of the logarithmic transformations in simple regression analysis is investigated. In practice, it is very common that exponential and power models' parameters are estimated by means of a logarithmic transformation which reduces them to a linear form. The estimations in the initial models are obtained by applying the exponential function to the intercept estimation. In this work, it is found that this procedure does not generate least squares solutions for the initial model and introduces variations in the way in which relationships between variables are conceived. Because of the popularity of software tools, the risk of using inappropriate models for the data may be unnoticed.

***Keywords*:** exponential model, power model, least squares, non linear regression, regression models.

[a]Docente. Facultad de Estadística, Universidad Santo Tomás, Colombia.
[b]Estudiante, Carrera de Estadística, Universidad Santo Tomás, Colombia.

1. Introducción

Una práctica común en las aplicaciones de los métodos de regresión consiste en buscar transformaciones que permitan construir modelos lineales para describir las relaciones entre las variables. La mayoría de los textos básicos hacen esta recomendación y dan por resuelto el problema. Por ejemplo, Mendenhall & McClave (1981, p. 259) escriben

> When the transformed model is used to predict the value of $\log y$, the predicted value of y is the antilog, $\widehat{y} = e^{\widehat{\log y}}$.

Walpole et al. (2012), en el ejemplo 11.9 de la página 426, utilizan el mismo procedimiento de transformar con logaritmos tanto la *presión* como el *volumen* de un gas para estudiar empíricamente la ley del gas ideal. Después de obtener los coeficientes del modelo transformado, calculan la función exponencial al intercepto para "regresar" a la forma original del modelo potencial.

Las referencias anteriores han tenido un alto impacto en la enseñanza de la estadística en carreras universitarias como ingeniería, física, química y economía. Una de ellas data de 1981 y la otra de 2012. Durante este periodo, la estadística se ha consolidado como herramienta de uso cotidiano y masivo entre los investigadores, gracias al desarrollo de las computadoras personales y a la disponibilidad de *software* que incorpora procedimientos de análisis de datos. Por otra parte, las hojas electrónicas y las calculadoras científicas que incluyen análisis de regresión aplican el procedimiento descrito como la única opción: se transforma el modelo en uno lineal, se obtienen las estimaciones de los parámetros por el método de mínimos cuadrados y se reconstruye el modelo original aplicando la transformación inversa (exponencial) a los elementos que corresponda.

En estas circunstancias, el analista utiliza las herramientas y obtiene resultados sin ninguna señal de alerta que le advierta sobre el riesgo de tomar decisiones, con base en modelos que no describan en forma adecuada las tendencias de la nube de puntos. La popularidad de estas herramientas hace masivo el riesgo.

Por tratarse de la función logarítmica que es una transformación estrictamente monótona creciente, efectivamente el modelo transformado es equivalente al modelo original. Esto garantiza una interpretación adecuada de los coeficientes con el debido cuidado de las transformaciones requeridas.

No sucede lo mismo con las estimaciones de los parámetros. Unas resultan de minimizar la suma de cuadrados de los errores del modelo en las unidades originales utilizadas para tomar los datos, y otras, en unidades logarítmicas que atribuyen menor importancia a las diferencias entre los valores más grandes de la variable. Como consecuencia, el método de mínimos cuadrados aplicado al modelo transformado no produce estimaciones de mínimos cuadrados para el modelo original. Por lo tanto, el resultado obtenido puede ser inadecuado para pronosticar la respuesta esperada a partir de valores específicos de la variable X.

El propósito de este artículo es comparar los métodos que se utilizan para obtener las estimaciones de mínimos cuadrados de los modelos exponencial y potencial de manera directa con los que se basan en transformaciones logarítmicas. Como criterio de comparación se toma la suma de cuadrados residual, como indicador de la bondad del ajuste del modelo a los datos observados.

2. Modelo exponencial

Cuando el modelo planteado es de la forma

$$y = \beta_0 e^{\beta_1 x} \tag{1}$$

las estimaciones de mínimos cuadrados se obtienen buscando b_0 y b_1 correspondientes al menor valor de la función

$$g(b_0, b_1) = \sum_{i=1}^{n} \left(y_i - b_0 e^{b_1 x_i}\right)^2 \tag{2}$$

Se deriva $g(b_0, b_1)$ con respecto a b_0 y a b_1 y luego se iguala a cero cada derivada:

$$\frac{\partial g(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^{n} \left(y_i - b_0 e^{b_1 x_i}\right) e^{b_1 x_i}$$

Entonces:

$$b_0 = \frac{\sum\limits_{i=1}^{n} y_i e^{b_1 x_i}}{\sum\limits_{i=1}^{n} e^{2 b_1 x_i}} \tag{3}$$

Haciendo lo mismo para b_1,

$$\frac{\partial g(b_0, b_1)}{\partial b_1} = -2 \sum_{i=1}^{n} \left(y_i - b_0 e^{b_1 x_i}\right) b_0 e^{b_1 x_i} x_i$$

$$\sum_{i=1}^{n} x_i y_i e^{b_1 x_i} - b_0 \sum_{i=1}^{n} x_i e^{2 b_1 x_i} = 0$$

Reemplazando b_0 por la expresión obtenida en (3), se llega a la siguiente ecuación que solo tiene b_1 como incógnita:

$$\sum_{i=1}^{n} x_i y_i e^{b_1 x_i} - \left(\frac{\sum y_i e^{b_1 x_i}}{\sum e^{2 b_1 x_i}} \right) \sum_{i=1}^{n} x_i e^{2 b_1 x_i} = 0 \tag{4}$$

La complejidad de esta ecuación solo permite encontrar sus soluciones por métodos numéricos. Si las denotamos como $\tilde{\beta}_0$ y $\tilde{\beta}_1$, el modelo ajustado por mínimos cuadrados directos es:

$$\tilde{y} = \tilde{\beta}_0 \, e^{\tilde{\beta}_1 \, x} \tag{5}$$

Las estimaciones mediante la transformación logarítmica se obtienen llevando el modelo (3) al equivalente:

$$y^* = \beta_0^* + \beta_1^* \, x^* \tag{6}$$

en donde

$$y^* = \ln(y), \quad \beta_0^* = \ln(\beta_0), \quad \beta_1^* = \beta_1 \, , \quad x^* = x \tag{7}$$

Como (6) es un modelo lineal, las estimaciones de β_0 y β_1 son:

$$b_1^* = \frac{cov(x^*, y^*)}{var(x^*)} \tag{8}$$

$$b_0^* = \overline{y^*} - b_1^* \, \overline{x^*} \tag{9}$$

Según las sugerencias de los autores citados, se "regresa" al modelo original (1) aplicando las transformaciones inversas acordes con (7):

$$y = e^{y^*}, \quad b_0 = e^{b_0^*}, \quad b_1 = b_1^* \, , \quad x = x^* \tag{10}$$

es decir,

$$\widehat{y} = e^{b_0^*} \, e^{b_1^* \, x} \tag{11}$$

Los dos procedimientos proveen soluciones diferentes. Resulta claro que si el primero es de mínimos cuadrados para el modelo original, el segundo no lo es. Por lo tanto, si se pasa al plano inferencial, los estimadores de los parámetros del modelo exponencial, obtenidos mediante la transformación logarítmica no son de mínimos cuadrados para el modelo original.

El siguiente ejemplo sirve para ilustrar la situación planteada:

Ejemplo 2.1. *Los siguientes datos fueron obtenidos de un modelo de la forma (1):*

x	y	x	y	x	y
6.7	*77.4*	*7.2*	*38.3*	*16.1*	*743.4*
14.9	*440.2*	*11.3*	*101.6*	*4.7*	*38.9*
7.0	*34.0*	*14.7*	*457.7*	*7.6*	*9.8*
5.2	*119.8*	*7.7*	*4.1*	*13.8*	*234.5*
7.6	*102.6*	*8.3*	*24.9*	*18.7*	*2367.9*
18.7	*2287.0*	*17.0*	*1186.4*	*11.8*	*167.8*
11.4	*177.3*	*10.8*	*109.5*	*5.3*	*24.2*
9.5	*65.0*	*18.1*	*1818.2*	*17.0*	*1201.4*
17.1	*1273.1*	*12.0*	*149.2*	*19.2*	*2892.6*
8.5	*124.1*	*9.3*	*94.5*	*12.0*	*135.4*

En la gráfica 1, el modelo construido con la transformación logarítmica se dibuja con trazos discontinuos y el obtenido por mínimos cuadrados directos con una curva continua. Desde el punto de vista de los datos, el primero presenta un desajuste importante en los valores más grandes y no describe adecuadamente la tendencia de la nube de puntos.

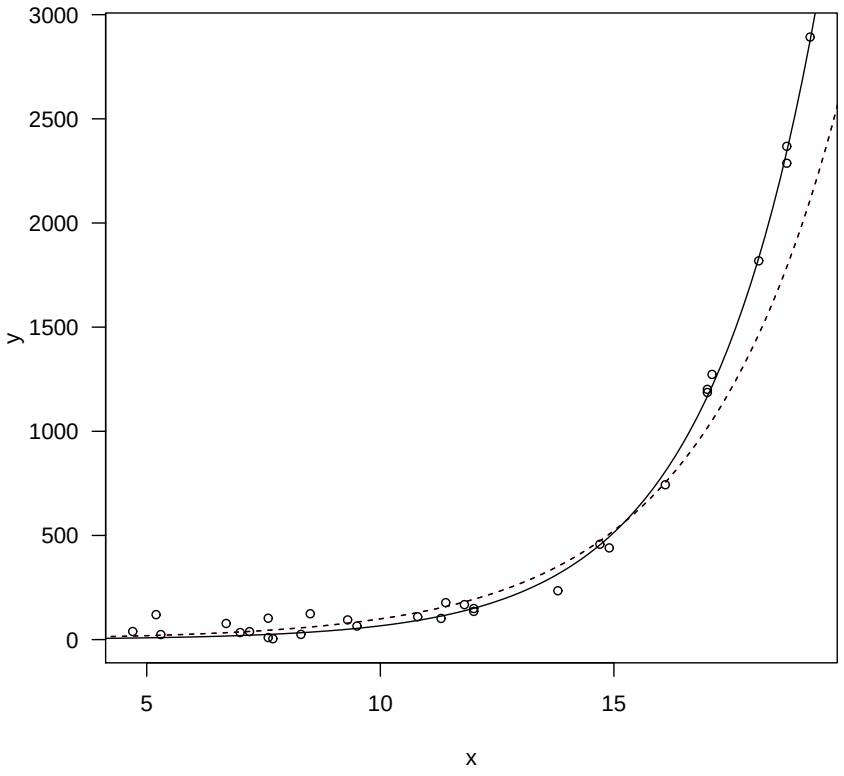

Figura 1: *Ajuste de un modelo exponencial por mínimos cuadrados directos (línea continua) y por linealización mediante transformación logarítmica de la variable Y (línea discontinua). Fuente: elaboración propia.*

Las estimaciones y las sumas de cuadrados residuales en la tabla siguiente muestran diferencias importantes en estos valores. En particular, la suma de cuadrados residual del modelo estimado por transformación logarítmica es más de 24 veces la de mínimos cuadrados.

	b0	*b1*	*Suma de cuadrados residual*
Mínimos cuadrados	*1.098125*	*0.4099219*	*61709.12*
Transformación Log.	*3.598670*	*0.3319955*	*1484642.58*

En el modelo exponencial los errores son de la forma:

$$\varepsilon = Y - \beta_0 e^{\beta_1 x} \tag{12}$$

mientras que en el modelo transformado son:

$$\begin{aligned} \varepsilon &= \ln(Y) - \ln(\beta_0 e^{\beta_1 x}) \\ &= \ln\Big(\frac{Y}{\beta_0 e^{\beta_1 x}}\Big) \end{aligned} \tag{13}$$

Por otra parte, el supuesto de normalidad de los errores trae consecuencias muy diferentes para los dos procedimientos. En el caso de los mínimos cuadrados directos, los errores son de carácter aditivo para Y y $Y \sim N(\beta_0 e^{\beta_1 x}, \sigma^2)$. En el modelo transformado, son aditivos para $\ln(Y)$, es decir, multiplicativos para Y. Si se asume que $\varepsilon \sim N(0, \sigma^2)$, entonces de (13) se deduce que $\dfrac{Y}{\beta_0 e^{\beta_1 x}}$ tiene distribución *log-normal* con valor esperado $e^{\sigma^2/2}$ y varianza $e^{\sigma^2}(e^{\sigma^2} - 1)$. Por lo tanto, la distribución de Y bajo el modelo transformado es *log-normal* con media $\beta_0 e^{\beta_1 x + \sigma^2/2}$ y varianza $(e^{\sigma^2} - 1)e^{2\ln\beta_0 + 2\beta_1 x + \sigma^2}$.

Es claro que, dependiendo del procedimiento que se utilice, se ajustan modelos diferentes en cuanto al papel que cumplen los errores y a los supuestos acerca de su distribución, y en cuanto a las consecuencias que traen sobre la distribución condicional de la variable dependiente.

3. Modelo potencial

El modelo se llama *potencial* cuando la relación entre las variables es de la forma:

$$y = \beta_0 x^{\beta_1} \tag{14}$$

Igual que para el modelo exponencial, las estimaciones de mínimos cuadrados se obtienen buscando b_0 y b_1 correspondientes al menor valor de la función

$$\begin{aligned} g(b_0, b_1) &= \sum_{i=1}^{n} \big(y_i - \widehat{y_i}\big)^2 \\ &= \sum_{i=1}^{n} \big(y_i - b_0 x_i^{b_1}\big)^2 \end{aligned} \tag{15}$$

Se aplica el método tradicional de derivarla con respecto a b_0 y a b_1 y luego igualar a 0 cada derivada:

$$\frac{\partial g(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^{n} \; bigl(y_i - b_0 x_i^{b_1}\big) x_i^{b_1} \tag{16}$$

Entonces:

$$b_0 = \frac{\sum y_i x_i^{b_1}}{\sum x_i^{2b_1}} \tag{17}$$

Haciendo lo mismo para b_1,

$$\frac{\partial g(b_0, b_1)}{\partial b_1} = -2\sum_{i=1}^{n}\left(y_i - b_0 x_i^{b_1}\right) b_0 x_i^{b_1} \log(x_i) \tag{18}$$

$$\sum_{i=1}^{n} y_i x_i^{b_1} \log(x_i) - b_0 \sum_{i=1}^{n} x_i^{2b_1} \log(x_i) = 0 \tag{19}$$

Reemplazando b_0, se obtiene la siguiente ecuación que se resuelve por métodos numéricos para encontrar el valor de b_1.

$$\sum_{i=1}^{n} y_i x_i^{b_1} \log(x_i) - \left(\frac{\sum y_i x_i^{b_1}}{\sum x_i^{2b_1}}\right) \sum_{i=1}^{n} x_i^{2b_1} \log(x_i) = 0 \tag{20}$$

Ejemplo 3.1. *Los datos siguientes son utilizados por Walpole et al. (2012, ejemplo 11.9, p.420) para ilustrar el uso de la regresión potencial. Según la ley del gas ideal, $PV^{\gamma} = C$, donde P es la presión, V es el volumen y C y γ son constantes por estimar. En el ejemplo, P es la variable dependiente y V es la variable independiente. C asume el papel de β_0 y γ el de β_1 en el modelo potencial y sus estimaciones se denotan como b_0 y b_1.*

x (Volumen)	*50*	*60*	*70*	*90*	*100*
y (Presión)	*64.7*	*51.3*	*40.5*	*25.9*	*7.8*

	b0	*b1*	*Suma de cuadrados residual*
Lineal	*116.1616*	*-1.055698*	*37.53616*
Mín.Cuadr	*112451.3806*	*-1.894926*	*164.33431*
Transf.Log	*2568862.8877*	*-2.653472*	*399.26979*

Aunque las diferencias en las sumas de cuadrados residuales no son tan grandes como en el ejemplo de la regresión exponencial, la obtenida con el procedimiento de la transformación logarítmica es más del doble de la de mínimos cuadrados directos.

Se incluyó un ajuste lineal que curiosamente arroja una suma de cuadrados residual menor que las de los modelos potenciales. Este resultado sirve para advertir que no siempre el mejor ajuste corresponde a la respuesta más adecuada. La orientación del análisis estadístico debe tener sus bases en los aspectos teóricos de la disciplina respectiva. Por otra parte, los puntos observados son seguramente insuficientes para garantizar estimaciones adecuadas de las constantes que indica la ley del gas ideal.

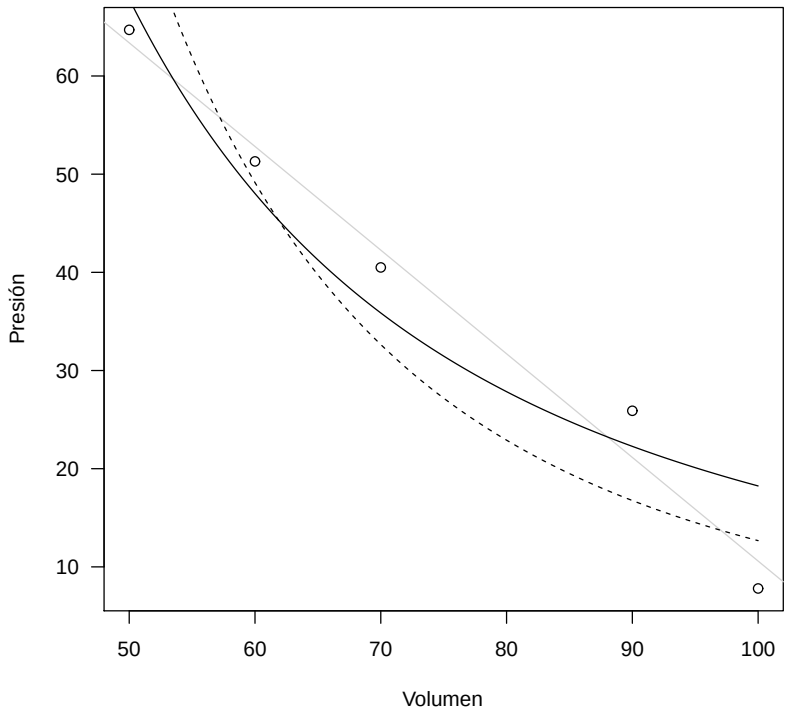

Figura 2: *Ajuste de un modelo potencial por mínimos cuadrados directos (línea continua) y por linealización mediante transformación logarítmica de la variable Y (línea discontinua). En color gris claro se muestra el modelo lineal que se comenta en el texto. Fuente: elaboración propia.*

Los comentarios del final de la sección anterior son válidos para el modelo potencial. Cuando se aplica el método directo de mínimos cuadrados, se considera que los errores son de la forma $\varepsilon = Y - \beta_0\ x^{\beta_1}$, es decir, son *aditivos*. Cuando se emplea el método de la transformación logarítmica, los errores se calculan como $\varepsilon = \ln(Y) - \ln(\beta_0\ x^{\beta_1}) = \ln\Big(\dfrac{Y}{\beta_0\ x^{\beta_1}}\Big)$, es decir que son de carácter multiplicativo. Igualmente, si en un contexto inferencial se asume que $\varepsilon \sim N(0, \sigma^2)$, entonces para los mínimos cuadrados directos, la variable Y tiene distribución normal condicional para cada x, mientras que para la transformación logarítmica la distribución condicional de Y para cada x es de tipo *log-normal*.

4. Conclusiones

1. La aplicación de transformaciones sobre la variable dependiente en los modelos con el fin de linealizarlos no conduce a soluciones de mínimos cuadrados.

2. Algunos residuos del modelo pueden resultar falsamente atípicos.

3. La proporción de varianza explicada por el modelo puede ser un indicador inadecuado de la bondad de ajuste.

4. Dependiendo del procedimiento que se utilice, se ajustan modelos diferentes en cuanto al papel que cumplen los errores y a los supuestos acerca de su distribución y en cuanto a las consecuencias que traen sobre la distribución condicional de la variable dependiente.

4.1. Recomendaciones

1. La observación rutinaria de la gráfica de puntos con la curva del modelo es fundamental para ver su calidad.

2. Si se trata de ejercicios de interpolación dentro del rango de los datos observados, el procedimiento de mínimos cuadrados directos es más adecuado que el de la transformación logarítmica.

3. El uso de *software* no especializado en estadística debe ser especialmente cuidadoso, en particular, las hojas electrónicas y las calculadoras científicas.

4. En la actualidad, tanto el desarrollo teórico como el computacional permiten dar respuesta adecuada a la búsqueda de modelos conocidos como linealizables.

4.2. Otros estudios

1. El estudio de propiedades generadas en función de supuestos distribucionales para los errores del modelo, en particular el insesgamiento.

2. La comparación de los procedimientos cuando se utiliza el método de máxima verosimilitud para estimar los parámetros.

3. Las implicaciones del uso de los procedimientos en problemas de regresión múltiple.

4. El uso de otros criterios de comparación que exigen supuestos distribucionales para los errores, como AIC de Akaike.

5. El estudio de las transformaciones para otros modelos no lineales entre variables, como las de la familia *Box-Cox*.

Agradecimientos

Los autores agradecen a los evaluadores la dedicación y el cuidado en la lectura del artículo y los comentarios que permitieron corregir algunos errores y mejorar el contenido.

Recibido: 21 de marzo de 2014
Aceptado: 28 de abril de 2014

Referencias

Mendenhall, W. & McClave, J. (1981), *A Second Course in Business Statistics: Regression Analysis*, Dellen Publishing Company, Santa Clara, California.

Walpole, R., Myers, R., Myers, S. & Ye, K. (2012), *Probability & Statistics for Engineers & Scientists*, Prentice Hall, New York.