CAPÍTULO 4
TRANSFORMACIONES EN REGRESIÓN

Edgar Acuña Fernández

Departamento de Matemáticas
Universidad de Puerto Rico
Recinto Universitario de Mayagüez

Transformaciones para linealizar modelos

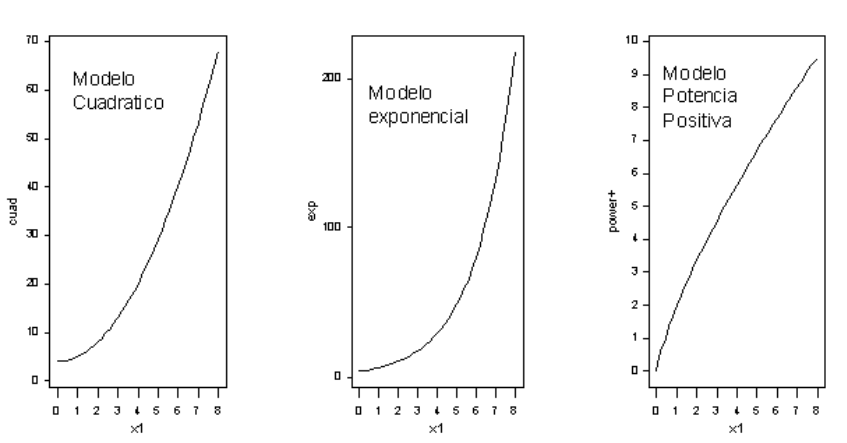
Consideremos por ahora solo modelos con una variable predictora.

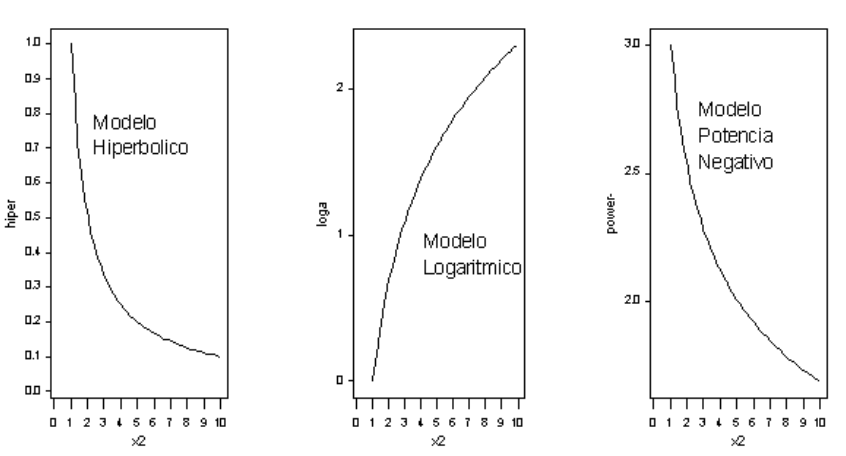
El objetivo es tratar de transformar las variables es aumentar la medida de ajuste R^2 del modelo, sin incluir variables predictoras adicionales.

Se recomienda hacer un plot para observar el tipo de tendencia.

Transformaciones de la variable predictora y/o respuesta para linealizar varios modelos.

Nombre del modelo	Ecuación del Modelo	Transformación	Modelo Linealizado
Exponencial	$Y=\alpha e^{\beta X}$	$Z=\text{Log}Y \quad X=X$	$Z=\text{Log}\,\alpha+\beta X$
Logaritmico	$Y=\alpha+\beta\,\text{Log}\,X$	$Y=Y \quad W=\text{Log}\,X$	$Y=\alpha+\beta W$
Doblemente Logarítmico o Potencia	$Y=\alpha X^{\beta}$	$Z=\text{Log}\,Y \quad W=\text{Log}\,X$	$Z=\text{Log}\,\alpha+\beta W$
Hiperbólico	$Y=\alpha+\beta/X$	$Y=Y \quad W=1/X$	$Y=\alpha+\beta W$
Doblemente Inverso	$Y=1/(\alpha+\beta X)$	$Z=1/Y \quad X=X$	$Z=\alpha+\beta X$





- Los modelos exponencial y doblemente logarítmico son válidos bajo la suposición de que los **errores son multiplicativos**, esto se debe cotejar haciendo análisis de residuales, si los logaritmos de los errores tiene una media de cero y varianza constante.
- Si los errores no son multiplicativos entonces deberían aplicarse técnicas de **regresión no lineal.**

4.2 Transformaciones de las variables predictoras en regresión múltiple

- Supongamos que uno tiene una variable de respuesta Y y varias variables predictoras y desea hacer transformaciones en las variables predictoras para mejorar la medida de ajuste del modelo.
- Del plot matricial se extrae las relaciones de y con cada una de las variables predictoras.
- Las transformaciones pueden ser afectadas por la colinealidad existente entre las variables predictoras.

Modelo basado en series de taylor

En 1962, Box y Tidwell , propusieron un método para transformar las variables predictoras, usando solamente potencias de ellas. ellos consideraron el modelo:

$$y = \beta_o + \beta_1 w_1 + ...\beta_k w_k + e$$

donde: $w_j = x_j^{\alpha_j}$ si $\alpha_j \neq 0$ y $w_j = ln(x_j)$ si $\alpha_j = 0$.

El desarrollo de la serie de Taylor se hace con respecto a $\boldsymbol{\alpha} = (\alpha_1,.....\alpha_k)$ y alrededor de $\boldsymbol{\alpha_0} = (\alpha_{1,0},....., \alpha_{k,0}) = (1,....,1)$

Luego se tiene

$$y \cong \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma_1 z_1 + \gamma_2 z_2 + ... + \gamma_k z_k$$

Donde: $\gamma_j = (\alpha_j - 1)\beta_j$ y $z_j = x_j \ln x_j$ para j=1,2….k.

Procedimiento para la estimación de los α_j

1. Hacer la regresión lineal múltiple considerando las variables predictoras originales x_j y denotar los estimados de los coeficientes por b_j.

2. Hacer una regresión lineal múltiple de y versus las variables predictoras originales mas las variables $z_j = x_j ln(x_j)$ y denotar los estimados de los coeficientes de z_j por $\hat{\gamma}_j$.

3. Estimar α_j por $\hat{\alpha}_j = \frac{\hat{\gamma}_j}{b_j} + 1$

Procedimiento para la estimación de los α_j

El procedimiento se puede repetir varias veces usando en cada etapa las nuevas variables transformadas y la siguiente relación de recurrencia:

$$\hat{\alpha}_j^{(m+1)} = (\frac{\hat{\gamma}_j^{(m)}}{b_j^{(m)}} + 1)\hat{\alpha}_j^{(m)}$$

El proceso termina cuando $|\alpha_j^{(m+1)} - \alpha_j^{(m)}| < TOL$, donde *TOL* es la tolerancia su valor es muy cercano a cero.

Sin embargo, muy a menudo un solo paso es suficiente.

Técnica sugerida por Box and Tidwell aplicado al conjunto de datos **millaje**

La regresión con las variables originales

MPG = 192 - 0.0156 VOL + 0.392 HP - 1.29 SP - 1.86 WT

Predictor	Coef	SE Coef	T	P
Constant	192.44	23.53	8.18	0.000
VOL	-0.01565	0.02283	-0.69	**0.495**
HP	0.39221	0.08141	4.82	0.000
SP	-1.2948	0.2448	-5.29	0.000
WT	-1.8598	0.2134	-8.72	0.000

R-Sq = 87.3% R-Sq(adj) = 86.7%

continuación…

Creamos cuatro variables predictoras $x_1 lnx_1$, $x_2 lnx_2$, $x_3 lnx_3$ y $x_4 lnx_4$.
La regresión múltiple con las 8 variables predictoras es

MPG = 1048 - 1.00 VOL + 5.47 HP - 38.9 SP - 17.9 WT + 0.180 x1lnx1
- 0.801 x2lnx2 + 6.36 x3lnx3 + 3.33 x4lnx4

Predictor	**Coef**	**SE Coef**	**T**	**P**
Constant	1048.2	268.4	3.91	0.000
VOL	-1.0023	0.5916	-1.69	**0.094**
HP	5.468	1.849	2.96	0.004
SP	-38.85	11.81	-3.29	0.002
WT	-17.902	4.324	-4.14	0.000
x1lnx1	0.1803	0.1086	1.66	0.101
x2lnx2	-0.8006	0.2744	-2.92	0.005
x3lnx3	6.362	1.971	3.23	0.002
x4lnx4	3.3263	0.8739	3.81	0.000

S = 3.247 R-Sq = 90.5% R-Sq(adj) = 89.5%

continuación …

Estimando α_j segun el paso 3 se tiene
α_1=0.1803/(-0.01565)+1=-10.52, α_2 =8006/0.39221)+1=-1.04,
α_3=(6.362)/(-1.2948)+1=-3.91, α_4=3.3263/(-1.8598)+1=-0.79.

Asi la regresión con las nuevas variables $vol^{-10.52}$, $hp^{-1.04}$, $sp^{-3.91}$ y $wt^{-0.79}$ es:

MPG = - 2.298 + 1668 hp1 +1.843E+18 -1.465E+08 sp1 + 332.9 wt1

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.298e+00	4.420e+00	-0.520	0.604656	
sp1	-1.465e+08	4.698e+08	-0.312	0.755972	
wt1	3.329e+02	9.382e+01	3.548	0.000665	***
vol1	1.843e+18	8.827e+17	2.088	0.040082	*
hp1	1.668e+03	8.078e+02	2.065	0.042325	*

Residual standard error: 3.095 on 77 degrees of freedom

Multiple R-Squared: 0.909, Adjusted R-squared: 0.9043

Se puede repetir el proceso, eliminando VOL antes de aplicar el método de Box and Tidwell

Transformaciones para estabilizar la varianza.

Transformación	Situación
$\sqrt{y}$	$Var(e_i) \propto E(y_i)$
$\sqrt{y}+\sqrt{y+1}$	$Var(e_i) \propto E(y_i)$
$Log(Y)$	$Var(e_i) \propto (E(y_i))^2$
$Log(y+1)$	$Var(e_i) \propto (E(y_i))^2$
$1/y$	$Var(e_i) \propto (E(y_i))^4$
$1/(y+1)$	$Var(ei) \propto (E(yi))4$
$Sin^{-1}(\sqrt{y})$	$Var(e_i) \propto E(y_i)[1-E(y_i)]$

Justificación de las transformaciones

Expandiendo en series de Taylor una función $h(Y)$ alrededor de $\mu=E(Y)$ se obtiene:

$$h(Y) \approx h(\mu) + h'(\mu)(Y-\mu) + h''(\mu)(Y-\mu)^2/2$$

Tomando varianza a ambos lados y considerando solamente la aproximación lineal se obtiene:

$$Var(h(Y)) \approx [h'(E(y))]^2 Var(Y)$$

Ejemplo: Si $Var(Y) \propto [E(y)]^2$ hallar h(Y) tal que su varianza sea constante: $[h'(E(Y))]^2 \approx constante/[E(y)]^2$, luego, $h'(\mu) \approx 1/\mu$, de donde $h(\mu) \approx \log(\mu)$.

Transformaciones para mejorar la normalidad de la variable de respuesta

En 1964, Box y Cox introdujeron una transformación de la variable de respuesta (transformación potencia) con el objetivo de satisfacer la suposición de normalidad del modelo de regresión.

la transformación está definida por $w = \frac{y^{\lambda} - 1}{\lambda}$

si $\lambda \neq 0$ y $w = ln(y)$ si $\lambda = 0$.

Notar que: $\lim_{\lambda \to 0} \frac{y^{\lambda} - 1}{\lambda} = \ln y$

Estimación del parámetro λ

Se estima conjuntamente con los coeficientes del modelo de regresión lineal múltiple

usando el método de **Máxima verosimilitud**,

$$w = \beta_o + \beta_1 x_1 + ...\beta_k x_k + e$$

La transformación estandarizada de los *w's* se define por $z_i = \frac{w_i}{\tilde{y}^{\lambda-1}}$ donde $\tilde{y} = (\prod_{i=1}^{n} y_i)^{1/n}$ es la media geométrica de las *y's*.

El método asume que para algún λ las z_i*'s* son normales e independientes con varianza común σ^2

Función de verosimilitud en términos de las z_i's

La función de verosimilitud esta dada por :

$$L(\boldsymbol{\beta},\lambda)] = \frac{e^{-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})}}{(2\pi\sigma^2)^{n/2}}$$

Luego

$$\max[LnL(\boldsymbol{\beta},\lambda)] = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

donde $\hat{\sigma}^2 = SSE/n = (\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})/n$

Así

$$\max[LnL(\boldsymbol{\beta},\lambda)] = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2} \equiv -\frac{n}{2}\ln(\hat{\sigma}^2)$$

Procedimiento para estimar el parámetro λ

1. Seleccionar un conjunto de valores de λ entre –2 y 2, usualmente entre 10 y 20 valores
2. Para cada valor de λ, ajustar el modelo

$$\mathbf{z=X\beta+e}$$

3. Plotear *max*[*Ln* L(β,λ)] versus λ.
4. Escoger como parámetro λ aquel que otorgue el mayor valor para *max*[*Ln* L(β,λ)].

Mínimos cuadrados ponderados.

Es otra manera de tratar de remediar la falta de homogeneidad de varianza de los errores. suponiendo que los errores son todavía no correlacionados.

Se minimiza $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$, donde w_i representa el peso asignado a la *i-ésima* observación.

Cómo escoger los pesos?

Si en el plot de, residuales versus la variable predictora se observa que la dispersión aumenta cuando x aumenta sería conveniente usar $w_i = \frac{1}{\sigma_i^2}$

Donde, σ_i^2 son las varianzas poblacionales de la Y (estimadas por s^2) para cada observación x_i en caso de regresión lineal simple, o para cada combinación de las variables predictoras en el caso de regresión lineal múltiple.

La idea de dar a las observaciones anómalas un menor peso.

Tambien se pueden calcular los pesos basado en los diagnósticos de regresión.

Consideremos el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

con $Var(e)=V\sigma^2$, donde **V** es una matríz diagonal

$$\mathbf{V} = \begin{bmatrix} k_1^2 & 0 & . & . & 0 \\ 0 & k_2^2 & . & . & 0 \\ 0 & 0 & k_3^2 & . & 0 \\ . & . & . & . & 0 \\ 0 & 0 & 0 & 0 & k_n^2 \end{bmatrix}$$

Sea $\mathbf{W}=(\mathbf{V}^{1/2})^{-1}$, luego $\mathbf{Wy} = \mathbf{WX\beta} + \mathbf{We}$

Sea $y^*=Wy$, $e^*=We$ y $X^*=WX$, **luego** $\mathbf{y^*=X^*\beta + e^*}$

Algunas Propiedades

La varianza de los errores es constante.

$$Var(\mathbf{e}^*)=Var(\mathbf{We})=\mathbf{W}Var(\mathbf{e})\mathbf{W'}=\mathbf{WVW'}\sigma^2=\mathbf{I}\sigma^2$$

El estimador mínimo cuadrático de β es

$$\boldsymbol{\beta}^* = (\mathbf{X}^{*\prime}\mathbf{X})^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

Para el cual se tiene

$$E(\beta^*)=\beta$$

$$Var(\beta^*)= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Var(\mathbf{Y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$$

Mínimos Cuadrados generalizados

Considera la situación más general donde:
Los errores no tienen varianza constante y ademas son correlacionados.

Sea el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Supongamos que: $Var(\mathbf{e}) = \mathbf{V}\sigma^2$
donde $\mathbf{V}$ es una matriz simétrica y definida positiva. Sea $\mathbf{T}$ una matríz nosingular y simétrica tal que $\mathbf{TT} = \mathbf{T}^2 = \mathbf{V}$, luego se tiene

$$\mathbf{T}^{-1}\mathbf{y} = \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{T}^{-1}\mathbf{e}$$

Mínimos Cuadrados generalizados

Sea $\mathbf{e^*= T^{-1}e}$,

$Var(\mathbf{e^*})=Var(\mathbf{T^{-1}e})=\mathbf{T^{-1}}Var(\mathbf{e})\mathbf{T^{-1}}=\mathbf{I}\sigma^2$ entonces el estimador mínimo cuadrático de β se obtiene minimizando

$$\mathbf{e^{*\prime}e^*} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

luego

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

$E(\beta^*)=\beta$

$Var(\beta^*)=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Var(\mathbf{Y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$