# Capítulo 6

# Transformaciones

## 6.1. Introducción

Existen problemas para los que se conoce, a partir de la teoría, que la función media es lineal. Sin embargo no siempre es posible decidir sobre la forma correcta de la función media y cualquier elección constituye una *aproximación* ya que se espera que sea adecuada al problema. Transformar la variable respuesta y los predictores permite al analista usar la metodología de regresión lineal para abordar aquellos problemas que incluyen relaciones no lineales. Lo anterior permite formular dos preguntas: ¿cómo escoger estas transformaciones? ¿cómo decidir si un modelo aproximado es adecuado para la data?

## 6.2. Transformaciones y diagramas de dispersión

El propósito de las transformaciones es lograr que una función media sea aproximadamente lineal. En problemas con solo un predictor y una variable respuesta se puede observar en un diagrama de dispersión el tipo de relación que tienen las variables transformadas. Con muchos predictores la selección de las transformaciones puede ser más difícil. Se busca una transformación de tal forma que si $X^*$ es el *predictor transformado* y $Y^*$ es la *respuesta transformada*, entonces la función media es:

$$\mathrm{E}(Y^*|X^* = x^*) = \beta_0 + \beta_1 x^*$$

La Figura 6.1 muestra el gráfico del peso del cuerpo y el peso del cerebro para 62 especies que se encuentran en la data `Mamiferos`. Además de los tres puntos separados para dos especies de elefantes y los humanos, la distribución desigual de los puntos esconde cualquier información útil acerca de la relación entre `PesoCerebro` y `PesoCuerpo`. Existe poca evidencia en el diagrama de dispersión que justifique el uso de una función media lineal. El rango para ambas variables va desde especies pequeñas que pesan pocos gramos hasta animales por encima de los 6600 kilogramos. Las transformaciones pueden ayudarnos en este problema.

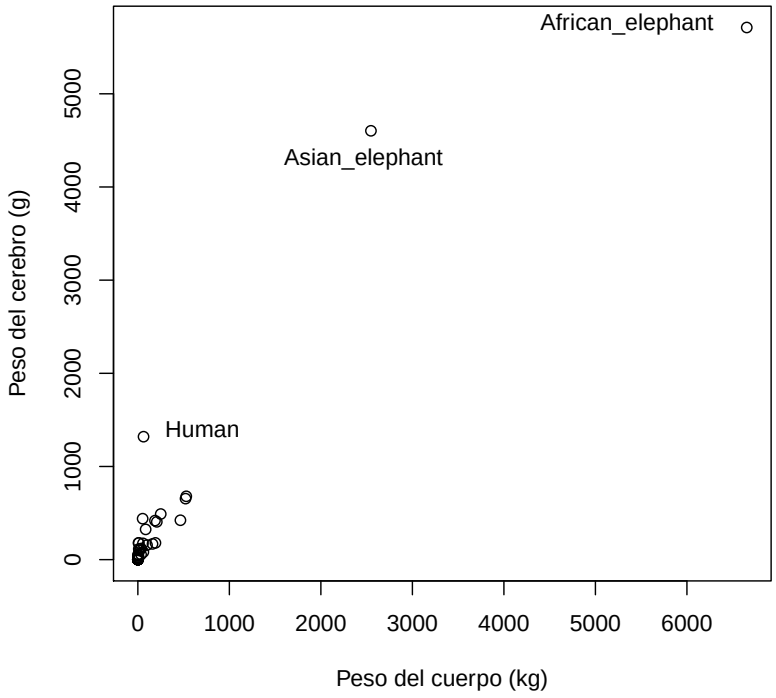


Figura 6.1: Diagrama de dispersión para la data `Mamiferos`

### 6.2.1. Transformaciones potencia

Una *familia de transformaciones* es una colección de transformaciones indexadas por uno o más parámetros que el analista debe seleccionar. La familia más usada es llamada *familia potencia*, definida para una variable estrictamente positiva $U$ por:

$$\psi(U, \lambda) = U^{\lambda} \tag{6.2.1}$$

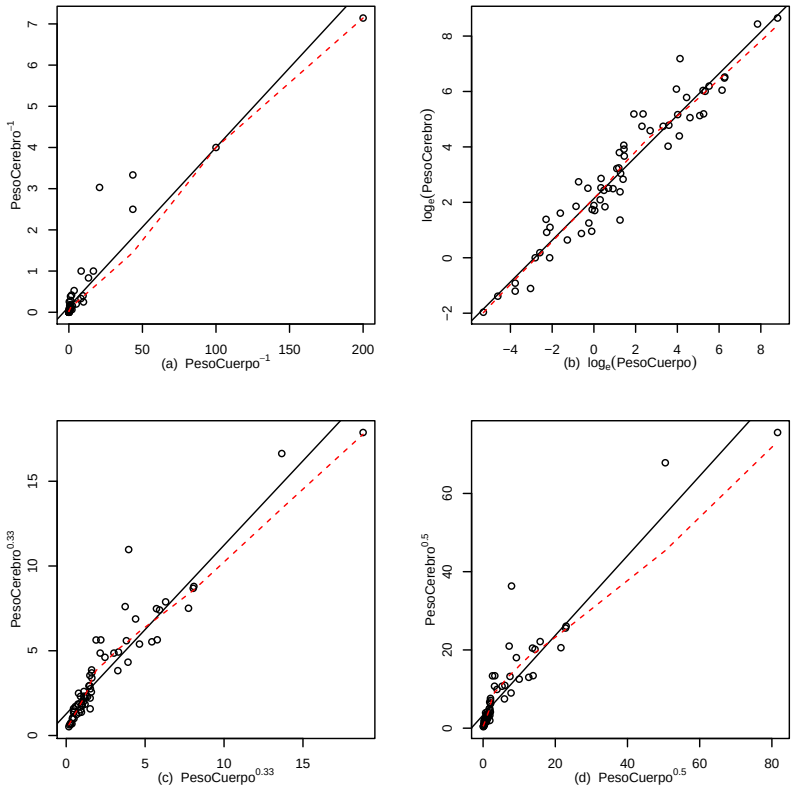


Figura 6.2: Data `Mamiferos` con cuatro posibles transformaciones

Conforme el parámetro $\lambda$ cambia, se obtienen los miembros de esta familia incluyendo las transformaciones raíz cuadrada, cúbica e inversa. Se interpreta el valor de $\lambda = 0$ como la *transformación logarítmica*. Los valores de $\lambda$ se

encuentran por lo general en el rango de $-1$ y 1. El valor de $\lambda = 1$ *no corresponde a transformación alguna.*

La Figura 6.2 muestra los gráficos para las transformaciones obtenidas usando los valores de $\lambda = -1, 0, 1/3, 1/2$ para ambas variables. No es necesario aplicar la misma transformación a $X$ y $Y$ sin embargo en esta situación es razonable ya que ambas representan el mismo tipo de medición. Si se permite que cada variable tenga su propia transformación, la búsqueda visual de la más apropiada podría ser complicada ya que se deben considerar muchas más posibilidades.

A partir de los cuatro gráficos anteriores la elección apropiada es la transformación logarítmica. El uso de logaritmos para la data `Mamiferos` no es sorprendente, de acuerdo a dos reglas empíricas que son bastante útiles:

- *La regla de logaritmos* Si el valor del rango de la variable es mayor que uno y la variable es estrictamente positiva, entonces podría ser útil reemplazar la variable por su logaritmo.

- *La regla del rango* Si el rango de una variable es mucho menor que uno, entonces cualquier transformación que se use sobre la variable es poco probable que resulte útil.

La regla del logaritmo se satisface para `PesoCuerpo` con rango de 0.005 kilogramos a 6654 kilogramos y para `PesoCerebro`, con rango de 0.14 gramos a 5712 gramos. La transformación logarítmica es un buen punto de partida para examinar cualquier otra transformación sobre las variables.

La regresión lineal simple parece ser apropiada para ambas variables en la escala logarítmica. Lo anterior corresponde al *modelo físico*:

$$\text{PesoCerebro} = \alpha \times \text{PesoCuerpo}^{\beta_1} \times \delta \tag{6.2.2}$$

donde $\delta$ es un *error multiplicativo* y se espera que tenga media 1 y una distribución concentrada en valores cercanos a él. Utilizando logaritmos y tomando $\beta_0 = \log \alpha$ y $e = \log \delta$,

$$\log(\text{PesoCerebro}) = \beta_0 + \beta_1 \log(\text{PesoCuerpo}) + e$$

que es un modelo de regresión lineal simple. Los científicos que estudian la relación entre atributos de individuos o especies llaman a 6.2.2 un modelo *alométrico* y el valor de $\beta_1$ juega un rol importante en este tipo de estudios. Sin embargo, no todas las transformaciones corresponden a modelos físicos interpretables.

### 6.2.2. Transformando solo la variable predictora

En el ejemplo de la data Mamiferos se requiere transformar ambas variables para obtener una función media lineal. En otros problemas, podría ser necesario trasformar solo una variable. Si se desea usar una familia de transformaciones potencia es conveniente introducir la familia de *transformaciones potencia escaladas* definida para $X$ estrictamente positiva por:

$$\varphi_S(X,\lambda) = \begin{cases} \left(X^\lambda - 1\right)/\lambda & \text{si } \lambda \neq 0 \\ \log(X) & \text{si } \lambda = 0 \end{cases} \tag{6.2.3}$$

Las transformaciones potencia escaladas difieren de las transformaciones potencia en algunos aspectos. Primero $\varphi_S(X,\lambda)$ es una función continua de $\lambda$. Como $\lim_{\lambda\to 0}\varphi_S(X,\lambda) = \log(X)$ entonces la transformación logarítmica es un miembro de esta familia cuando $\lambda = 0$. Además, $\varphi_S(X,\lambda)$ preserva la dirección de la asociación, en el sentido que si $(X,Y)$ se encuentran relacionadas de manera positiva entonces $(\varphi_S(X,\lambda), Y)$ mantiene el mismo tipo de relación para todos los valores de $\lambda$. Con las transformaciones potencia la dirección de la asociación cambia cuando $\lambda < 0$.

Si se consigue una potencia adecuada usando $\varphi_S(X,\lambda)$ en la práctica podría usarse la transformación potencia $\varphi(X,\lambda)$ en el modelamiento de la regresión ya que ambos difieren solo en escala, locación y posiblemente cambios de signo. Las transformaciones escaladas son usadas solo para seleccionar la transformación a utilizar.

Si se transforma el predictor usando un miembro de la familia potencia escalada se tiene la siguiente función media:

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1\varphi_S(X,\lambda) \tag{6.2.4}$$

Si se conoce $\lambda$ se puede estimar 6.2.4 usando mínimos cuadrados ordinarios y luego obtener la suma de cuadrados residual. El estimado $\hat{\lambda}$ es el valor que minimiza $SCRes(\lambda)$. Como regla práctica, no es necesario conocer exactamente $\lambda$ y por lo general es suficiente seleccionar su valor de:

$$\lambda \in \{-1, -1/2, 0, 1/3, 1/2, 1\} \tag{6.2.5}$$

Como ejemplo, considere la dependencia de la Altura del árbol en decímetros sobre el diámetro del árbol, Dbh en mm para una muestra de árboles de cedro en 1991. La data se encuentra en Arboles.

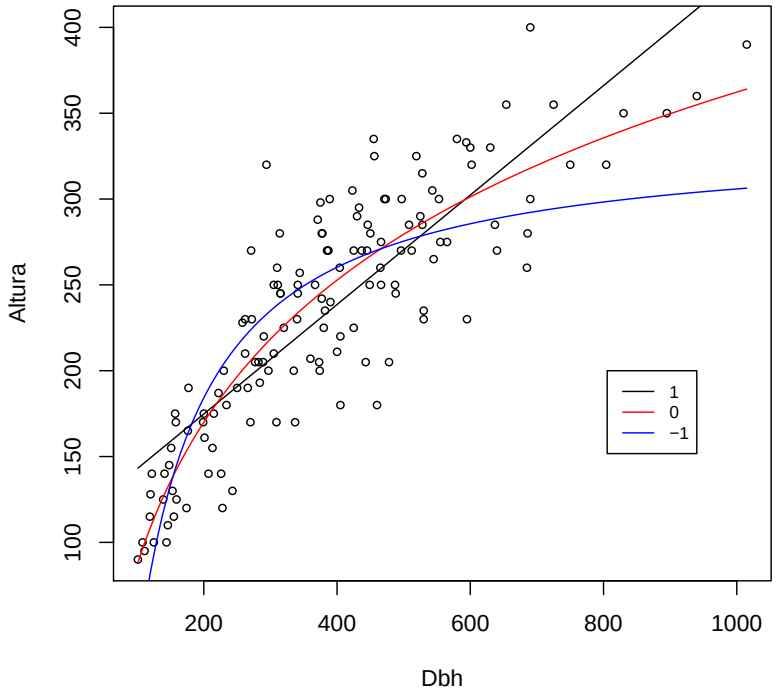


Figura 6.3: `Altura` versus `Dbh` para la data `Arboles`

La Figura 6.3 muestra el gráfico de dispersión de la data sobre el que se ha superpuesto tres curvas. Para cada $\lambda$ se calculan los valores estimados $\hat{y}(\lambda)$ de la regresión por mínimos cuadrados ordinarios de `Altura` sobre $\varphi_S(\text{Dbh}, \lambda)$. Entre los tres valores de $\lambda$ utilizados en la figura, $\lambda = 0$ parece estimar mejor la data. Para $\lambda = 1$ no se logra un buen estimado para árboles grandes y pequeños, mientras que la inversa es demasiado curva para árboles grandes. Lo anterior sugiere reemplazar `Dbh` por su logaritmo tal como se observa en la Figura 6.4.

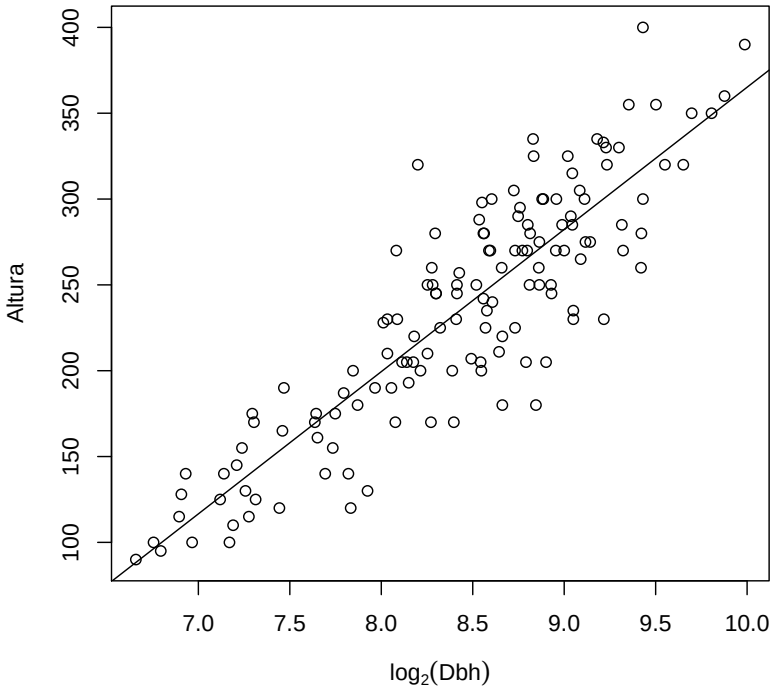


Figura 6.4: La data `Arboles` transformada

### 6.2.3. Transformando solo la variable respuesta. El método de Box y Cox

Box y Cox (1964) proporcionan un método general para seleccionar la transformación a utilizar sobre la variable respuesta para resolver el problema de inclumplimiento del supuesto de normalidad y que es aplicable tanto a la regresión simple como múltiple. Así como en los métodos previos, se debe seleccionar la transformación a partir de una familia indexada por el parámetro $\lambda$. Para el método de Box y Cox se necesita una versión ligeramente más complicada de las familias potencia, llamadas *familia potencia modificada* definidas para una variable respuesta $Y$, estrictamente positiva, por:

$$\varphi_M(Y,\lambda_y) = \begin{cases} \mathrm{gm}(Y)^{1-\lambda_y}\left(Y^{\lambda_y}-1\right)/\lambda_y & \text{si } \lambda_y \neq 0 \\ \mathrm{gm}(Y)\log(Y) & \text{si } \lambda_y = 0 \end{cases} \quad (6.2.6)$$

donde $\mathrm{gm}(Y)$ es la *media geométrica* de la variable no transformada. Si los valores de $Y$ son $y_1, \cdots, y_n$ entonces la media geométrica de $Y$ es $\mathrm{gm}(Y) = \exp\{\sum \log(y_i)/n\}$.

En el método de Box y Cox se asume que la función media:

$$\mathrm{E}(\varphi_M(Y,\lambda_y)|\mathbf{X}=\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \quad (6.2.7)$$

es correcta para cualquier $\lambda_y$. Se estima $\lambda_y$ como el valor de la transformación potencia que minimiza $SCRes(\lambda_y)$. Desde un punto de vista práctico se puede seleccionar $\lambda_y$ a partir de 6.2.5.

## 6.3. Transformaciones y matrices de dispersión

La data `Carreteras` descrita en la Tabla 6.1 fue tomada de un paper en ingeniería civil. En este trabajo se relaciona la tasa de accidentes automovilísticos con algunos términos potenciales. La data incluye 39 secciones de carreteras en el estado de Minnesota en 1973.

Tabla 6.1: Data `Carreteras`

| Variable | Descripción |
|---|---|
| `Tasa` | Tasa de accidentes en 1973 por millón de millas |
| `Longitud` | Longitud de los segmentos en millas |
| `TDP` | Tráfico diario promedio estimado (en miles) |
| `Volumen` | Volúmen del auto como porcentaje del total |
| `Borde` | Ancho del borde externo de la vía (en pies) |
| `Señales` | Número de semáforos por milla |

El objetivo de este análisis fue entender el impacto de algunas variables que se encuentran bajo el control del departamento de carreteras sobre la tasa de accidentes. Sin embargo, también se incluyeron variables adicionales para reducir la variabilidad debido a aquellos factores no controlables.

Un primer paso, importante, en este análisis es examinar la matriz de dispersión para todos los predictores y la variable respuesta, tal como se muestra en la Figura 6.5.

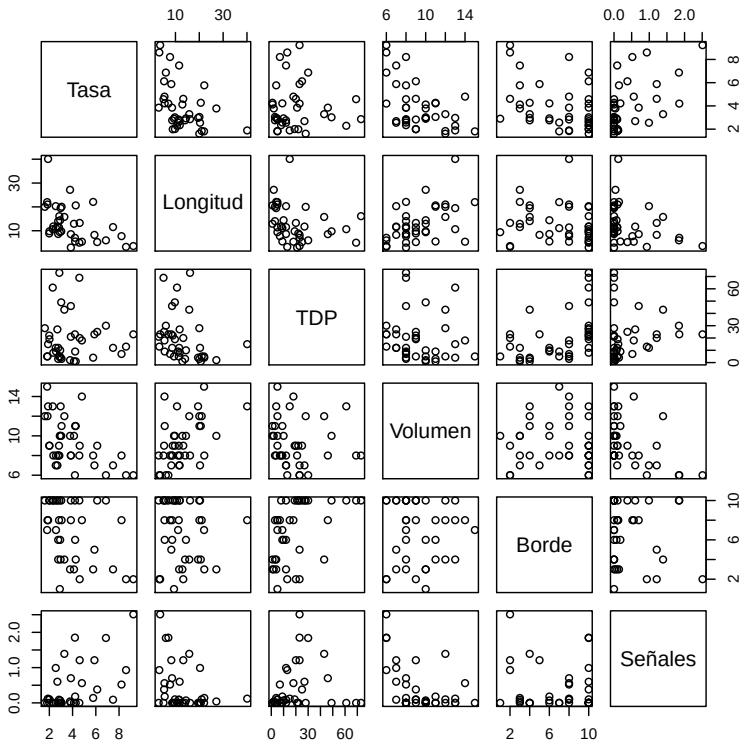


Figura 6.5: La data `Carreteras` no transformada

Existen algunas observaciones sobre este gráfico que podrían ayudar a seleccionar las transformaciones:

1. Se define la variable `Señales1` como el número total de señales dividido por `Longitud`, sumando uno al numerador para evitar que tome valores iguales a cero y no sea posible aplicar las transformaciones potencia.

$$\text{Señales1} = \frac{\text{Señales} \times \text{Longitud} + 1}{\text{Longitud}}$$

2. TDP y Longitud tienen un rango alto y probablemente los logaritmos puedan resultar apropiados para ellos.

3. Cada uno de los predictores se encuentran modestamente asociados con Tasa.

### 6.3.1. Selección automática de las transformaciones para los predictores

Usando los resultados anteriores, se busca una transformación de los predictores que permita que los gráficos de dispersión de cualquier predictor versus la variable respuesta tengan una función media aproximadamente lineal. Sin herramientas gráficas interactivas o algún método automático de selección de las transformaciones, lo anterior podría ser una tarea desalentadora ya que el analista necesitaría obtener muchas matrices de dispersión hasta lograr un conjunto útil de transformaciones.

Velilla (1993) propuso una extensión multivariada del método de Box-Cox para obtener un punto de partida que permita seleccionar las transformaciones. Para un conjunto de $k$ predictores estrictamente positivos $\mathbf{X} = (X_1, \cdots, X_k)$ se aplicara una transformación potencia modificada a cada $X_j$ a partir de $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)'$.

Sea $\varphi_M(\mathbf{X}, \boldsymbol{\lambda})$ el conjunto de variables:

$$\varphi_M(\mathbf{X}, \boldsymbol{\lambda}) = (\varphi_M(X_1, \lambda_1), \cdots, \varphi_M(X_k, \lambda_k))$$

y $\mathbf{V}(\boldsymbol{\lambda})$ la matriz de covariancias muestrales de la data transformada $\varphi_M(\mathbf{X}, \boldsymbol{\lambda})$. El valor $\hat{\boldsymbol{\lambda}}$ se selecciona como el valor de $\boldsymbol{\lambda}$ que minimiza el logaritmo del determinante de $\mathbf{V}(\boldsymbol{\lambda})$.

Retornado a la data Carretera en la Tabla 6.2 se muestra el resumen de las transformaciones usando el método multivariado de Box y Cox. La tabla muestra los valores de $\hat{\boldsymbol{\lambda}}$ y los errores estándar. Las siguientes dos columnas sirven para verificar si el parámetro de transformación es igual a cero o uno. Al final de la tabla se encuentran las *pruebas de razón de verosimilitud*. La primera sirve para probar que todas las potencias son cero, $\boldsymbol{\lambda} = \mathbf{0}$, y la segunda para probar que no es necesario realizar transformaciones, $\boldsymbol{\lambda} = \mathbf{1}$. La última fila de resultados sirve para probar que las tres primeras variables deberían estar en escala logarítmica y la última no transformada, con un $p$-valor de 0.29.

Tabla 6.2: Transformación de Box y Cox para la data Carreteras

```
bcPower Transformations to Multinormality

          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Longitud     0.1437   0.2127          -0.2732           0.5607
TDP          0.0509   0.1206          -0.1854           0.2872
Volumen     -0.7028   0.6177          -1.9134           0.5078
Borde        1.3456   0.3630           0.6341           2.0570
Señales1    -0.2408   0.1496          -0.5341           0.0525

Likelihood ratio tests about transformation parameters
                                       LRT df         pval
LR test, lambda = (0 0 0 0 0)   23.324467  5 0.0002926014
LR test, lambda = (1 1 1 1 1) 132.857421  5 0.0000000000
LR test, lambda = (0 0 0 1 0)   6.088599  5 0.2976930877
```

## 6.4. Transformando la variable respuesta

Una vez transformados los términos, es turno de transformar la variable respuesta. La Figura 6.6 es el gráfico de valor estimado inverso para la data Carretera usando los términos transformados determinados en la sección anterior. En este gráfico la variable respuesta Tasa se encuentra en el eje horizontal y los valores estimados de la regresión sobre los términos transformados en el eje vertical. Cook y Weisberg (1994) mostraron que si los predictores tienen relación lineal de manera aproximada entonces pueden usarse los métodos de la Sección 6.2.3 para seleccionar una transformación para la variable respuesta. Entre las tres curvas mostradas en este gráfico, la transformación logarítmica parece ser la más apropiada.

```
      lambda      RSS
1  0.1197427 32.32520
2 -1.0000000 35.70143
3  0.0000000 32.36532
4  1.0000000 34.24662
```

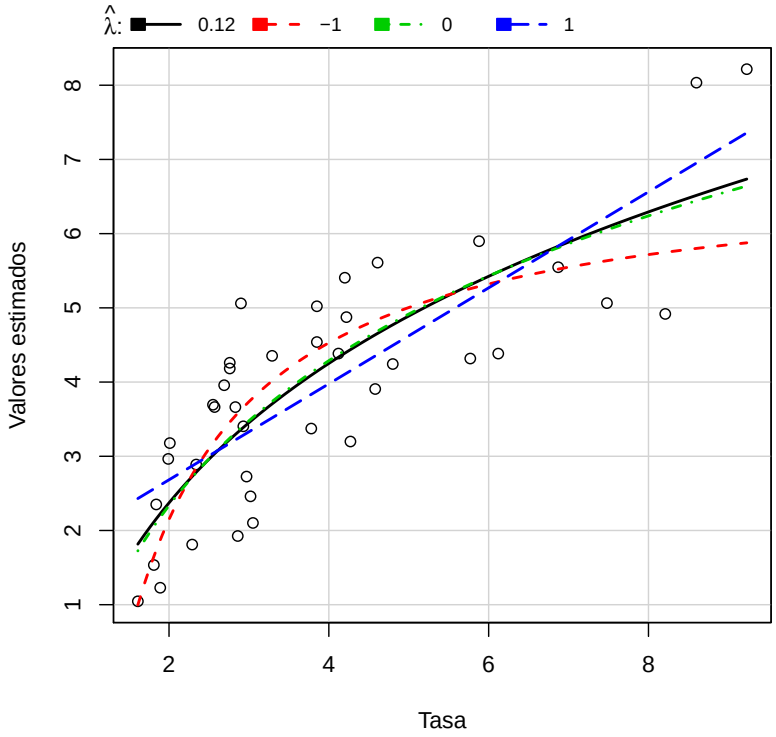


Figura 6.6: Gráfico inverso de valor estimado para la data `Carretera`

El método de Box y Cox proporciona un procedimiento alternativo para encontrar una transformación de la variable respuesta. Este método se resume mediante un gráfico con $\lambda_y$ en el eje horizontal y $-(n/2)\log(SCRes(\lambda_y)/n)$ sobre el eje vertical. En el último caso el estimado $\hat{\lambda}_y$ es el punto que maximiza la curva.

Este gráfico para la data `Carretera` se muestra en la Figura 6.7, con $\hat{\lambda} \approx -0{,}2$ y un intervalo de confianza que en el que se encuentra la transformación

logarítmica.

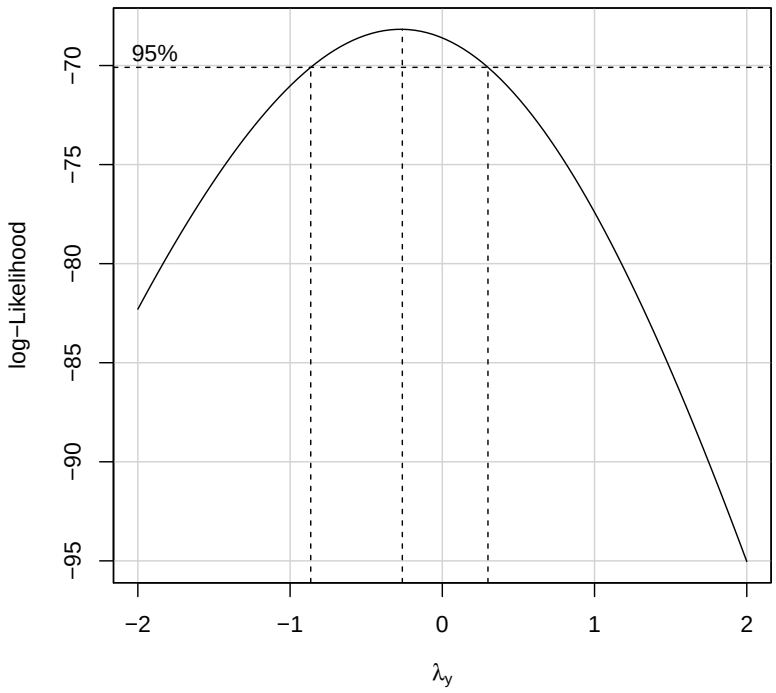


Figura 6.7: Resumen gráfico Box-Cox para la data `Carretera`

Para la data `Carretera`, se tiene un punto de partida razonable para el análisis de regresión con casi todos los predictores y la variable respuesta transformados a la escala logarítmica.

## 6.5. Transformaciones para variables no positivas

Se han sugerido familias de transformaciones para una variable $U$ que incluye valores negativos. La idea central es usar los métodos discutidos en

este capítulo para seleccionar una transformación que permita que $U$ tome valores negativos. Una posibilidad es considerar transformaciones de la forma $(U+\gamma)^{\lambda}$, donde $\gamma$ es lo suficientemente grande para asegurar que $U+\gamma$ es estrictamente positivo. Se usa una variante de este método con la variable `Señales` en la data `Carretera`. En principio, $(\gamma, \lambda)$ puede ser estimado simultáneamente, aunque en la práctica los estimados de $\gamma$ son muy variables y poco confiables. Alternativamente, Yeo y Jonson (2000) propusieron una familia de transformaciones que pueden ser usadas sin restricciones sobre $U$ y tienen muchas de las buenas propiedades de la familia potencia de Box y Cox. Estas transformaciones están definidas por:

$$\varphi_{YJ}(U,\lambda) = \begin{cases} \varphi_M(U+1,\lambda) & \text{si } U \geq 0 \\ \varphi_M(-U+1, 2-\lambda) & \text{si } U < 0 \end{cases} \tag{6.5.1}$$

Si $U$ es estrictamente positiva, entonces la transformación de Yeo-Johnson es la misma que la transformación potencia de Box y Cox para $(U+1)$. Si es estrictamente negativa, entonces la transformación de Yeo-Johnson es la transformación potencia de Box y Cox para $(-U+1)$, pero con potencia $2-\lambda$. Con valores positivos y negativos, la transformación es una mixtura de los dos, es decir potencias diferentes para valores positivos y negativos. En este último caso, la interpretación del parámetro de transformación es difícil, ya que tiene diferentes significados para $U \geq 0$ y para $U < 0$.