CAPÍTULO 4
TRANSFORMACIONES EN REGRESIÓN

4.1 Transformaciones para linealizar modelos

Consideremos por ahora solo modelos con una variable predictora. La idea es tratar de aumentar la medida de ajuste R^2 del modelo, sin incluir variables predictoras adicionales. Lo primero que hay que hacer es un plot para obervar el tipo de tendencia, pueden resultar plots como los que aparecen en las figuras 4.1 y 4.2.

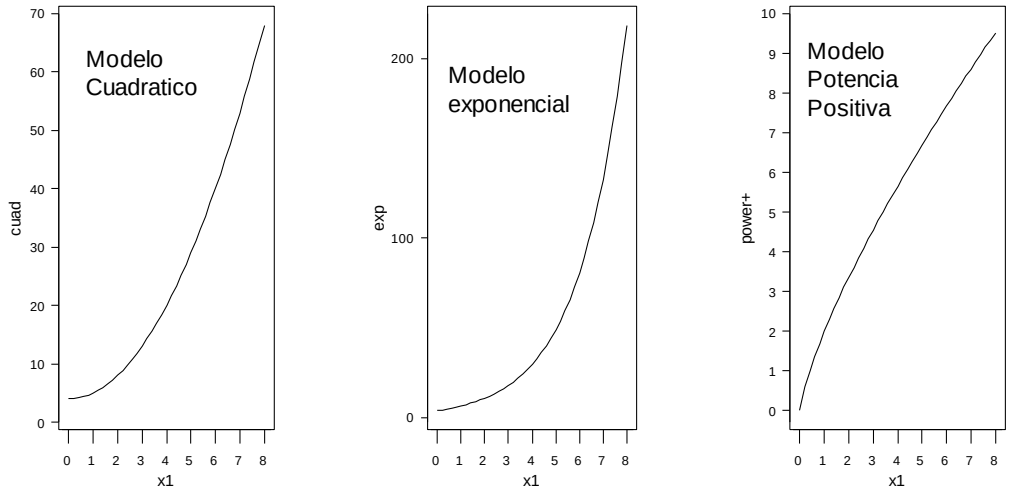

Figura 4.1. Gráficas de tres modelos no lineales.

En la primera gráfica de la figura 4.1 se ha ajustado un modelo cuadrático, que es de la forma general $y=a+bx+cx^2$ y es el caso más sencillo de regresión polinómica. Esto puede ser modelado como una regresión múltiple con dos variables predictoras.
La segunda gráfica corresponde a un modelo exponencial de la forma $y=\alpha e^{\beta x}$ con α y β positivos. Este modelo es muy adecuado para modelar crecimientos poblacionales.
La tercera gráfica corresponde a un modelo potencial o doblemente logarítmico de la forma $y=\alpha x^{\beta}$, con β positivo.

La primera gráfica de la figura 4.2 corresponde a un modelo hiperbólico o inverso de la forma $y=\alpha+\beta/x$, con $x>0$.
La segunda gráfica corresponde a un modelo logarítmico de la forma $y==\alpha+\beta\log(x)$ con $x>0$.
La tercera gráfica corresponde a un modelo potencia pero con $\beta>0$.

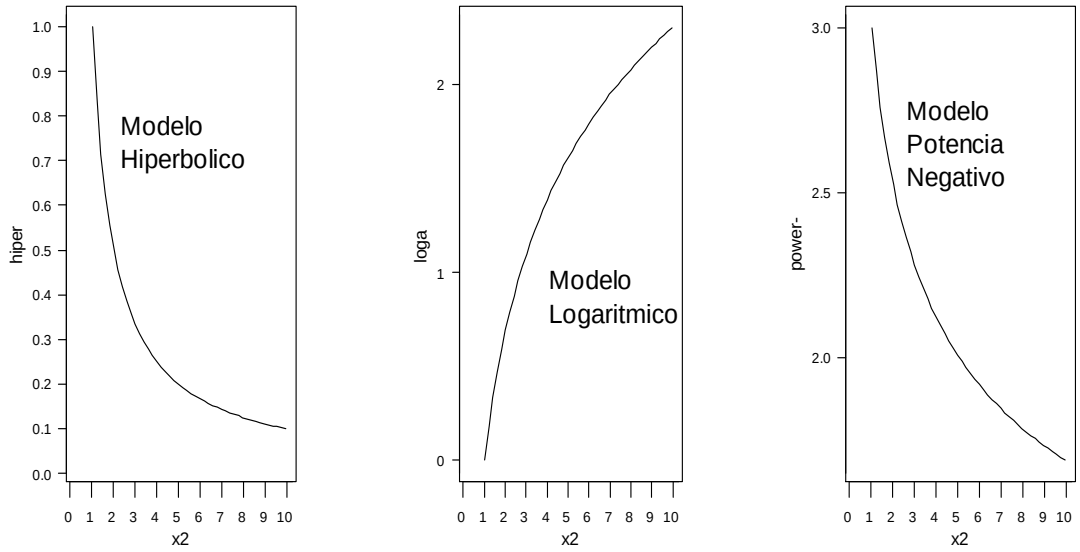

Figura 4.2. Mas gráficas de modelos no lineales

La siguiente tabla muestra las transformaciones de la variable predictora y/o respuesta que se requieren para linealizar varios modelos.

Nombre del modelo	**Ecuación del Modelo**	**Transformación**	**Modelo Linealizado**
Exponencial	$Y=\alpha e^{\beta X}$	Z=LogY X=X	Z=Log α +βX
Logaritmico	Y= α +βLog X	Y=Y W=Log X	Y= α +βW
Doblemente Logarítmico o Potencia	$Y=\alpha X^{\beta}$	Z=Log Y W=Log X	Z= Log α +βW
Hiperbólico	Y= α +β/X	Y=Y W=1/X	Y= α +βW
Doblemente Inverso	Y=1/(α +βX)	Z=1/Y X=X	Z=α +βX

El primer y tercer modelo son válidos bajo la suposición de que los errores son multiplicativos y habría que cotejar haciendo análisis de residuales. Si el logaritmos de los errores tiene una media de cero y varianza constante entonces se cumplirian los supuestos . Si los errores no son multiplicativos entonces deberían aplicarse técnicas de regresión no lineal las cuales no son consideradas en este texto.

Ejemplo 1. Los siguientes datos representan como ha cambiado la población en Puerto Rico desde 1930.

```
year  poblacion

1930    1552000
1940    1877800
```

```
1950    2218000
1960    2359800
1970    2716300
1980    3196520
1990    3527796
```

Se desea establecer un modelo para predecir la población de Puerto Rico en el año 2000.

Solución: Observando el diagrama de puntos de población versus años que aparece en la figura de abajo

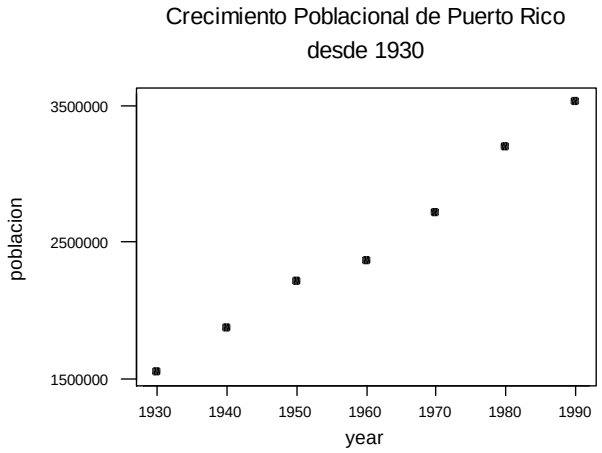

Figura 4.3 Crecimiento poblacional de Puerto Rico desde 1930

El plot sugiere que podemos ajustar los datos al modelo exponencial

$Poblac=\alpha e^{\beta year}$

Y el modelo linealizado da como ecuación

Ln(Poblac) = - 11.4 + 0.0133 year

con un R^2 del 98.9%, mejorando el R^2 del modelo lineal que era de 98.7%. Para predecir la población para el año 2000 se obtiene que

Ln(Poblac)=-11.4 + 0.0133(2000)=-11.4+26.6=15.2

luego $Poblac=e^{15.2}$=3,992,787 será la población de PR estimada para el año 2000.

4.2 Transformaciones para estabilizar la varianza.

Algunas veces el comportamiento de la varianza varía según la variable de respuesta. Una de las medidas remediales para hacer constante la varianza es transformar la variable de respuesta. La siguiente tabla muestra las transformaciones de la variable de respuesta que hay que hacer para hacer que la varianza sea constante

Situación	**Transformación**
$Var(e_i) \propto E(y_i)$	$\sqrt{y}$
Igual que el caso anterior	$\sqrt{y}+\sqrt{y+1}$
$Var(e_i) \propto (E(y_i))^2$	Log(Y)
Igual que el caso anterior	Log(y+1)
$Var(e_i) \propto (E(y_i))^4$	1/y
Igual que el caso anterior	1/(y+1)
$Var(e_i) \propto E(y_i)[1-E(y_i)]$	$Sin^{-1}(\sqrt{y}\,)$

Las transformaciones se justifican de la siguiente manera:

Expandiendo en series de Taylor una función h(Y) alredededor de μ=E(Y) se obtiene

$$h(Y) \approx h(\mu) + h'(\mu)(Y-\mu) + h''(\mu)(Y-\mu)^2/2 \qquad (4.1)$$

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

$$Var(h(Y)) \approx [h'(E(y)]^2 Var(Y) \qquad (4.2)$$

Por ejemplo, si $Var(Y) \propto [E(y)]^2$ se tendrá que $[h'(E(Y))]^2 \approx$constante/$[E(y)]^2$. Luego, $h'(\mu) \approx 1/\mu$, de donde por integración resulta $h(\mu) \approx \log(\mu)$.

Haciendo un plot de residuales versus los valores ajustados de Y se puede estimar la transformación más adecuada. Aunque es mejor agrupar la variable Y, y calcular medias y desviaciones estándar para cada uno de los grupos y luego estimar la mejor linea que pasa por los puntos ($\log \overline{Y}_g$, $\log S_g^2$).

Ejemplo 2. Aplicar una transformación para establizar la varianza en el modelo de regresión para el conjunto de datos **millaje**

Solución. Si observamos el plot de residuales versus valores ajustados por el modelo de regresión, el cual aparece en la figura 4.4 podemos ver que la varianza está cambiando de alguna manera con los valores $\hat{y}$ Se ha explorado varias transformaciones del tipo potencia para la variable de respuesta y la que ha dado mejores resultados es la transformación $h(y)=y^{-1/2}$ que es

aquella correspondiente a la situación cuando la varianza de los errores es proporcional al cubo de la media de la variable de respuesta.

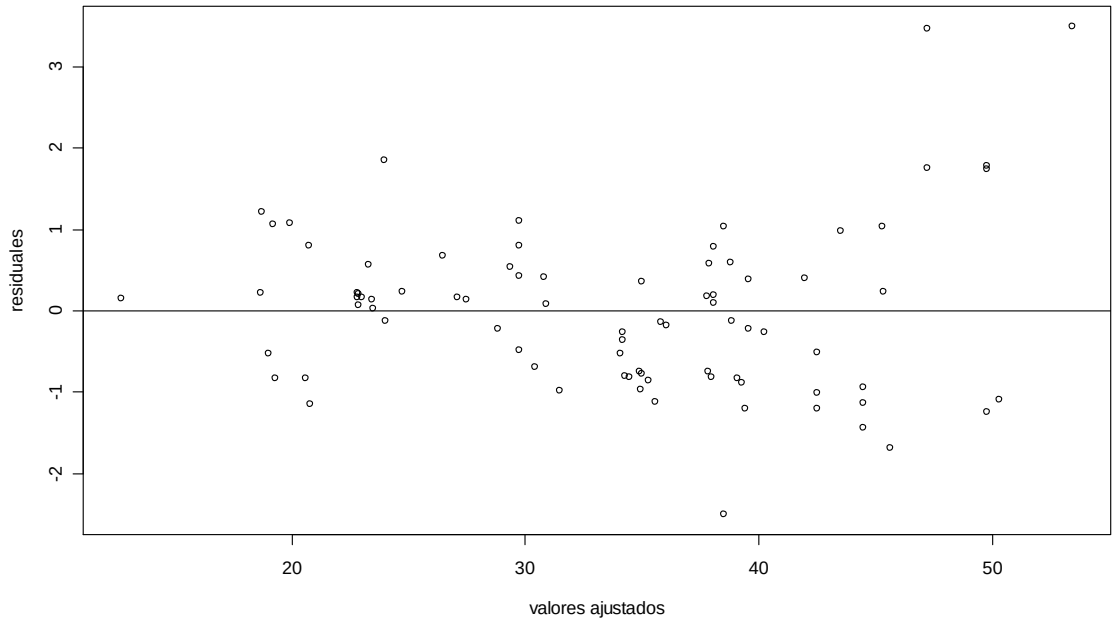

Figura 4.4. Plot de residuales estandarizados versus valores ajustados para el conjunto de datos **millaje**

```
> # El lsfit indica que la varianza es proporcional a la media al cuadrado
> # una transformacion logaritmica en la variable de respuesta es recomendada
> mpglog<-log(millaje$mpg)
> millaje1<-cbind(millaje,mpglog)

> l2<-lm(mpglog~sp+wt+vol+hp,data=millaje1)
> summary(l2)

Call:
lm(formula = mpglog ~ sp + wt + vol + hp, data = millaje1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.273816 -0.058032 -0.008837  0.038624  0.253079

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7725247  0.5647743  10.221 5.49e-16 ***
sp          -0.0130542  0.0058747  -2.222   0.0292 *
wt          -0.0370088  0.0051209  -7.227 3.08e-10 ***
vol         -0.0003088  0.0005478  -0.564   0.5746
hp           0.0029479  0.0019540   1.509   0.1355
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08767 on 77 degrees of freedom
Multiple R-Squared: 0.9211,     Adjusted R-squared: 0.917
F-statistic: 224.8 on 4 and 77 DF,  p-value: < 2.2e-16
```

```
# Considerando que la varianza es proporcional a la media al cubo
# una transformacion  h(y)=y^-0.5 es realizada
mpg05<-millaje$mpg^-0.5
millaje2<-cbind(millaje,mpg05)
l3<-lm(mpg05~sp+wt+vol+hp,data=millaje2)
summary(l3)
```

```
> summary(l3)

Call:
lm(formula = mpg05 ~ sp + wt + vol + hp, data = millaje2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.019083 -0.003005  0.001039  0.003944  0.024431

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.141e-02  5.014e-02   1.823   0.0722 .
sp          -7.386e-05  5.215e-04  -0.142   0.8878
wt           2.398e-03  4.546e-04   5.275 1.18e-06 ***
vol          1.751e-05  4.863e-05   0.360   0.7198
hp           1.621e-04  1.735e-04   0.935   0.3529
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007783 on 77 degrees of freedom
Multiple R-Squared: 0.9266,     Adjusted R-squared: 0.9228
F-statistic: 243.1 on 4 and 77 DF,  p-value: < 2.2e-16
```

El plot de residuales versus valores ajustados es como en la Figura 4.5. Notar que no se observa ningún patron de los puntos y hay dos “outliers” bien distinguibles.

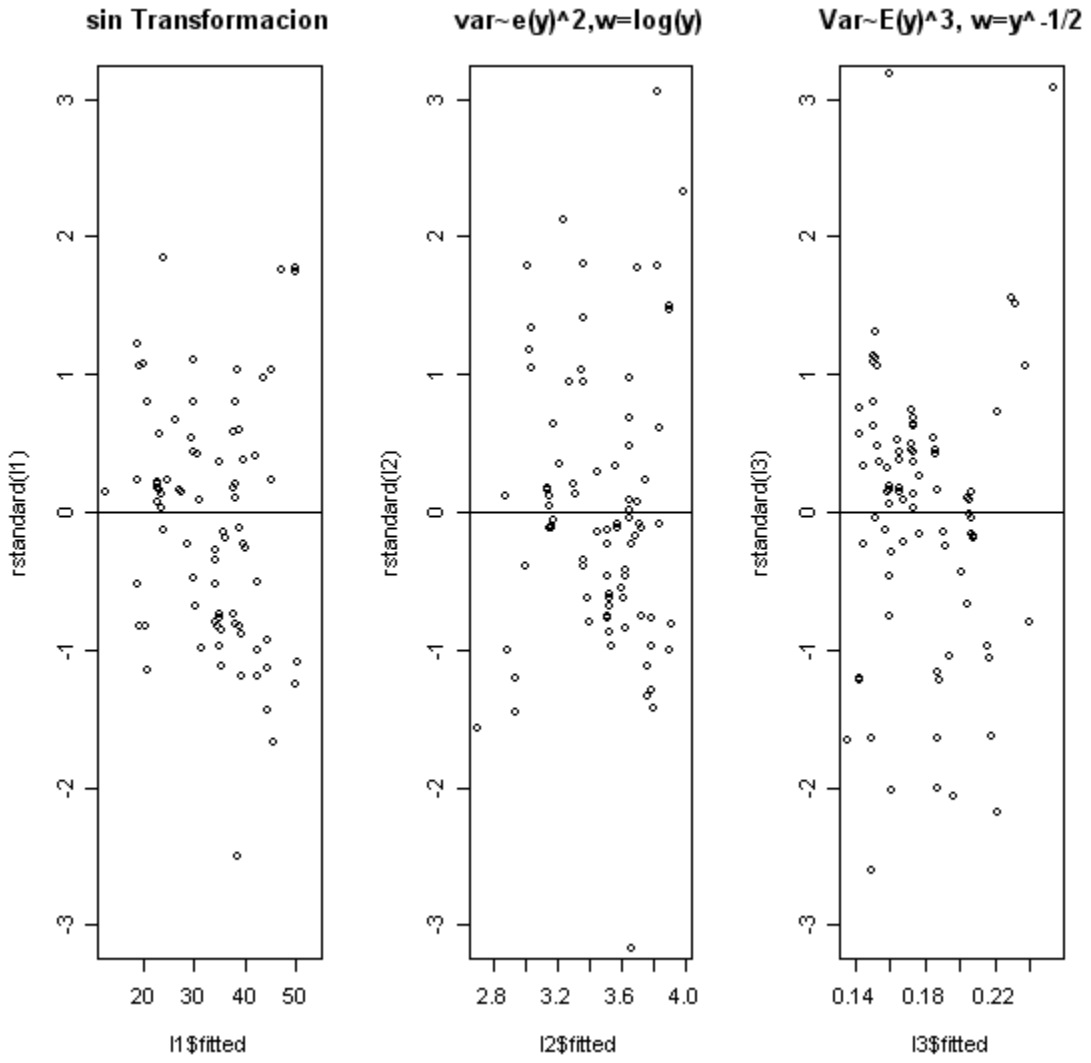

Figura 4.5. Plot de residuales versus valores ajustados después de la transformación.

4.3 Transformaciones de las variables predictoras en regresión múltiple

Supongamos que se tiene una variable de respuesta Y y varias variables predictoras, las cuales asumen valores positivos. Se desea hacer transformaciones en las variables predictoras para mejorar la medida de ajuste del modelo. Lo primero que uno intenta es hacer un plot matricial y de allí extraer las relaciones de y con cada una de las variables predictoras. Pero estas transformaciones se pueden ver afectadas por la colinealidad (dependencia lineal) existente entre las variables predictoras. Este mismo problema afecta al plot de regresión parcial o de variables añadidas.

En 1962, Box y Tidwell, propusieron un método para transformar las variables predictoras pero solo usando potencia de ellas. Mas especificamente, ellos consideraron el modelo

$$y = \beta_o + \beta_1 w_1 + ...+ \beta_k w_k + e \qquad (4.3)$$

donde $w_j = x_j^{\alpha j}$ si $\alpha_j \neq 0$ y $w_j=\ln(x_j)$ si $\alpha_j = 0$. El método está basado en el desarrollo en series de Taylor del modelo anterior con respecto a $\boldsymbol{\alpha} = (\alpha_1,.....\alpha_k)$ y alrededor del punto $\boldsymbol{\alpha_0} = (\alpha_{1,0},.....,\alpha_{k,0}) = (1,....,1)$. Haciendo las derivaciones respectivas, el modelo (4.1) se reduce a:

$$y \cong \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (\alpha_1 - 1)\beta_1 x_1 \ln x_1 + (\alpha_2 - 1)\beta_2 x_2 \ln x_2 + ... + (\alpha_k - 1)\beta_k x_k \ln x_k$$

el cual es equivalente a

$$y \cong \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma_1 z_1 + \gamma_2 z_2 + ... + \gamma_k z_k \qquad (4.4)$$

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

El procedimiento para la estimación de los α_j se puede resumir como sigue:

a) Hacer la regresión lineal múltiple considerando las variables predictoras originales x_j y denotar los estimados de los coeficientes por b_j.
b) 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 $\hat{\gamma}_j$.
c) Estimar α_j por $\hat{\alpha}_j = \dfrac{\hat{\gamma}_j}{b_j} + 1$

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

$$\hat{\alpha}_j^{(m+1)} = (\frac{\hat{\gamma}_j^{(m)}}{b_j^{(m)}} + 1)\hat{\alpha}_j^{(m)} \qquad (4.5)$$

Terminando el proceso cuando $|\alpha_j^{(m+1)} - \alpha_j^{(m)}| < TOL$, donde TOL es una cantidad de tolerancia muy cercana a cero.

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

Ejemplo 3. Aplicar la técnica sugerida por Box and Tidwell al conjunto de datos **millaje.**

Solución. Usando R se obtiene

```
> l1<-lm(mpg~.,data=millaje)
> betas<-l1$coeff
> betas
 (Intercept)          sp          wt         vol          hp
192.43775332  -1.29481848  -1.85980373  -0.01564501   0.39221231
> summary(l1)

Call:
```

```
lm(formula = mpg ~ ., data = millaje)

Residuals:
   Min     1Q Median     3Q    Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
sp           -1.29482    0.24477  -5.290 1.11e-06 ***
wt           -1.85980    0.21336  -8.717 4.22e-13 ***
vol          -0.01565    0.02283  -0.685    0.495
hp            0.39221    0.08141   4.818 7.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733,     Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

Notar que la predictora VOL no es significativa.

La regresión con las variables originales resulta ser

```
MPG = 192.4 - 0.0156 VOL + 0.392 HP - 1.294 SP - 1.859 WT
```

Ahora creamos cuatro variables predictoras $z_1=x_1lnx_1$, $z_2=x_2lnx_2$, $z_3=x_3lnx_3$ y $z_4=x_4lnx_4$. Haciendo la regresión múltiple con las 8 variables predictoras se obtiene

```
> l2<-lm(mpg~.,data=millaje1)
> betas2<-l2$coeff
> betas2
 (Intercept)          sp          wt         vol          hp          z1
1048.2022263  -38.8522423  -17.9023484   -1.0023285    5.4675149    6.3624693
          z2          z3          z4
   3.3262799    0.1803016   -0.8006012
> summary(l2)

Call:
lm(formula = mpg ~ ., data = millaje1)

Residuals:
   Min     1Q Median     3Q    Max
-8.0797 -1.4479 -0.1852  1.4320 10.1958

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1048.2022   268.3693   3.906 0.000208 ***
sp           -38.8522    11.8106  -3.290 0.001546 **
wt           -17.9023     4.3238  -4.140  9.2e-05 ***
vol           -1.0023     0.5916  -1.694 0.094470 .
```

```
hp           5.4675     1.8491   2.957 0.004185 **
z1           6.3625     1.9713   3.228 0.001871 **
z2           3.3263     0.8739   3.806 0.000291 ***
z3           0.1803     0.1086   1.660 0.101185
z4          -0.8006     0.2744  -2.917 0.004690 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 73 degrees of freedom
Multiple R-Squared: 0.905,      Adjusted R-squared: 0.8946
F-statistic: 86.97 on 8 and 73 DF,  p-value: < 2.2e-16
```

La ecuación de regression estimada resulta ser

$$MPG = 1048.2 - 38.852\ SP - 17.902\ WT - 1.002\ VOL + 5.467\ HP + 6.362\ x_1 lnx_1 + 3.326\ x_2 lnx_2 + 0.180\ x_3 lnx_3 - 0.800\ x_4 lnx_4$$

Notar que tanto VOL como la variable z_3 , relacionada a ella, son no significativas.

Aplicando el paso c) del algoritmo se tendría que

```
> gammas<-betas2[c(6:9)]
> #Hallando los alfas
> alfas<-(gammas/betas1)+1
> alfas
       z1         z2         z3         z4
 -3.9137925  -0.7885113 -10.5245410  -1.0412443
```

Haciendo la regresión con las nuevas variables $vol^{-10.52}$, $hp^{-1.04}$, $sp^{-3.91}$ y $wt^{-0.79}$ se obtiene

```
> sp1<-millaje1$sp^alfas[1]
> wt1<-millaje1$wt^alfas[2]
> vol1<-millaje1$vol^alfas[3]
> hp1<-millaje1$hp^alfas[4]
> #regresion con todas las variables transformadas
> l3<-lm(millaje1$mpg~sp1+wt1+vol1+hp1)
> summary(l3)

Call:
lm(formula = millaje1$mpg ~ sp1 + wt1 + vol1 + hp1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.34348 -1.62938 -0.07744  1.35872 10.15980

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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.095 on 77 degrees of freedom
Multiple R-Squared: 0.909,      Adjusted R-squared: 0.9043
F-statistic: 192.4 on 4 and 77 DF,  p-value: < 2.2e-16
```

Hay problemas con la variable transformada de VOL, su coeficiente estimado es enormemente grande, porque todas sus entradas se hacen demasiado pequeñas.

Repitiendo el proceso, eliminado VOL antes de aplicar el método de Box and Tidwell se obtiene que

```
> millaje2<-cbind(millaje2,z11,z21,z31)
> l21<-lm(mpg~.,data=millaje2)
> betas22<-l21$coeff
> gammas1<-betas22[c(5:7)]
> #Hallando los alfas1
> alfas1<-(gammas1/betas12)+1
> alfas1
      z11       z21       z31
-4.3033518 -0.9219605 -1.0966294
```

Luego, α_1=-1.09, α_2 =-4.30 y α_3=-0.92

```
> #Creando las nuevas variables
> sp11<-millaje2$sp^alfas1[1]
> wt11<-millaje2$wt^alfas1[2]
> hp11<-millaje2$hp^alfas1[3]
> l5<-lm(millaje2$mpg~sp11+wt11+hp11)
> summary(l5)

Call:
lm(formula = millaje2$mpg ~ sp11 + wt11 + hp11)

Residuals:
    Min      1Q  Median      3Q     Max
-8.60068 -1.61086  0.08952  1.18229 12.43902

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.084e+00  3.681e+00   0.566  0.57286
sp11        -1.507e+09  2.800e+09  -0.538  0.59186
wt11         4.503e+02  1.345e+02   3.348  0.00125 **
hp11         2.146e+03  1.052e+03   2.040  0.04475 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.167 on 78 degrees of freedom
Multiple R-Squared: 0.9035,     Adjusted R-squared: 0.8998
```

F-statistic: 243.4 on 3 and 78 DF, p-value: < 2.2e-16

Luego la regresión estimada es

```
MPG = 2.084 + 2146 hp11 + 450.3 wt11 -1.507e+09 sp11
```

Observese que la predictora SP11 no es significativa y podríamos sacarla del modelo. El cual se reduciría ahora a

```
> #Haciendo la regresion con solo las dos variables significativas
> l6<-lm(millaje2$mpg~wt11+hp11)
> summary(l6)

Call:
lm(formula = millaje2$mpg ~ wt11 + hp11)

Residuals:
    Min      1Q  Median      3Q     Max
-8.67047 -1.66461  0.04419  1.21415 12.53739

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3744     1.8522   0.202     0.84
wt11         511.1816    72.4007   7.060 5.73e-10 ***
hp11        1600.0983   280.2234   5.710 1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.153 on 79 degrees of freedom
Multiple R-Squared: 0.9031,     Adjusted R-squared: 0.9007
F-statistic: 368.3 on 2 and 79 DF,  p-value: < 2.2e-16

La ecuación de regression estimada es:
MPG = 0.374 + 1600.0 hp1 + 511.1 wt1
```

Donde $hp1=1/hp^{1.09}$ y $wt1=1/wt^{0.92}$. En la sección 2.3.5 habíamos llegado a establecer que el mejor modelo era de MPG versus hpo=1/hp y wto=1/wt. Los resultados eran como sigue:

```
> reg1=lm(mpg~hp0+wt0)
> summary(reg1)

Call:
lm(formula = mpg ~ hp0 + wt0)

Residuals:
    Min      1Q  Median      3Q     Max
-8.70343 -1.70579 -0.01131  1.20593 12.60609

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.217      1.573   0.773    0.442
```

```
hp0         1131.489    200.586   5.641 2.54e-07 ***
wt0          610.370     87.913   6.943 9.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.165 on 79 degrees of freedom
Multiple R-Squared: 0.9024,     Adjusted R-squared: 0.8999
F-statistic: 365.3 on 2 and 79 DF,  p-value: < 2.2e-16
```

Notar pues que la transformación de Box y Tidwell parece ser bastante eficiente.

4.4. Transformaciones para mejorar la normalidad de la variable de respuesta

En 1964, Box y Cox introdujeron una transformación de la variable de respuesta con el objetivo de satisfacer la suposición de normalidad del modelo de regresión. La transformación es de la forma y^{λ} (transformación potencia), donde λ es estimada con los datos tomados. Más especificamente, la transformación está definida por $w=\frac{y^{\lambda}-1}{\lambda}$ si $\lambda \neq 0$ y $w=\ln(y)$ si $\lambda=0$. Notar que $\lim_{\lambda \to 0} \frac{y^{\lambda}-1}{\lambda} = \ln y$. En la figura 4.6 se muestra la gráfica de la transformación Box-Cox para cinco valores distintos de lambda.

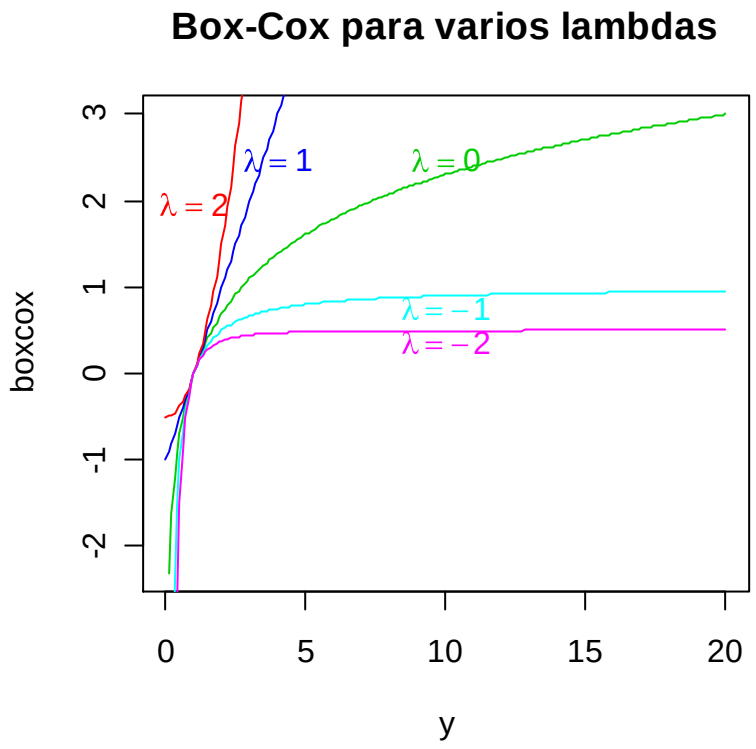

Figure 4.6 Transformación Box-Cox para varios valores de lambda

El parámetro λ se estima conjuntamente con los coeficientes del modelo de regresión lineal múltiple usando el método de Máxima verosimilitud

$$w = \beta_o + \beta_1 x_1 + ...+ \beta_k x_k + e \qquad (4.6)$$

Notar que $Var(w) \approx \mu_Y^{2(\lambda-1)}\sigma_Y^2$. Como se quiere Var($w_1$)=…..=Var($w_n$)=C, se tiene que $\prod_{i=1}^{n} \mu_{y_i}^{2(\lambda-1)}\sigma_{y_i}^2 = C^n$ .La transformación estandarizada de los w's se define por

$$z_i = \frac{w_i}{\tilde{y}^{\lambda-1}} \qquad (4.7)$$

donde $\tilde{y} = (\prod_{i=1}^{n} y_i)^{1/n}$, es la media geométrica de las y's . Luego, el modelo (4.6) se convierte en **z=Xβ+*e* .** El método asume que para algún λ las z_i's son normales e independientes con varianza común σ^2 .
Escribiendo la función de verosimilitud, correspondiente al modelo transformado, en términos de las z_i's se tiene que

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

Luego se puede establecer que el máximo del logaritmo de la función de verosimílitud está dado por:

$$LnL(\hat{\boldsymbol{\beta}},\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (4.8)$$

donde $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$, y $\hat{\sigma}^2 = SSE/n = (\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}})/n$. Luego,

$$LnL(\hat{\boldsymbol{\beta}},\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2} \equiv -\frac{n}{2}\ln(\hat{\sigma}^2) \qquad (4.9)$$

Claramente (4.9) depende de λ puesto que $\hat{\sigma}^2$ depende de **z** y ésta a su vez de λ.

El procedimiento para estimar el parámetro λ es el siguiente:

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

   z=Xβ+e

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

Varios programas estadísticos, entre ellos S-Plus y R, tienen funciones que permiten estimar el parámetro λ de la transformación Box-Cox. Además del plot del paso 3 producen un intervalo de confianza para λ.

Ejemplo 4. Aplicar la transformación de Box y Cox al conjunto de datos **millaje**

Solución: Haremos uso de R, cuya librería MASS incluye la función **boxcox**. Los resultados usando las variables originales son como sigue:

```
> reg1<-lm(MPG~VOL+HP+SP+WT,data=MILLAJE)
> summary(reg1)

Call: lm(formula = MPG ~ VOL + HP + SP + WT, data = MILLAJE)
Residuals:
   Min     1Q Median    3Q   Max
-9.011 -2.773 0.2733 1.836 11.99

Coefficients:
             Value Std. Error   t value  Pr(>|t|)
(Intercept)  192.4378   23.5316     8.1778    0.0000
        VOL   -0.0156    0.0228    -0.6854    0.4951
         HP    0.3922    0.0814     4.8176    0.0000
         SP   -1.2948    0.2448    -5.2899    0.0000
         WT   -1.8598    0.2134    -8.7166    0.0000

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733
F-statistic: 132.7 on 4 and 77 degrees of freedom, the p-value is 0

Correlation of Coefficients:
   (Intercept)     VOL      HP      SP
VOL  0.1049
 HP  0.9814      0.2324
 SP -0.9961     -0.1501 -0.9837
 WT -0.8658     -0.4260 -0.9228  0.8555
```

Aplicando la función boxcox

```
> boxcox(reg1,lambda=seq(-.6,.6,length=20),plotit=T)
```

Se obtiene el plot de la siguiente figura, donde el parámetro λ puede ser estimado por –0.22, con un intervalo de confianza de (-0.54,0.10)

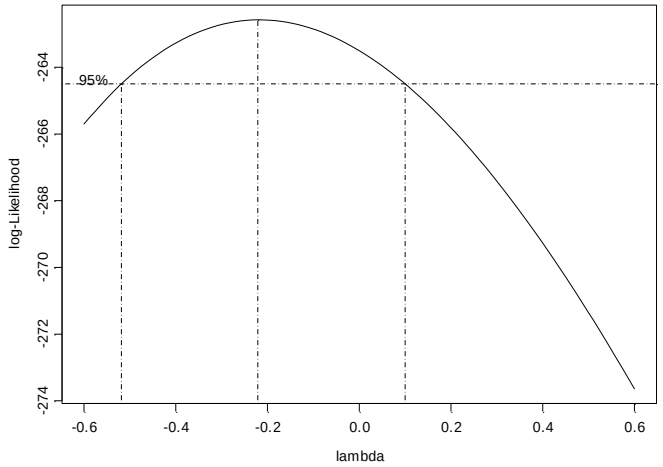

Figura 4.7. Plot de log-likelihood para varios valores de λ

Ahora veremos el efecto de la transformación

```
> millaje1=millaje
> millaje1$mpg<-((millaje$mpg)^-0.22-1)/-0.22
> reg2<-lm(mpg~vol+hp+sp+wt,data=millaje1)
> summary(reg2)

Call: lm(formula = mpg ~ vol + hp + sp + wt, data = millaje1)
Residuals:
    Min      1Q    Median      3Q    Max
 -0.128 -0.0237 -0.004189 0.01595 0.1096
Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept)  3.2290   0.2600   12.4214   0.0000
        vol -0.0001   0.0003   -0.4866   0.6279
         hp  0.0004   0.0009    0.4573   0.6488
         sp -0.0033   0.0027   -1.2145   0.2283
         wt -0.0152   0.0024   -6.4641   0.0000

Residual standard error: 0.04035 on 77 degrees of freedom
Multiple R-Squared: 0.9252
F-statistic: 238.2 on 4 and 77 degrees of freedom, the p-value is 0
Correlation of Coefficients:
   (Intercept)     vol        hp         sp
vol  0.1049
hp   0.9814      0.2324
sp  -0.9961     -0.1501 -0.9837
wt  -0.8658     -0.4260 -0.9228  0.8555
```

Los plots para cotejar normalidad de los residuales se muestra en la figura 4.8

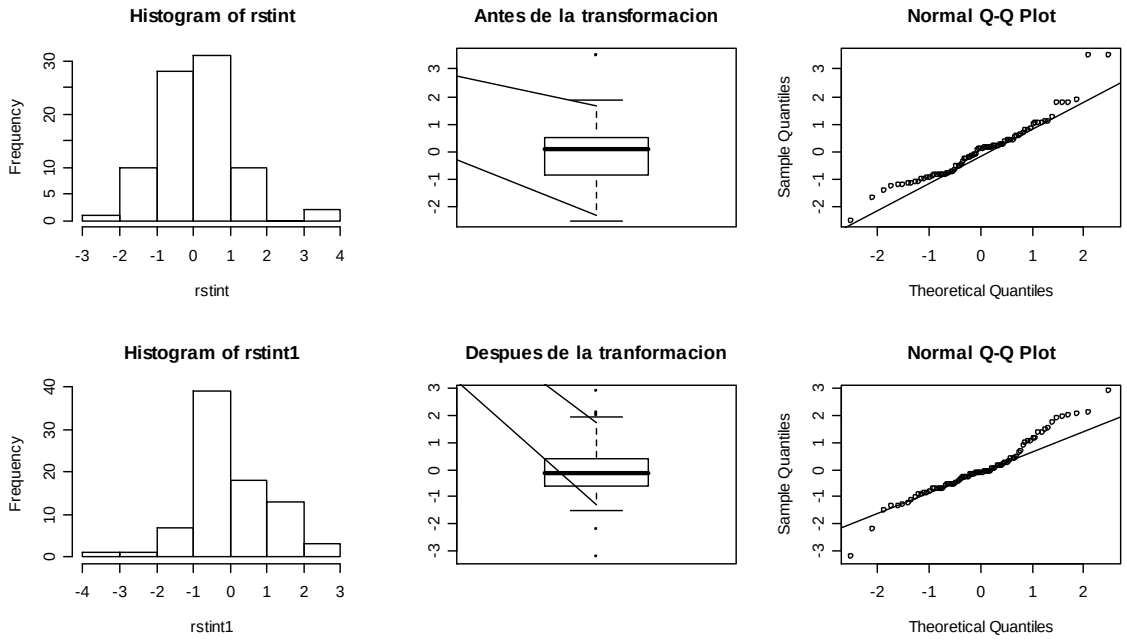

Figura 4.8. Plots para ver el efecto de la transformación Box-Cox en la distribución de los residuales de la regresión para el conjunto de datos millaje.

Notar que los puntos están mejor alineados que en plot con las variables originales (ver figura 4.8) especialmente en la parte central. Se observan claramente dos "outliers" inferiores y uno superior. Notar que el R^2 ha subido de 87.33% a 92.52%, mejorando el efecto de transformar las variables predictoras que se llevó a cabo en el ejemplo 3.

4.5 Mínimos cuadrados ponderados.

Otra manera de tratar de remediar la falta de homogeneidad de varianza de los errores es usar mínimos cuadrados ponderados, suponiendo que los errores son todavía no correlacionados. En este caso 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. Por ejemplo, si en el plot de residuales versus la variable predictora se observa que la dispersión aumenta cuando x aumenta sería conveniente usar $w_i = \frac{1}{\sigma_i^2}$. Aquí, σ_i^2 son las varianzas poblacionales de la Y 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. Obviamente estas varianzas no son conocidas y deben ser estimadas por sus varianzas muestrales s_i^2. Si hay solamente una observación *y* para el valor x_i entonces se consideran valores de y corespondientes a valores cercanos a x_i. En otras palabras la variable x es considerada agrupada.
Esta no es la única manera de escoger los pesos, en regresión robusta que será tratada en el capítulo 8, se hacen distintos cálculos de los pesos con la idea de dar a las observaciones

anómalas un menor peso. El cálculo de los pesos está basado mayormente en los diagnósticos de reesión.

Consideremos el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e} \qquad (4.10)$$

con $\mathbf{Var(e)=V}\sigma^2$, donde **V** es una matriz diagonal. Es decir,

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

Sea $\mathbf{W}=(\mathbf{V}^{1/2})^{-1}$, claramente, $\mathbf{W'W=V^{-1}}$. Multiplicando ambos lados del modelo lineal (4.10) por W se obtiene

$$\mathbf{Wy} = \mathbf{WX\beta} + \mathbf{We} \qquad (4.11)$$

Sea $\mathbf{y^*}=\mathbf{Wy}$, $\mathbf{e^*}=\mathbf{We}$ y $\mathbf{X^*}=\mathbf{WX}$, entonces el modelo (4.11) se convierte en el modelo

$$\mathbf{y}^*=\mathbf{X^*\beta} + \mathbf{e^*} \qquad (4.12)$$

Notar que $\text{Var}(\mathbf{e^*})=\text{Var}(\mathbf{We})=\mathbf{W}\text{var}(\mathbf{e})\mathbf{W'}=\mathbf{WVW'}\sigma^2=\mathbf{I}\sigma^2$, así que la varianza de los errores es constante. Luego el estimador mínimo cuadrático de $\boldsymbol{\beta}$ será

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

Se puede ver fácilmente que $\mathbf{E}(\boldsymbol{\beta}^*)=\boldsymbol{\beta}$ y que

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

Ejemplo 5: Consideremos las variables MPG y WT del conjunto de datos **millaje** y que además la primera y última observacion han sido eliminadas. El plot de residuales versus la variable predictora WTO (WT excluyendo la primera y última observación) aparece en la fiura 4.9.

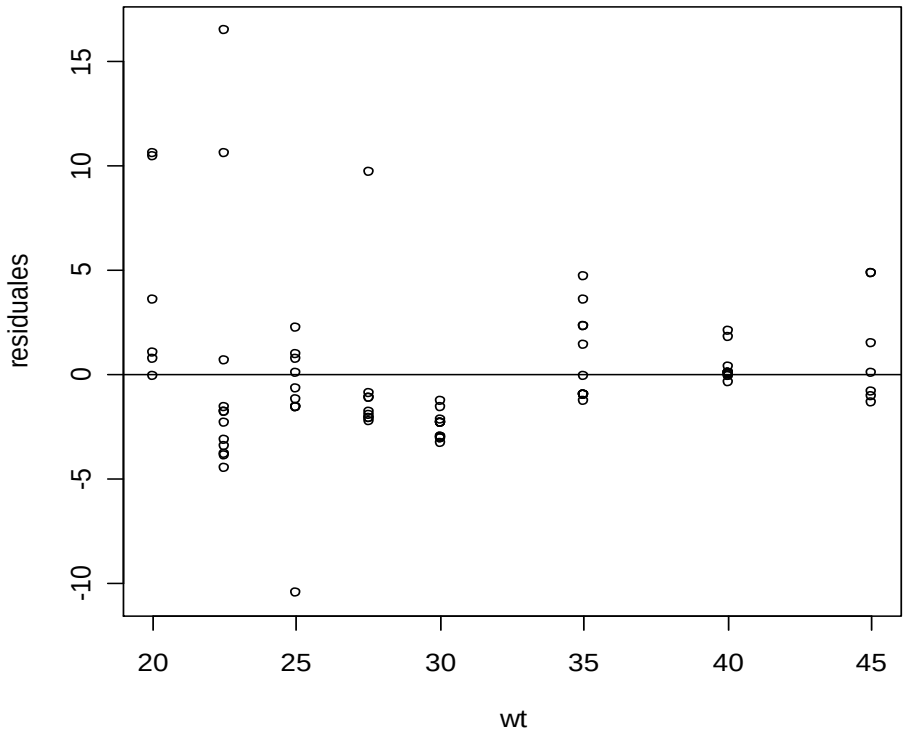

Figura 4.9. Plot de residuales versus la variable WTO donde se observa que la varianza no es homogénea

Aunque es díficil verlo en forma definitiva la variabilidad de los residuales está disminuyendo cuando la variable predictora aumenta. Algo más formal sería calcular la varianza de las y's por cada valor de X. esto produce la siguiente tabla de valores.

X_i	n_i	s_i^2
20.0	6	23.8987
22.5	12	42.7533
25.0	10	12.0049
27.5	9	14.5586
30.0	12	0.3961
35.0	12	4.4627
40.0	12	0.5973
45.0	7	7.2948

Haciendo un plot de x_i versus s_i^2 parece haber una buena relación cuadrática entre ambas variables. En la figura 4.9 se observa el plot de puntos y la regresión cuadrática La ecuación del modelo resulta ser

s_i^2 = 148.482 - 7.81488 Xi + 0.103609 Xi**2

El R^2=64.7.
Para determinar los pesos hay dos alternativas:

Primera alternativa: (Myers pag.281 Weisberg pag 85). Usar $w_i = \frac{1}{s_i^2}$, los s_i^2 están dados en la tabla anterior. Para usar esta alternativa debería haber un número razonable de observaciones y's para cada X_i.

Segunda Alternativa: (Draper y Smith, pag 226). Usar la ecuación de la regresión cuadrática para estimar las varianzas muestrales s_i^2 para cada x_i. Luego, escogemos los pesos como el reciproco de la varianzas muestrales estimadas. Es decir, $w_i = \frac{1}{\hat{s}_i^2}$, donde $\hat{s}_i^2$ es el valor correspondiente a un X_i en el modelo cuadrático.

Los resultados que se obtienen en R son los siguientes:

a) Análisis sin usar regresión ponderada.

```
> millaje1<-millaje[-c(1,82),c(1,3)]
> l1=lm(mpg~wt,data=millaje1)
> summary(l1)

Call:
lm(formula = mpg ~ wt, data = millaje1)

Residuals:
   Min     1Q Median     3Q    Max
-10.484  -1.992  -1.017   0.720  16.474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.40171    1.78634   37.73   <2e-16 ***
wt          -1.09671    0.05635  -19.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.825 on 78 degrees of freedom
Multiple R-Squared: 0.8293,     Adjusted R-squared: 0.8271
F-statistic: 378.8 on 1 and 78 DF,  p-value: < 2.2e-16
```

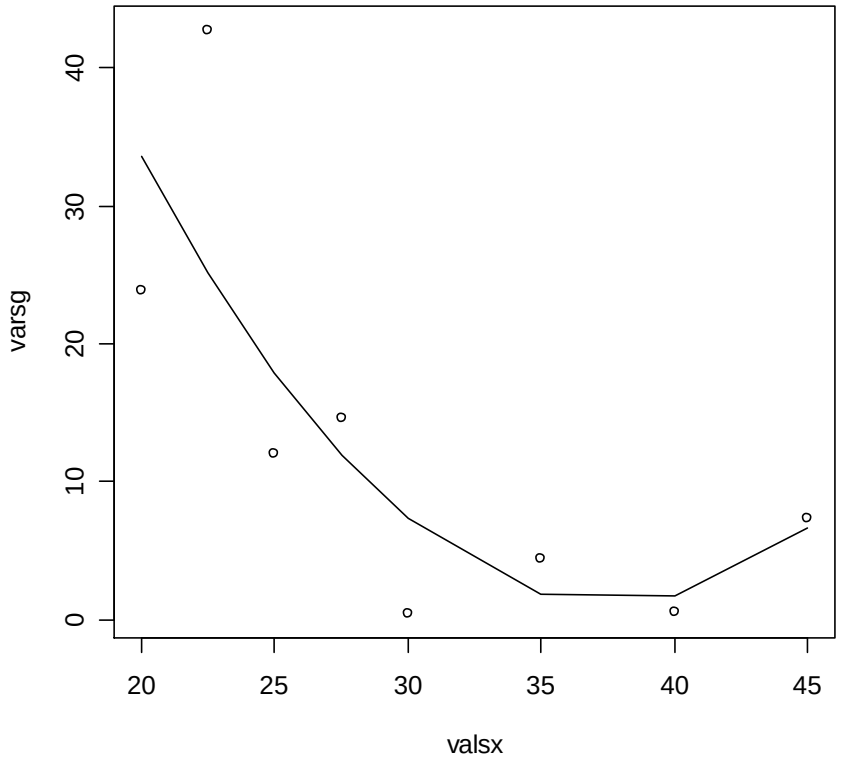

Figura 4.10. Ajuste cuadrático de la varianza versus la variable predictora

B) Análisis de regresión ponderada con la alternativa a).

```
> summary(lw1)

Call:
lm(formula = mpg ~ ., data = millaje1, weights = pesos)

Residuals:
   Min     1Q  Median     3Q    Max
-2.1007 -0.1768  0.1953  1.0526  3.2224

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.51692    1.06539   54.92   <2e-16 ***
wt          -0.86954    0.03107  -27.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.206 on 78 degrees of freedom
Multiple R-Squared: 0.9094,     Adjusted R-squared: 0.9083
```

F-statistic: 783.4 on 1 and 78 DF, p-value: < 2.2e-16

C) Análisis de Regresión ponderada usando la alternativa b)

```
> lw2<-lm(mpg~.,data=millaje1,weights=pesos1)
> summary(lw2)

Call:
lm(formula = mpg ~ ., data = millaje1, weights = pesos1)

Residuals:
   Min     1Q  Median     3Q    Max
-2.3321 -0.7047 -0.3122  0.2965  3.4499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.94848    1.72500   37.65   <2e-16 ***
wt          -1.02365    0.04738  -21.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.082 on 78 degrees of freedom
Multiple R-Squared: 0.8568,     Adjusted R-squared: 0.855
F-statistic: 466.8 on 1 and 78 DF,  p-value: < 2.2e-16
```

Observese que cuando se hace la regresión ponderada con la alternativa a) se obtiene una mejora del 7.0% en el R^2 mientras que con la alternativa b) solo se mejora un 2.8% .

4.6 Mínimos Cuadrados generalizados

Consideremos ahora la situación más general de que los errores no tiene varianza constante y además que son correlacionados. Sea el modelo de regresión lineal múltiple

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

Supongamos ahora que $\mathbf{Var(e)=V}\sigma^2$, donde $\mathbf{V}$ es una matriz simétrica y definida positiva. Un caso particular de $\mathbf{V}$ es cuando los errores tienen distinta varianza y no están correlacionados.

Siempre es posible encontrar una matriz nosingular y simétrica $\mathbf{T}$ tal que $\mathbf{TT=T^2=V.}$ Mutiplicando ambos lados del modelo anterior por $\mathbf{T^{-1}}$ se obtiene

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

Sea $\mathbf{e^*}= \mathbf{T^{-1}e,}$ notando que $\text{Var}(\mathbf{e^*})=\text{Var}(\mathbf{T^{-1}e})=\mathbf{T^{-1}}\text{Var}(\mathbf{e})\mathbf{T^{-1}}=\mathbf{I}\sigma^2$ entonces el estimador mínimo cuadrático de $\boldsymbol{\beta}$ se obtiene minimizando

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

lo cual produce $\boldsymbol{\beta^* = (X' V^{-1} X)^{-1} X' V^{-1} Y}$. Se puede ver fácilmente que $\mathbf{E(\beta^*)=\beta}$ y que $\mathbf{Var(\beta^*)= (X' V^{-1} X)^{-1} X' V^{-1} } Var(\mathbf{Y}) \mathbf{V^{-1} X (X' V^{-1} X)^{-1}} = \mathbf{(X' V^{-1} X)^{-1}} \sigma^2$.

Ejercicios

1. Usar el conjunto de datos **Highway**, con variable de respuesta es RATE y todas las otras como variables predictoras para responder las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R^2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para establizar la varianza
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

2. Usar el conjunto de datos **Fuel** con variable de respuesta es Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para estabilizara la varianza
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

3. Usar el conjunto de datos **Headcirc** con variable de respuesta es headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas.

a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el R^2 de la regresión usando como predictora la que tiene mas alta correlacion
b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
d) Aplicar una transformación tipo potencia para estabilizar la varianza.
e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.

4. Verificar la relación 4.7

5. Prueba para detectar varianza no constante (Cook y Weisberg, 1983). Consiste de los siguientes pasos:

a) Calcular la regression de Y versus todas las variables predictoras y guardar los residuales $\hat{e}_i$.

b) Calcular los residuales cuadrados escalados u_i definidos por $u_i = \frac{\hat{e}_i}{\tilde{\sigma}^2}$, donde $\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n}$ es el estimado máximo verosímil de σ^2.

c) Calcular la regresión de u_i versus las $\mathbf{z}_i$ incluyendo el intercepto. Los $\mathbf{z}_i$ son las variables de las que se sospecha que depende la varianza σ^2. Así, $\mathbf{z}_i = \hat{y}_i$ indica que la varianza varia con la variable de respuesta, $\mathbf{z}_i = x_i$ indica que la varianza varía con la predictora x_i. También $\mathbf{z}_i$ pueden contener todas las variables predictoras o un subconjunto de ellas. Guardar la Suma de cuadrado de la regresión (SSR).

d) Calcular la prueba S=SSR/2. Si S es grande entonces hay indicación de varianza no constante. Mas formalmente S se distribuye asintóticamente como una Ji cuadrado con q grados de libertad, donde q es el número de componentes de z_i, bajo la hipótesis nula de varianza constante.
Aplicar la prueba definida por los pasos a –d a los datos de los ejercicios 1 y 3.

6) Deducir que transformación de la variable de respuesta hace que la varianza σ^2 sea constante cuando ella es proporcional a $[E(y)]^4$