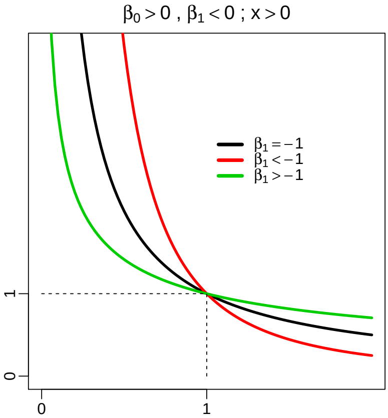
3 RLS – Transformaciones



−ebz

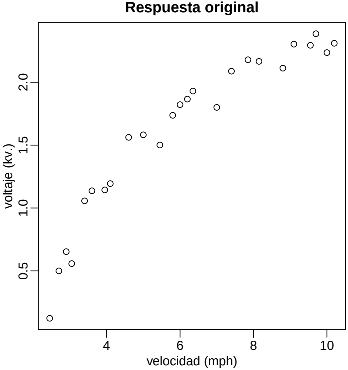
Contenido

Transformaciones a una Línea Recta

Ejemplo: Generación de electricidad mediante molino de viento[1]
Un ingeniero investiga la posibilidad de generar electricidad mediante un molino de viento. Después de un tiempo ha registrado la corriente eléctrica que sale del molino y la velocidad del viento.

Datos:

obs.	velocidad mph.	voltaje kv.	obs.	velocidad mph.	voltaje kv.
1	5.00	1.58	14	5.80	1.74
2	6.00	1.82	15	7.40	2.09
3	3.40	1.06	16	3.60	1.14
4	2.70	0.50	17	7.85	2.18
5	10.00	2.24	18	8.80	2.11
6	9.70	2.39	19	7.00	1.80
7	9.55	2.29	20	5.45	1.50
8	3.05	0.56	21	9.10	2.30
9	8.15	2.17	22	10.20	2.31
10	6.20	1.87	23	4.10	1.19
11	2.90	0.65	24	3.95	1.14
12	6.35	1.93	25	2.45	0.12
13	4.60	1.56			



[1] Montgomery and Peck (1991)

Ejemplo: Generación de Electricidad (cont.)

Ajuste del Modelo: $y = \beta_0 + \beta_1 x + \epsilon$

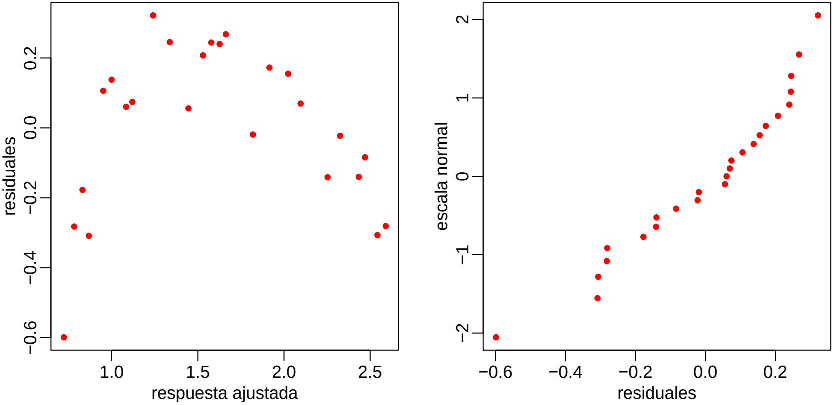
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13088    0.12599   1.039     0.31
velocidad    0.24115    0.01905  12.659 7.55e-12

Residual standard error: 0.2361 on 23 degrees of freedom
Multiple R-Squared: 0.8745,     Adjusted R-squared: 0.869
F-statistic: 160.3 on 1 and 23 DF,  p-value: 7.546e-12

Analysis of Variance Table

Response: voltaje
          Df Sum Sq Mean Sq F value    Pr(>F)
velocidad  1 8.9296  8.9296  160.26 7.546e-12
Residuals 23 1.2816  0.0557
```

−ebz

Ejemplo: Generación de Electricidad (cont.)

Validación del Modelo

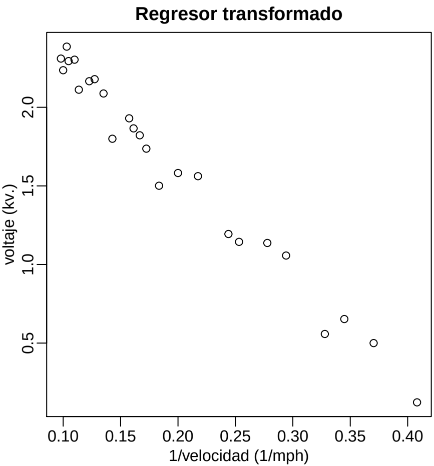
Análisis de residuales datos originales



−ebz

Ejemplo: Generación de Electricidad (cont.)

Identificación del Modelo – Transformación $X = 1/x$



–ebz

Ejemplo: Generación de Electricidad (cont.)

Ajuste del Modelo: $y = \beta_0 + \beta_1/x + \epsilon$

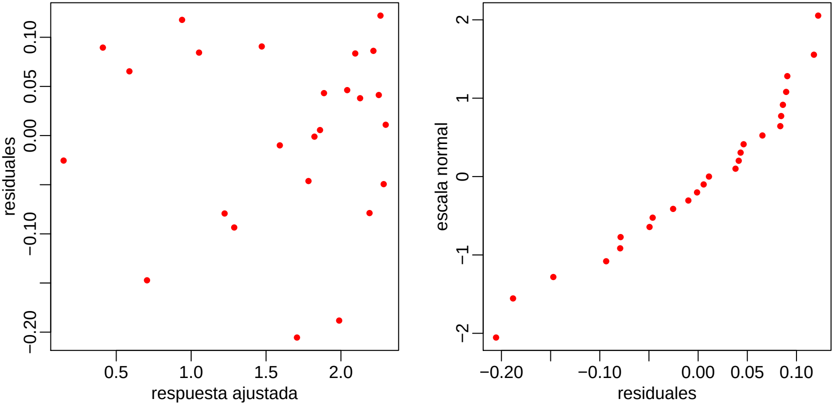
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9789     0.0449   66.34   <2e-16
invX         -6.9345     0.2064  -33.59   <2e-16

Residual standard error: 0.09417 on 23 degrees of freedom
Multiple R-Squared:  0.98,      Adjusted R-squared: 0.9792
F-statistic:  1128 on 1 and 23 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: voltaje
          Df  Sum Sq Mean Sq F value    Pr(>F)
invX       1 10.0072 10.0072  1128.4 < 2.2e-16
Residuals 23  0.2040  0.0089
```

–ebz

Ejemplo: Generación de Electricidad (cont.)

Validación del Modelo

Análisis de residuales datos transformados



−ebz

Transformaciones a una Línea Recta

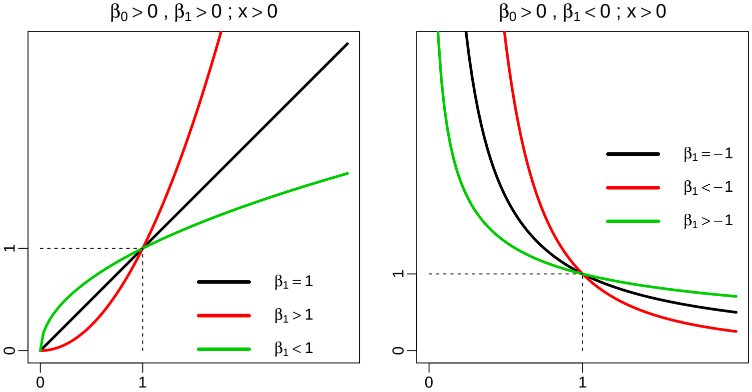
Funciones Linealizables y Formas Lineales

Función Linealizable	Transformación	Forma Lineal
$y = \beta_0 x^{\beta_1}$	$Y = \log y,\ X = \log x$	$Y = \log \beta_0 + \beta_1 X$
$y = \beta_0 e^{\beta_1 x}$	$Y = \log y$	$Y = \log \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$X = \log x$	$y = \beta_0 + \beta_1 X$
$y = \frac{x}{\beta_0 x + \beta_1}$	$Y = \frac{1}{y},\ X = \frac{1}{x}$	$Y = \beta_0 + \beta_1 X$

–ebz

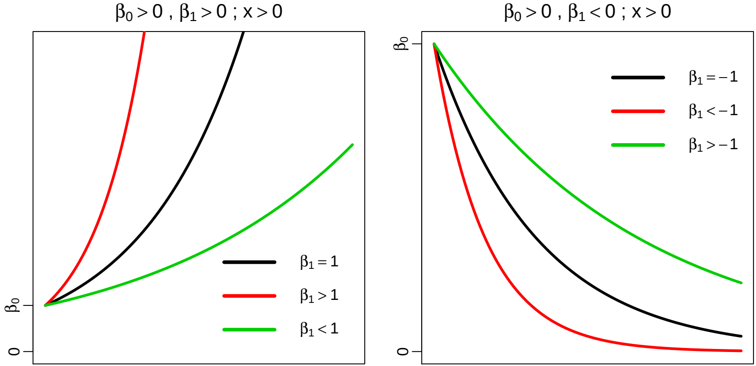
Modelo:

$$y = \beta_0 \, x^{\beta_1}$$



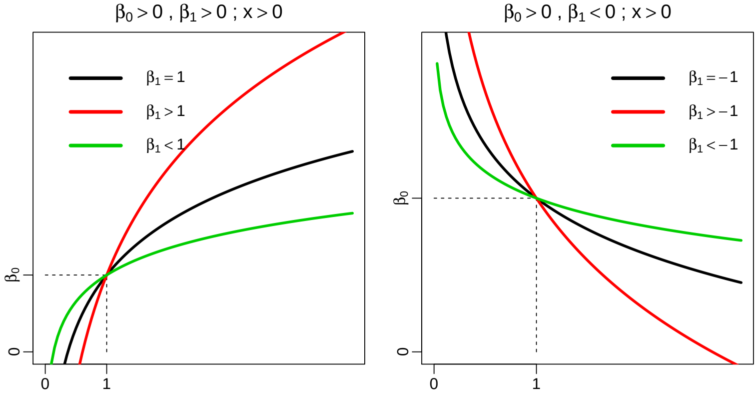
Modelo:

$$y = \beta_0 \, e^{\beta_1 x}$$



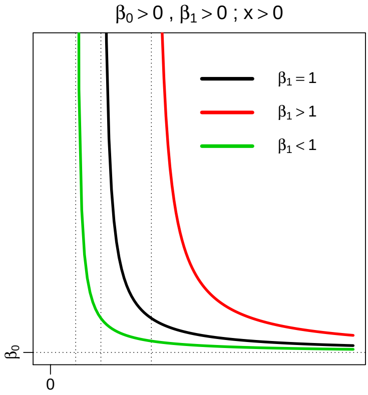
Modelo:

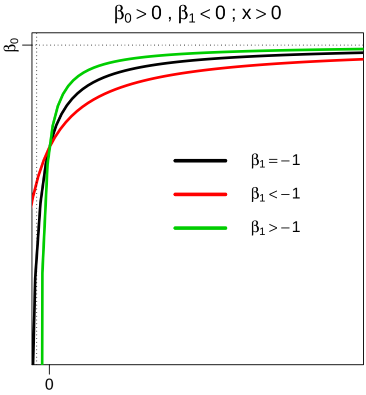
$$y = \beta_0 + \beta_1 \log(x)$$



Modelo:

$$y = \frac{x}{\beta_0 x + \beta_1}$$





−ebz

Transformaciones estabilizadoras de la varianza

Relación	Transformación	Notas
$\sigma^2 \propto k$	$Y = y$	Sin transformación
$\sigma^2 \propto \mathbb{E}(y)$	$Y = \sqrt{y}$	Raíz cuadrada (datos Poisson)
$\sigma^2 \propto \mathbb{E}(y)[1 - \mathbb{E}(y)]$	$Y = \arcsin\sqrt{y}$	Arco seno (proporciones binomiales) $0 \leq y \leq 1$
$\sigma^2 \propto [\mathbb{E}(y)]^2$	$Y = \log y$	Logaritmo
$\sigma^2 \propto [\mathbb{E}(y)]^3$	$Y = 1/\sqrt{y}$	Recíproco raíz cuadrada
$\sigma^2 \propto [\mathbb{E}(y)]^4$	$Y = 1/y$	Recíproco

−ebz

Fórmula de Transmisión de Error - Método Delta [2]

Método Delta

Sea Y una variable aleatoria (*v. a.*) con al menos sus primeros 2 momentos finitos ($\mathbb{E}[Y] = \mu_Y < \infty$ y $\text{var}(Y) = \sigma_Y^2 < \infty$) y sea $h(\cdot)$ una función *suave*, al menos 2 veces diferenciable. Luego,

$$\begin{aligned} \mathbb{E}[h(Y)] &\approx h(\mu_Y) + h^{(2)}(\mu_Y)\frac{\sigma_Y^2}{2} \\ \text{var}(h(Y)) &\approx \left(h^{(1)}(\mu_Y)\right)^2 \sigma_Y^2 \end{aligned}$$

Ejemplo: Sea Y *v. a.* con varianza $\sigma_Y^2 \propto \mu_Y^2$. Entonces $W = \log(Y)$ tiene varianza constante. Sea $W = h(Y) = \log(Y)$. Entonces $h'(y) = 1/y$

$$\sigma_W^2 \approx [h'(\mu_Y)]^2 \sigma_Y^2 = \left[\frac{1}{\mu_Y}\right]^2 \sigma_Y^2 \propto \frac{1}{\mu_Y^2} \mu_Y^2 \equiv 1$$

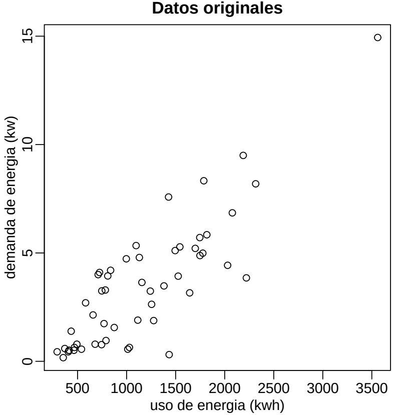
Esto es, $\sigma_W^2 \propto k$.

[2] Dudewicz and Mishra (1988), Casella and Berger (2002)

Ejemplo: Demanda de energía y uso de energía [3]

Una compañía generadora de electricidad está interesada en modelar la demanda en *horas pico* (y) como función del uso mensual total (x).

obs.	x (kwh)	y (kw)	obs.	x (kwh)	y (kw)
1	679	0.79	27	837	4.20
2	292	0.44	28	1748	4.88
3	1012	0.56	29	1381	3.48
4	493	0.79	30	1428	7.58
5	582	2.70	31	1255	2.63
6	1156	3.64	32	1777	4.99
7	997	4.73	33	370	0.59
8	2189	9.50	34	2316	8.19
9	1097	5.34	35	1130	4.79
10	2078	6.85	36	463	0.51
11	1818	5.84	37	770	1.74
12	1700	5.21	38	724	4.10
13	747	3.25	39	808	3.94
14	2030	4.43	40	790	0.96
15	1643	3.16	41	783	3.29
16	414	0.50	42	406	0.44
17	354	0.17	43	1242	3.24
18	1276	1.88	44	658	2.14
19	745	0.77	45	1746	5.71
20	435	1.39	46	468	0.64
21	540	0.56	47	1114	1.90
22	874	1.56	48	413	0.51
23	1543	5.28	49	1787	8.33
24	1029	0.64	50	3560	14.94
25	710	4.00	51	1495	5.11
26	1434	0.31	52	2221	3.85
			53	1526	3.93



[3] Montgomery and Peck (1992)

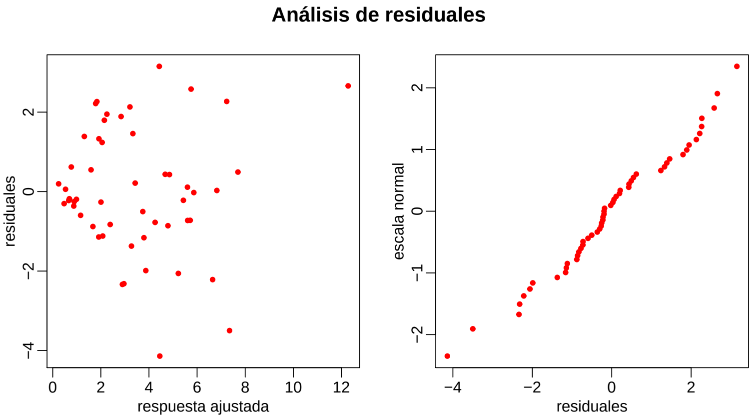
Ejemplo: Demanda de energía (cont.)

Ajuste del Modelo: $y = \beta_0 + \beta_1 x + \epsilon$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8313037  0.4416121  -1.882   0.0655
x            0.0036828  0.0003339  11.030 4.11e-15

Residual standard error: 1.577 on 51 degrees of freedom
Multiple R-Squared: 0.7046,     Adjusted R-squared: 0.6988
F-statistic: 121.7 on 1 and 51 DF,  p-value: 4.106e-15

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 302.633 302.633  121.66 4.106e-15
Residuals 51 126.866   2.488
```

Ejemplo: Demanda de energía (cont.)

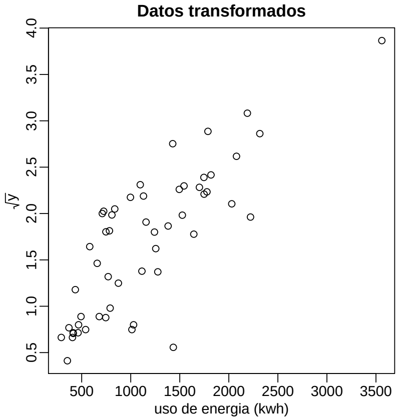
Validacion del Modelo:



−ebz

Ejemplo: Demanda de energía (cont.)

Identificación del Modelo – Transformación $Y = \sqrt{y}$



Ejemplo: Demanda de energía (cont.)

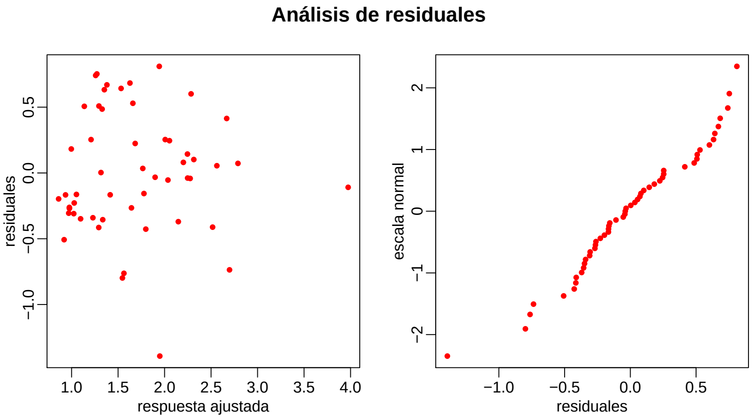
Ajuste del Modelo: $\sqrt{y} = \beta_0 + \beta_1 x + \epsilon$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.822e-01  1.299e-01   4.481 4.22e-05
x           9.529e-04  9.824e-05   9.699 3.61e-13

Residual standard error: 0.464 on 51 degrees of freedom
Multiple R-Squared: 0.6485,     Adjusted R-squared: 0.6416
F-statistic: 94.08 on 1 and 51 DF,  p-value: 3.614e-13

Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 20.2585 20.2585  94.078 3.614e-13
Residuals 51 10.9822  0.2153
```

–ebz

Ejemplo: Demanda de energía (cont.)

Validacion del Modelo:



−ebz

Box & Cox, 1964

1964] 211

An Analysis of Transformations

By G. E. P. Box and D. R. Cox

University of Wisconsin *Birkbeck College, University of London*

[Read at a Research Methods Meeting of the Society, April 8th, 1964, Professor D. V. Lindley in the Chair]

Summary

In the analysis of data it is often assumed that observations $y_1, y_2, \ldots, y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in a set of parameters $\boldsymbol{\theta}$. In this paper we make the less restrictive assumption that such a normal, homoscedastic, linear model is appropriate after some suitable transformation has been applied to the y's. Inferences about the transformation and about the parameters of the linear model are made by computing the likelihood function and the relevant posterior distribution. The contributions of normality, homoscedasticity and additivity to the transformation are separated. The relation of the present methods to earlier procedures for finding transformations is discussed. The methods are illustrated with examples.

–ebz

Transformación estabilizadora de la varianza: Box-Cox[4]

Transformación potencia Box-Cox

Ajuste el modelo de regresión lineal simple a la respuesta

$$Y = \begin{cases} y^{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

Para determinar qué λ utilizar, considere

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \log y, & \lambda = 0 \end{cases}$$

donde, $\dot{y} = \left(\Pi_{i=1}^{n} y_i\right)^{1/n}$, es el *promedio geométrico* de las respuestas y_i.

Así pues, ajuste

$$y^{(\lambda)} = \beta_0 + \beta_1 x + \epsilon$$

para varios valores de λ y elija λ^* que minimice la suma de cuadrados de los residuales $\mathrm{SC}_{\mathrm{Res}}(\lambda)$.

Intervalo (aproximado) del $100(1-\alpha)$ *% de confianza* para λ:

$$\mathrm{SC}^* = \mathrm{SC}_{\mathrm{Res}}(\lambda^*)\left(1 + \frac{t^2_{(1-\alpha/2,\nu)}}{\nu}\right)$$

donde $\nu(= n-2)$ son los grados de libertad de los residuales.

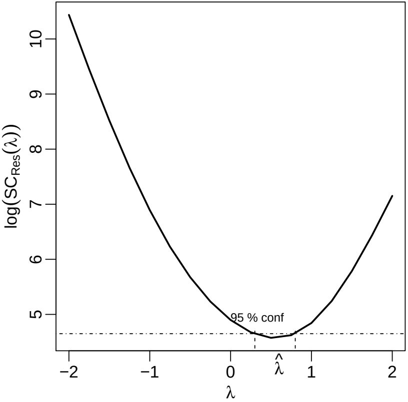
[4] Box and Cox (1964)

Ejemplo: Demanda de energía (cont.)

Variable respuesta: $y^{(\lambda)}$

λ	$SC_{Res}(\lambda)$	$\log[SC_{Res}(\lambda)]$
−2.00	34100.6	10.44
−1.75	12716.2	9.45
−1.50	5014.7	8.52
−1.25	2126.2	7.66
−1.00	986.0	6.89
−0.75	507.3	6.23
−0.50	291.6	5.68
−0.25	187.3	5.23
0.00	134.1	4.90
0.25	107.2	4.67
0.50	**96.9**	**4.57**
0.75	101.7	4.62
1.00	126.9	4.84
1.25	188.8	5.24
1.50	325.7	5.79
1.75	623.5	6.44
2.00	1275.6	7.15

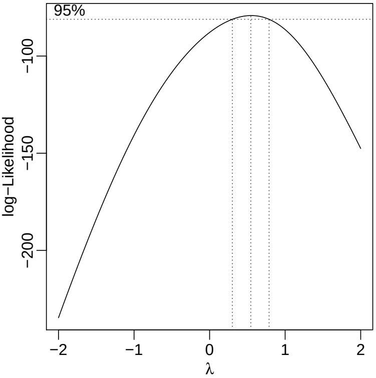
$$SC^* = 96.9 \cdot \left(1 + \frac{2.007^2}{51}\right) = 104.46$$



Ejemplo: Demanda de energía (cont.)

Ajuste y transformación:

```
print(summary(mod <- lm(y ~ x,data=dat)))
library(MASS)
boxcox(mod)

Call:
lm(formula = y ~ x, data = dat)

Residuals:
Min       1Q  Median       3Q     Max
-4.1399 -0.8275 -0.1934  1.2376  3.1522

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8313037  0.4416121  -1.882   0.0655
x            0.0036828  0.0003339  11.030 4.11e-15

Residual standard error: 1.577 on 51 degrees of freedom
Multiple R-squared:  0.7046,Adjusted R-squared:  0.6988
F-statistic: 121.7 on 1 and 51 DF,  p-value: 4.106e-15
```



Transformación estabilizadora de la varianza: Box-Cox (cont.)

Observaciones:

1. Cuando el error tiene varianza constante se llama *homoscedástico*; si no es constante, *heteroscedástico*.
2. ¿Cómo saber qué función proponer para estabilizar la varianza?
   - Si el intervalo de confianza incluye al cero, el modelo es multiplicativo.
   - Si el intervalo incluye al uno, indica que no hay que hacer una transformación.
3. La transformación de Box-Cox ayuda a *estabilizar la varianza* y *normalizar* los datos.
4. La transformación de Box-Cox es continua en $\lambda = 0$.
5. En la práctica,cuando lo permita el intervalo de confianza, utilice valores de λ *"fáciles de interpretar"*. Por ejemplo,

$$\begin{array}{lcccc} \lambda = 0.45 & \rightarrow & \lambda \equiv 0.5 & \Longrightarrow & y^{\lambda} = \sqrt{y} \\ \lambda = -0.10 & \rightarrow & \lambda \equiv 0.0 & \Longrightarrow & y^{\lambda} = \log(y) \end{array}$$

6. El método de estimación de $\hat{\lambda}$ es el mismo en el caso de la *regresión lineal múltiple*.

Transformaciones Potencia de Box-Cox

7. Box y Cox sugieren la estimación de λ, β_0 y β_1 de manera conjunta y por máxima verosimilitud.
8. En la práctica, se utiliza el *perfil de la verosimilitud*: para distintos valores de λ, se obtienen $\hat{\beta}_0$ y $\hat{\beta}_1$ y se elige $\hat{\lambda}$ tal que minimice la SC_{Error}.
9. Una partición del intervalo $[-2, 2]$ de longitud 0.25 es apropiada.
10. Transformación utilizada en varias áreas estadísticas, no solamente en modelos lineales.
11. Box y Cox (1964) es de los artículos más referenciados en la literatura estadística. Es la transformación más usada pero no la única.
12. Hay otras familias de transformaciones o procedimientos utilizados en la regresión. Vea por ejemplo Carroll and Ruppert (1988) y el paquete `car` de R.

Modelo no lineal de Michaelis-Menten [5]

Ejemplo: Velocidad de reacción como función de la concentración

El modelo de *Michaelis-Menten* es utilizado en química cinética para modelar la velocidad inicial y de una reacción enzimática con la concentración x del substrato. El modelo está dado por:

$$y = f(x, \theta) + \epsilon = \frac{\theta_1 x}{\theta_2 + x} + \epsilon$$

que se puede linealizar de la siguiente manera:

$$Y = \frac{1}{f(x, \theta)} = \beta_0 + \beta_1 X$$

donde

$$Y = \frac{1}{y}; \qquad X = \frac{1}{x}; \qquad \beta_0 = \frac{1}{\theta_1}; \qquad \beta_1 = \frac{\theta_2}{\theta_1}$$

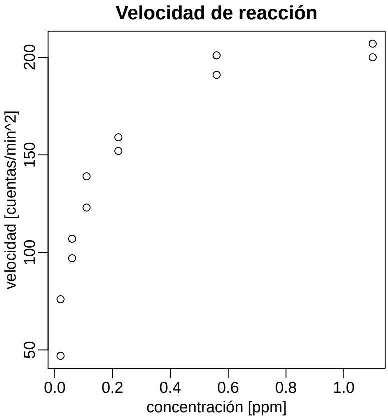
[5] Bates and Watts (1988)

Modelo no lineal de Michaelis-Menten

Velocidad de reacción (cont.)

Datos:

obs.	concentración (x)	velocidad (y)
1	0.02	47
2	0.02	76
3	0.06	97
4	0.06	107
5	0.11	123
6	0.11	139
7	0.22	152
8	0.22	159
9	0.56	191
10	0.56	201
11	1.10	200
12	1.10	207



Modelo no lineal de Michaelis-Menten

Velocidad de reacción (cont.)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Ajuste:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0051072  0.0007040   7.255 2.74e-05
X           0.0002472  0.0000321   7.700 1.64e-05

Residual standard error: 0.001892 on 10 degrees of freedom
Multiple R-squared: 0.8557,     Adjusted R-squared: 0.8413
F-statistic:  59.3 on 1 and 10 DF,  p-value: 1.642e-05

Analysis of Variance Table

Response: Y
          Df     Sum Sq    Mean Sq F value    Pr(>F)
X          1 2.1232e-04 2.1232e-04  59.297 1.642e-05
Residuals 10 3.5806e-05 3.5810e-06
```

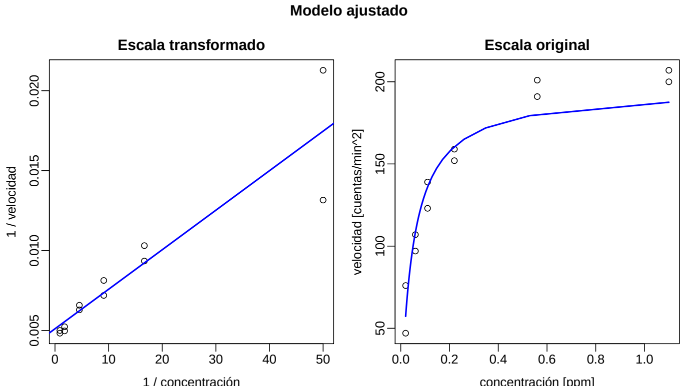
−ebz

Modelo no lineal de Michaelis-Menten

Velocidad de reacción (cont.)

Modelos:

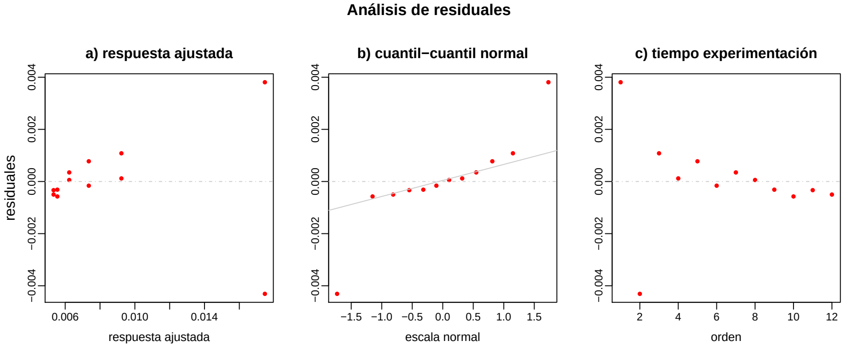
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \qquad\qquad \hat{y} = \frac{\hat{\theta}_1 x}{\hat{\theta}_2 + x}$$



Modelo No lineal de Michaelis-Menten

Velocidad de reacción (cont.)

Análisis de Residuales



−ebz

Modelo no lineal de Michaelis-Menten

Ajuste mínimos cuadrados no lineales

$$y = \frac{\theta_1 x}{\theta_2 + x} + \epsilon$$

Ajuste:

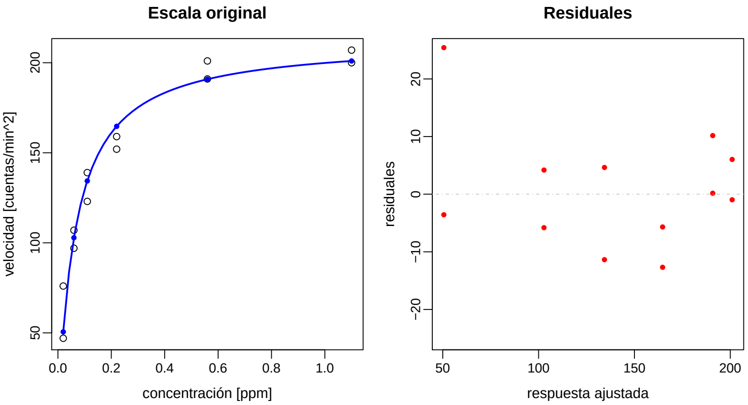
```
puromycin.mod <- nls( rate ~ (theta1 * conc)/(theta2 + conc), data = puromycin.data,
                    start = list(theta1 = 200, theta2 = 0.1), trace=1, model=TRUE)
print(summary(puromycin.mod))

 7964.19 :  200.0          0.1
 1593.16 :  212.02378921   0.05428736
1201.035 :  211.77279725   0.06232446
1195.509 :  212.56325867   0.06392648
1195.449 :  212.67158763   0.06410228
1195.449 :  212.68256678   0.06411945
1195.449 :  212.68362992   0.06412111

Formula: rate ~ (theta1 * conc)/(theta2 + conc)

Parameters:
        Estimate Std. Error t value Pr(>|t|)
theta1 2.127e+02  6.947e+00  30.615 3.24e-11
theta2 6.412e-02  8.281e-03   7.743 1.57e-05

Residual standard error: 10.93 on 10 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 6.085e-06
```

–ebz

Modelo No lineal de Michaelis-Menten

Ajuste mínimos cuadrados no lineales

Modelo ajustado RNL



−ebz