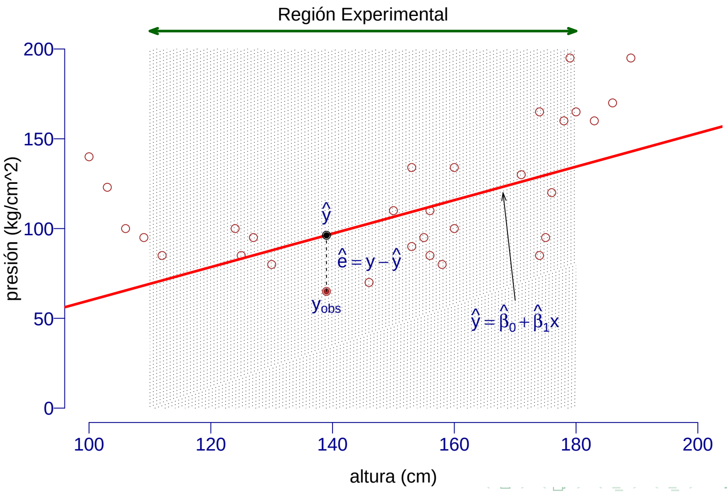
1 - Introducción



Temario

−ebz

Modelos y modelación

- Un *modelo* es una representación aproximada de una situación física.

G. E. P. Box, 1979

"Todos los modelos son erróneos, pero algunos modelos son útiles."

- En ocasiones, cuando los problemas son complejos (extensos), una opción práctica es el uso de *modelos probabilísticos* y *modelos estadísticos* que consideran "patrones regulares" de ruido. Muchas veces estos modelos son empíricos pero útiles en la práctica (*working models*).
- Los *modelos lineales* nos ofrecen una forma de explicar la variable de respuesta en términos de otras variables. Estos modelos sí tratan de explicar y aproximar la realidad a diferencia de los *modelos de series de tiempo* que sólo buscan predecir.

Modelos estadísticos lineales

- *Modelos de regresión lineal múltiple:*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad ((1))$$

- *Modelos polinomiales:*

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

  Este modelo es de la misma forma que el modelo (1) con $x_i = x^i$.

- *Modelos sinusoidales:*

$$y = \beta_0 + \beta_1 \sin\theta + \beta_2 \cos\theta + \epsilon$$

  similar al modelo (1) con $x_1 = \sin\theta$ y $x_2 = \cos\theta$.

- Modelos como

$$y = \beta_0 + \beta_1 \log\xi_1 + \beta_2 \frac{e^{\xi_2}}{\xi_3} + \epsilon$$

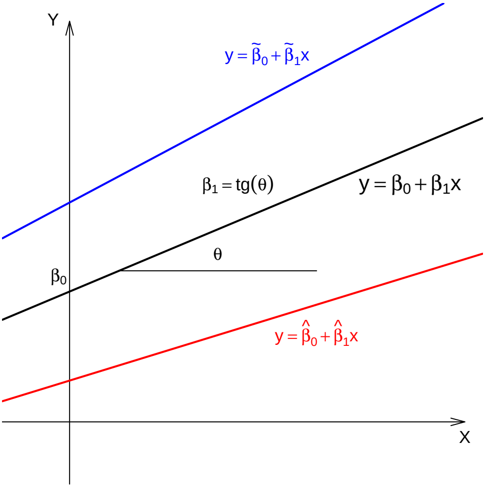
  donde $x_1 = \log\xi_1, \; x_2 = e^{\xi_2}/\xi_3$

- Ejemplo de un modelo *no lineal*:

$$y = \beta_0(1 - e^{-\beta_1 \xi}) + \epsilon$$

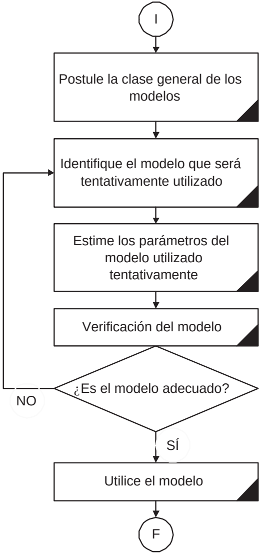
–ebz

La *"mejor"* línea recta



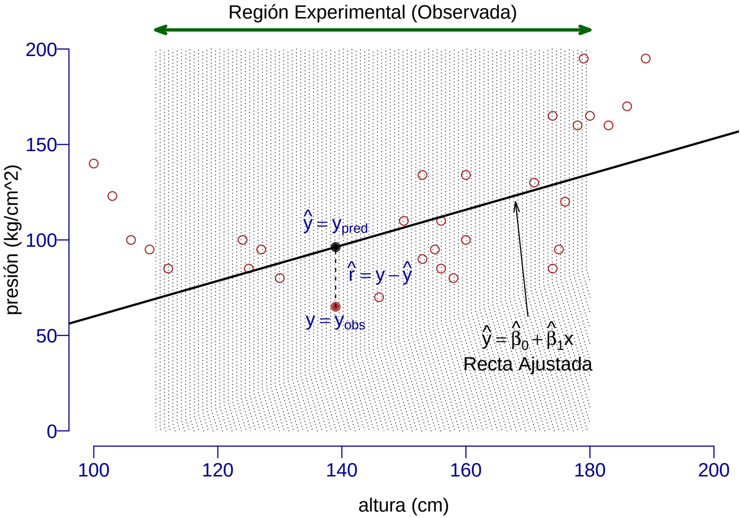
−ebz

Modelación estadística[1]



[1] Box and Jenkins, 1970, p. 19.

Datos observados y modelo ajustado



−ebz

Criterios para determinar la “mejor” línea recta

Modelo

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

1. Criterio L_0: Elija $(\tilde{\beta}_0, \tilde{\beta}_1)$, de modo que

$$\sum_i (y_i - \tilde{y}_i) = \sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0$$

2. Criterio L_1: *Mínima Desviación Absoluta*

$$\min_{\beta} \sum_i |y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i|$$

3. Criterio L_2: *Mínimos Cuadrados*

$$\min_{\beta} \sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

4. Criterio L_∞: *Mínima Desviación Máxima*

$$\min_{\beta} \left\{ \max_i \left| y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right| \right\}$$

−ebz

Modelo y supuestos de la regresión lineal

- *Modelo*

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

  O en notación matricial,

$$y = X\beta + \epsilon$$

- *Supuestos*
  - El modelo es correcto.
  - Los errores son *i.i.d.*: $\epsilon \sim N(0, \sigma^2 I)$
- *Ajuste del modelo* (estimación de parámetros)

  $\hat{\beta} = (X'X)^{-1}X'y \sim N(\beta, \sigma^2(X'X)^{-1})$

  $\hat{\sigma}^2 = \frac{1}{n-p-1} ||y - X\hat{\beta}||^2 \sim \chi^2_{n-p-1}$, independiente de $\hat{\beta}$
- *Validación del modelo* (Análisis de Residuales)

$$\hat{e} = y - \hat{y} = y - X\hat{\beta}$$

  Si el modelo es correcto los residuales se comportan como "errores estimados". Por tanto, ¿cumplen éstos con los supuestos en los que se basa el modelo?

Supuestos sobre la aleatoriedad de los errores y consecuencias

Modelo:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Supuestos:

$$\epsilon \sim \mathrm{N}(0, \sigma^2), \qquad \textit{i.i.d.}$$

Consecuencias:

respuesta observada: $y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$

pendiente: $\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \sigma^2 \frac{1}{S_{xx}}\right)$

ordenada al origen: $\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\sum x_i^2}{n S_{xx}}\right]\right)$

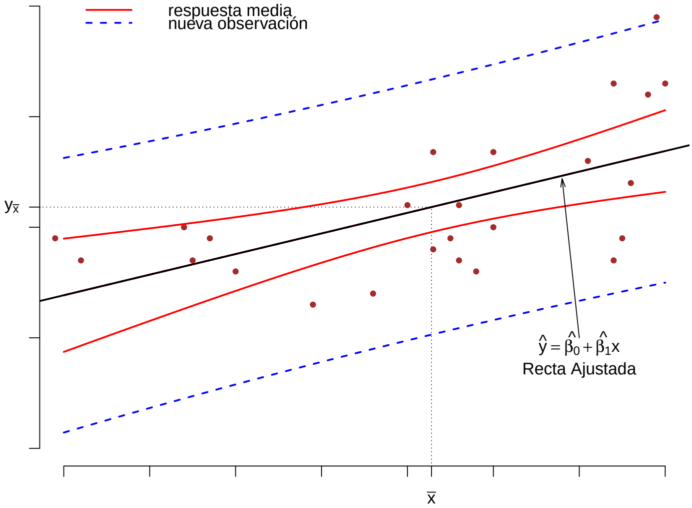
suma desviaciones: $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$

respuesta ajustada: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\right)$

nueva observación: $\dot{y} = \hat{y} + \epsilon \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[1 + \frac{1}{n} + \frac{\sum (x_i - \bar{x})^2}{S_{xx}}\right)\right]$

−ebz

Bandas de confianza y Bbndas de predicción



Análisis de varianza

Análisis de Varianza

Fuente	GL	Suma de Cuadrados	Cuadrados Medios	F
Debido regresión	1	$\sum(\hat{y}_i - \bar{y})^2$	$\frac{\text{SC}_{\text{Reg}}}{\text{gl}}$	$\frac{\text{CM}_{\text{Reg}}}{\text{CM}_{\text{Res}}}$
Residuales	$n-2$	$\sum(y_i - \hat{y}_i)^2$	s^2	
Total (Corregido)	$n-1$	$\sum(y_i - \bar{y})^2$		

Análisis de Varianza y Suma Extra de Cuadrados

Fuente	GL	Suma de Cuadrados	Cuadrados Medios	F
β_0	1	$n\bar{y}^2$		
$\beta_1 \mid \beta_0$	1	$\text{S}_{xy}^2/\text{S}_{xx}$	CM_{Reg}	
Residuales	$n-2$	por diferencia	s^2	
Total	n	$\sum y_i^2$		

Coeficiente de correlación múltiple R^2:

$$R^2 = \frac{\text{SC debido regresión}}{\text{SC corregida}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}, \qquad 0 \leq R^2 \leq 1$$

R^2 Ajustado:

$$R^2_{\text{Adj}} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right)$$

−ebz

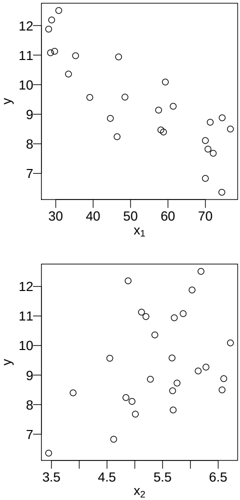
Ejemplo: Datos vapor[2]

obs	x_1	x_2	y
1	35.3	5.20	10.98
2	29.7	5.12	11.13
3	30.8	6.19	12.51
4	58.8	3.89	8.40
5	61.4	6.28	9.27
6	71.3	5.76	8.73
7	74.4	3.45	6.36
8	76.7	6.57	8.50
9	70.7	5.69	7.82
10	57.5	6.14	9.14
11	46.4	4.84	8.24
12	28.9	4.88	12.19
13	28.1	6.03	11.88
14	39.1	4.55	9.57
15	46.8	5.71	10.94
16	48.5	5.67	9.58
17	59.3	6.72	10.09
18	70.0	4.95	8.11
19	70.0	4.62	6.83
20	74.5	6.60	8.88
21	72.1	5.01	7.68
22	58.1	5.68	8.47
23	44.6	5.28	8.86
24	33.4	5.36	10.36
25	28.6	5.87	11.08

Donde,

x_1 : temperatura atmosférica (^oF)
x_2 : tiempo de operación promedio (hrs.)
y : consumo de vapor (lb/mes)

Diagramas de Dispersión



[2] Draper and Smith (1998)

Ejemplo: Datos vapor (cont.)

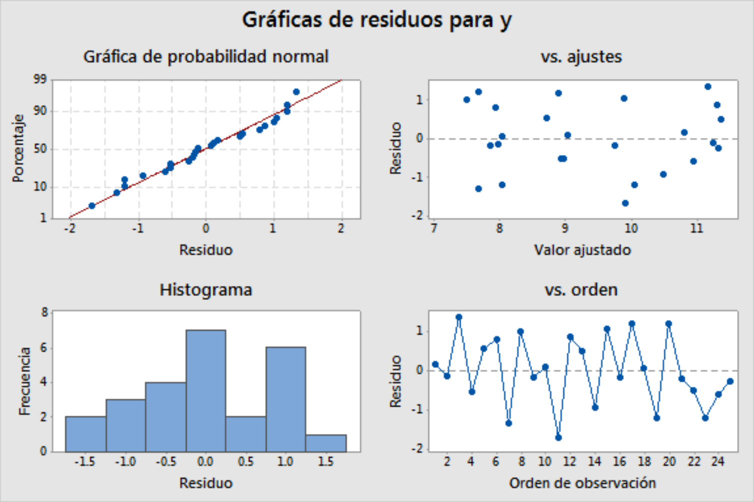
Ajuste del modelo $y = \beta_0 + \beta_1 x_1 + \epsilon$

Minitab: Stat $\rightarrow$ Regression

```
Regression Analysis: y versus x1

The regression equation is
y = 13.6 - 0.0798 x1

Predictor        Coef  SE Coef      T      P
Constant      13.6230   0.5815  23.43  0.000

x1           -0.07983  0.01052  -7.59  0.000

S = 0.890125   R-Sq = 71.4%   R-Sq(adj) = 70.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  45.592  45.592  57.54  0.000
Residual Error  23  18.223   0.792
```

Ejemplo: Datos vapor (cont.)

Análisis gráfico de residuales $\hat{e} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1$



Ejemplo: Datos vapor (cont.)

Ajuste del modelo $y = \beta_0 + \beta_1 x_1 + \epsilon$

R: `mod <- lm(y ~ x1, data=dat)`

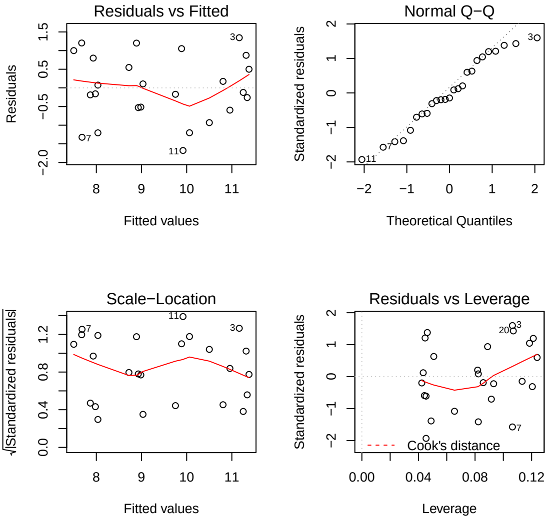
```
Call: lm(formula = y ~ x1, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62299    0.58146  23.429  < 2e-16
x1          -0.07983    0.01052  -7.586 1.05e-07

Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-Squared: 0.7144,     Adjusted R-squared: 0.702
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07
```

```
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 45.592  45.592  57.543 1.055e-07
Residuals 23 18.223   0.792
```

−ebz

Ejemplo: Datos vapor (cont.)

Análisis gráfico de residuales $\hat{e} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1$



Ejemplo: Datos vapor (cont.)

Ajuste del modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

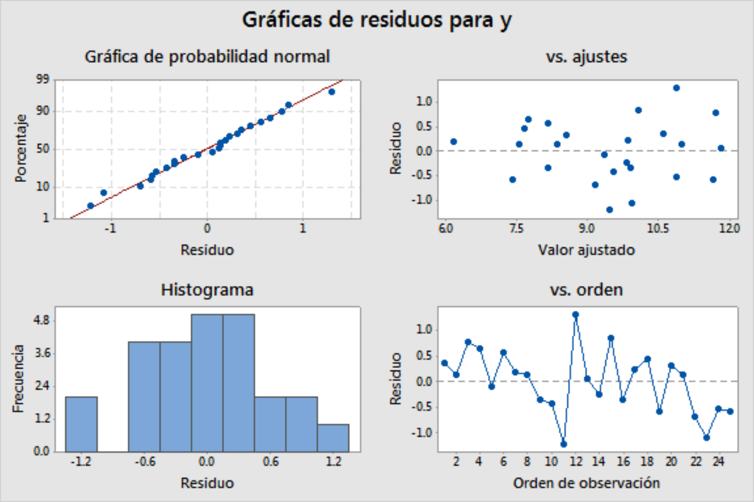
Minitab: Stat $\rightarrow$ Regression

```
Regression Analysis: y versus x1, x2

The regression equation is
y = 9.47 - 0.0798 x1 + 0.762 x2

Predictor        Coef   SE Coef       T      P
Constant       9.4742    0.9619    9.85  0.000
x1          -0.079761  0.007533  -10.59  0.000
x2             0.7616    0.1592    4.78  0.000

S = 0.637158   R-Sq = 86.0%   R-Sq(adj) = 84.7%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  54.884  27.442  67.60  0.000
Residual Error  22   8.931   0.406
Total           24  63.816
```

–ebz

Ejemplo: Datos vapor (cont.)

Análisis gráfico de residuales $\hat{e} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2$



−ebz

Ejemplo: Datos vapor (cont.)

Ajuste del modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

R: `mod <- lm(y ~ x1 + x2, data=dat)`

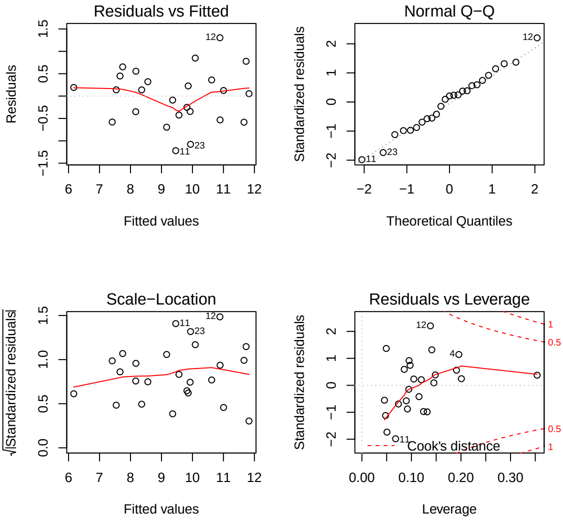
```
Call:
lm(formula = y ~ x1 + x2, data = dat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.474222   0.961894   9.850 1.59e-09
x1          -0.079761   0.007533 -10.588 4.22e-10
x2           0.761648   0.159201   4.784 8.90e-05

Residual standard error: 0.6372 on 22 degrees of freedom
Multiple R-Squared:  0.86,      Adjusted R-squared: 0.8473
F-statistic:  67.6 on 2 and 22 DF,  p-value: 4.035e-10
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 45.592  45.592 112.305 4.157e-10
x2         1  9.292   9.292  22.889 8.896e-05
Residuals 22  8.931   0.406
```

−ebz

Ejemplo: Datos vapor (cont.)

Análisis gráfico de residuales $\hat{e} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2$



Transformaciones a una línea recta

Funciones linealizables y formas lineales

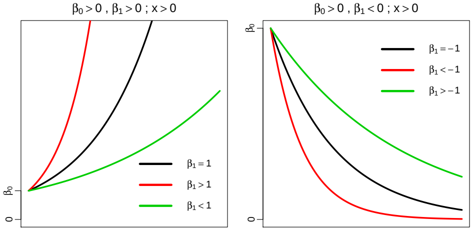
Función Linealizable	Transformación	Forma Lineal
$y = \beta_0 x^{\beta_1}$	$Y = \log y,\ X = \log x$	$Y = \log \beta_0 + \beta_1 X$
$y = \beta_0 e^{\beta_1 x}$	$Y = \log y$	$Y = \log \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$X = \log x$	$y = \beta_0 + \beta_1 X$
$y = \frac{x}{\beta_0 x + \beta_1}$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

−ebz

Transformaciones a una Línea Recta

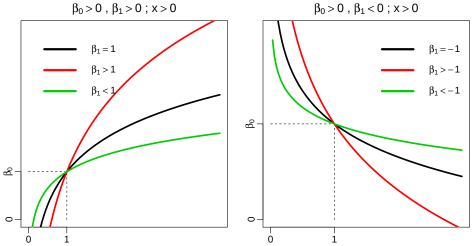
Modelo

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



Modelo

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



Transformación estabilizadora de la varianza Box-Cox

Ajuste el modelo de regresión lineal simple a la respuesta

Transformación potencia Box-Cox

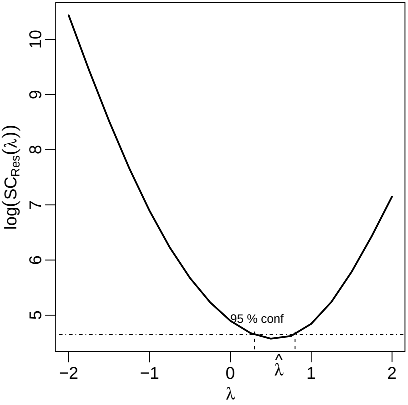
$$Y = \begin{cases} y^{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

Use λ^* que minimice la suma de cuadrados de los residuales $\text{SC}_{\text{Res}}(\lambda)$.

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

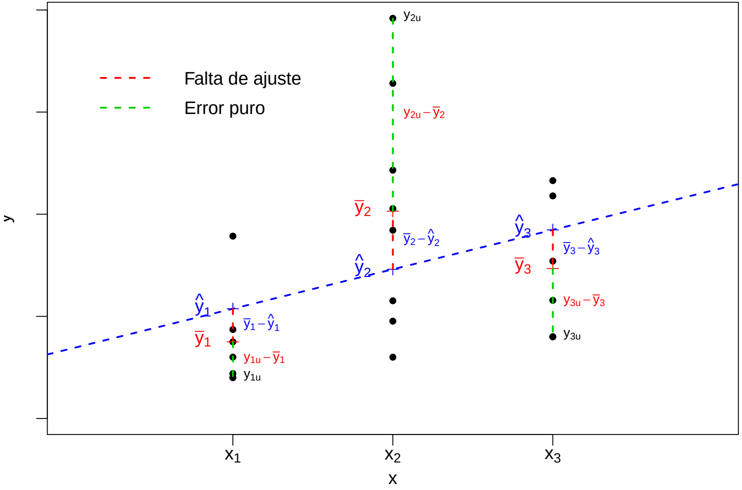
$$\text{SC}^* = \text{SC}_{\text{Res}}(\lambda^*) \left(1 + \frac{t^2_{(1-\alpha/2,\nu)}}{\nu}\right)$$

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



Validación del modelo

Modelo correcto: error puro y falta de ajuste



Modelo Correcto: Error Puro y Falta de Ajuste

En la presencia de réplicas puras la *suma de cuadrados de los residuales* se puede descomponer como

$$
\begin{array}{lcccccc}
 & \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_{iu}-\hat{y}_i)^2 & = & \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_{iu}-\bar{y}_{i\cdot})^2 & + & \sum_{i=1}^{m} n_i(\bar{y}_{i\cdot}-\hat{y}_i)^2 \\
 & \mathrm{SC}_{\text{Residuales}} & = & \mathrm{SC}_{\text{Error Puro}} & + & \mathrm{SC}_{\text{Falta de Ajuste}} \\
\text{g.l.} & (n-2) & = & (n-m) & + & (m-2)
\end{array}
$$

Bajo los supuestos del modelo, $\mathrm{CM}_{\mathrm{EP}} = \mathrm{SC}_{\mathrm{EP}}/(n-m)$ y $\mathrm{CM}_{\mathrm{FA}} = \mathrm{SC}_{\mathrm{FA}}/m-2$ son estimaciones independientes de σ^2 y su cociente sería aproximadamente 1. De hecho, bajo los supuestos del modelo:

$$\hat{F} = \frac{\mathrm{CM}_{\mathrm{FA}}}{\mathrm{CM}_{\mathrm{EP}}} \sim F_{(m-2,n-m)}$$

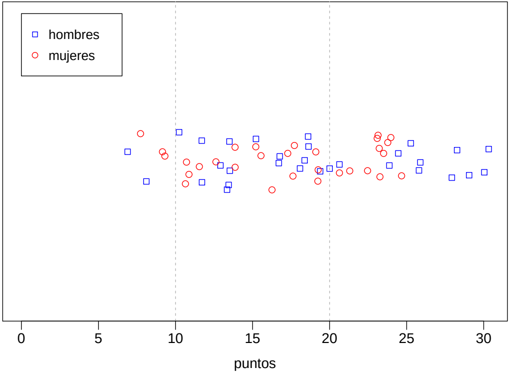
Entonces, si

$$\hat{F} > F_{(1-\alpha;m-2,n-m)} \implies \textit{El modelo no es correcto}$$

Modelos de análisis de covarianza

Ejemplo: Salarios por género y antigüedad

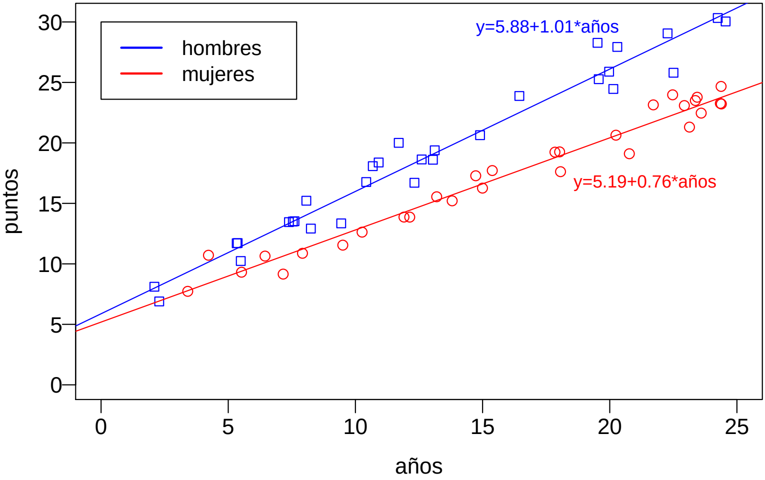
Una empresa tiene un *sistema de puntos*, que dependen básicamente de la antigüedad del empleado, y que están muy correlacionados con el salario. Se tomó una muestra aleatoria de 30 mujeres y 30 hombres y se observaron los puntos acumulados. ¿Hay diferencia de género en la asignación de puntos?



−ebz

Comparación de líneas rectas

Ejemplo: Salarios por género (cont.)

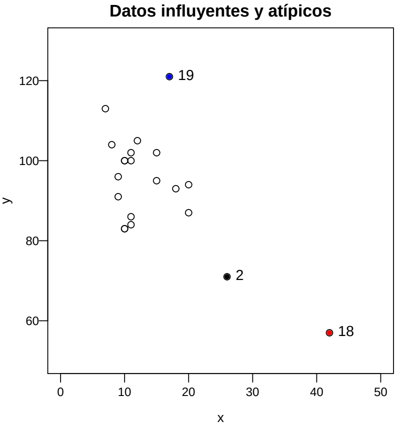


Datos influyentes y atípicos

Ejemplo: Score de aptitud de Gesell

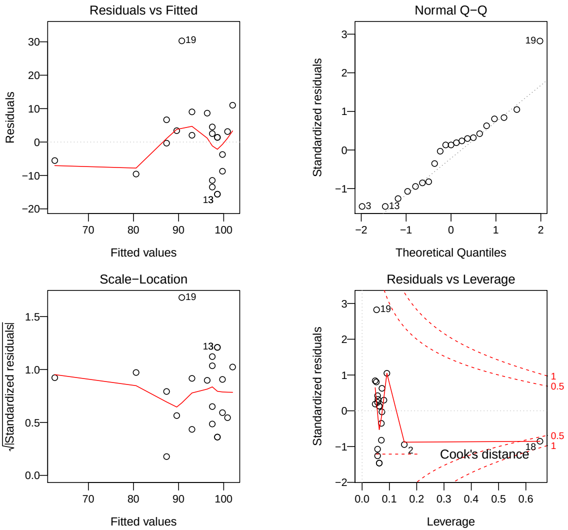
- n: Observación
- x: Edad primera palabra (meses)
- y: Score de aptitud de Gesell

i	x	y	n	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100



Análisis gráfico de residuales

Ejemplo: Score de Aptitud de Gesell (cont.)



Selección del modelo

Ejemplo: Datos sobre cemento de Hald[3]
El siguiente juego de datos es sobre el endurecimiento de cemento Portland, famoso por su nada fácil modelación.

variable	concepto
x_1	Cantidad de tricalcio de aluminiato, 3 $CaO \cdot Al_2O_3$.
x_2	Cantidad de tricalcio de silicato, 3 $CaO \cdot SiO_2$.
x_3	Cantidad de tricalcio de aluminio ferrito, 4 $CaO \cdot Al_2O_3 \cdot Fe_2O_2$.
x_4	Cantidad de dicalcio de silicato, 2 $CaO \cdot SiO_2$.
y	Calor en calorías por gramo de cemento.

Los regresores, x_1, x_2, x_3, x_4 son medidos como porcentaje del peso de las *ollas* donde se hace el cemento.

obs	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

[3] Draper & Smith (1998).

Criterios para la selección de modelos

Ejemplo: Datos sobre cemento de Hald (cont.)

			Estadísticos						Coeficientes				
k	p	q	s	S^2	R^2	$\bar{R}^2$	C_p	AIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
1	0	1	15.04	226.31	·	·	442.92	·	95.42	·	·	·	·
2	1	2	10.73	115.06	0.53	0.49	202.55	102.41	81.48	1.869	·	·	·
3	1	2	9.08	82.39	0.67	0.64	142.49	98.07	57.42	·	0.789	·	·
4	1	2	13.28	176.31	0.29	0.22	315.15	107.96	110.20	·	·	−1.256	·
5	1	2	8.96	80.35	0.67	0.64	138.73	97.74	117.57	·	·	·	−0.738
6	2	3	2.41	5.79	0.98	0.97	2.68	64.31	52.58	1.468	0.662	·	·
7	2	3	11.08	122.71	0.55	0.46	198.09	104.01	72.34	2.312	·	0.494	·
8	2	3	2.73	7.48	0.97	0.97	5.50	67.63	103.10	1.400	·	·	−0.614
9	2	3	6.45	41.54	0.85	0.82	62.44	89.93	72.08	·	0.731	−1.008	·
10	2	3	9.32	86.89	0.68	0.62	138.23	99.52	94.16	·	0.311	·	−0.457
11	2	3	4.19	17.57	0.94	0.92	22.37	78.74	131.28	·	·	−1.200	−0.724
12	3	4	2.31	5.35	0.98	0.98	3.04	63.90	48.19	1.696	0.657	0.250	·
13	3	4	2.31	5.33	0.98	0.98	3.02	63.87	71.65	1.452	0.416	·	−0.237
14	3	4	2.38	5.65	0.98	0.98	3.50	64.62	203.64	·	−0.923	−1.448	−1.557
15	3	4	2.86	8.20	0.97	0.96	7.34	69.47	111.68	1.052	·	−0.410	−0.643
16	4	5	2.45	5.98	0.98	0.97	5.00	65.84	62.41	1.551	0.510	0.102	−0.144

Violación de supuestos: normalidad

Distribución normal de errores

La prueba de *Jarque-Bera* se basa en la *prueba de score*. Compara de manera conjunta el coeficiente de asimetría y curtosis contra los correspondientes parámetros de la distribución normal, resumido en el siguiente estadístico de prueba

$$JB = n\left(\frac{s^2}{6} + \frac{(k-3)^2}{24}\right)$$

donde,

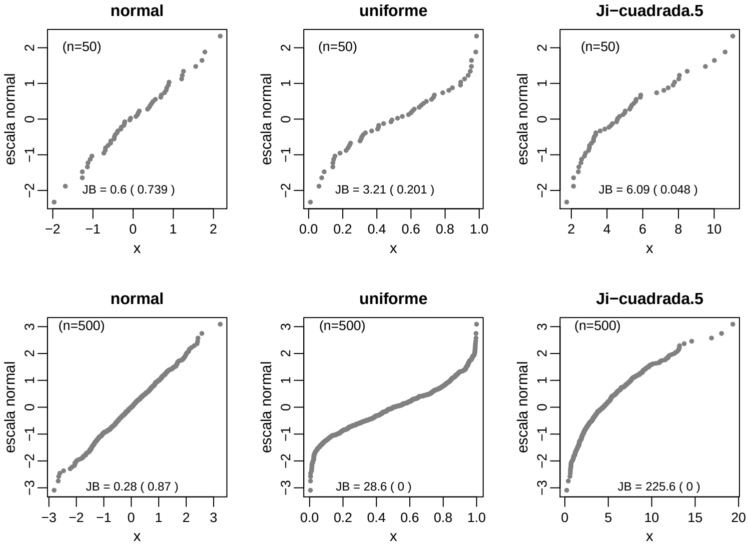
$$s = \frac{m_3}{(m_2)^{3/2}} = \frac{\frac{1}{n}\sum(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]^{3/2}} \quad \text{y} \quad k = \frac{m_4}{(m_2)^{4/2}} = \frac{\frac{1}{n}\sum(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]^{4/2}}$$

con $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$, el r-ésimo momento central muestral.
Los autores muestran que bajo el supuesto de normalidad,

$$JB \xrightarrow{d} \chi^2_2$$

y la prueba es asintóticamente eficiente, que no lo es para muestras pequeñas ($n \leq 100$).

Muestras simuladas y estadístico Jarque-Bera JB (valor-p)



−ebz

Violación de supuestos: homoscedasticidad

Mínimos Cuadrados Generalizados

Suponga el modelo $y = X\beta + \epsilon$ con

$$\mathbb{E}[\epsilon] = 0, \quad \text{y} \quad \text{var}(\epsilon) = \Sigma = \sigma^2 V \neq \sigma^2 I$$

En este caso, utilizar mínimos cuadrados ordinarios (MCO) no es lo apropiado pues las *condiciones de Gauss-Markov* no se cumplen. (Piense en la variación de observaciones o pesos de las observaciones.)
Puesto que $V = \frac{1}{\sigma^2}\text{var}(\epsilon)$, podemos suponer que la matriz V es definida positiva. Entonces, existe $R_{n\times n}$ no singular y simétrica tal que $V = R'R$.

Defina:

$$\begin{aligned} w &= R^{-1}y \\ Z &= R^{-1}X \\ \delta &= R^{-1}\epsilon \end{aligned}$$

Luego,

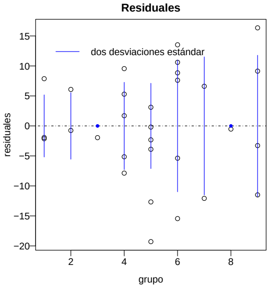
$$\begin{aligned} y &= X\beta + \epsilon \\ R^{-1}y &= R^{-1}X\beta + R^{-1}\epsilon \\ w &= Z\beta + \delta \end{aligned}$$

—ebz

Regresión lineal simple:

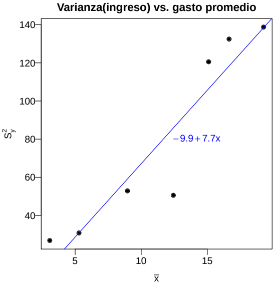
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.4434     4.2889   11.53 3.81e-12
gasto         8.0484     0.3265   24.65  < 2e-16

Residual standard error: 8.999 on 28 degrees of freedom
Multiple R-squared: 0.9559,  Adjusted R-squared: 0.9544
F-statistic: 607.5 on 1 and 28 DF,   p-value: < 2.2e-16
```

- Problema: Varianza creciente.



Regresión lineal S_Y^2 *vs.* $\bar{x}$:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.903     16.764  -0.591  0.58040
xbar           7.703      1.309   5.885  0.00201

Residual standard error: 19.26 on 5 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8739,  Adjusted R-squared: 0.8486
F-statistic: 34.64 on 1 and 5 DF,  p-value: 0.002012
```

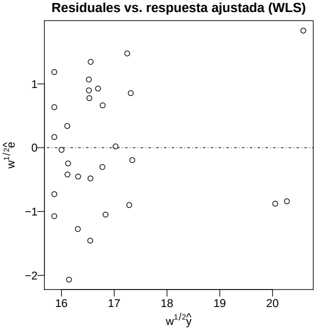


Ejemplo: Venta alimentos (cont.)

```
w <- 1/predict(mod2,data.frame(x=dat$gasto))
lm(formula = ingreso ~ gasto, data = dat, weights = w)
```

Mínimos cuadrados ponderados

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.0475     2.4095   21.19   <2e-16
gasto         7.9162     0.2503   31.62   <2e-16

Residual standard error: 0.9961 on 28 degrees of freedom
Multiple R-squared: 0.9728, Adjusted R-squared: 0.9718
F-statistic:  1000 on 1 and 28 DF,  p-value: < 2.2e-16
```



Violación de supuestos: autocorrelación

Suponga el modelo de primer orden para el error $\epsilon_t = \rho\epsilon_{t-1} + a_t$, donde $a_t \sim \mathrm{N}(0, \sigma_a^2 I)$ es *ruido blanco* y $|\rho| < 1$. Luego, el modelo completo queda:

$$\begin{aligned} y_t &= x_t'\beta + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + a_t \end{aligned}$$

Ahora bien,

$$\epsilon_t = \rho\epsilon_{t-1} + a_t = \rho(\rho\epsilon_{t-2} + a_{t-1}) + a_t = \cdots = \sum_{u=0}^{\infty} \rho^u a_{t-u}$$

De donde,

$$\begin{aligned} \mathbb{E}[\epsilon_t] &= 0 \\ \mathrm{var}(\epsilon_t) &= \mathrm{var}\left(\sum_{u=0}^{\infty} \rho^u a_{t-u}\right) = \sigma_a^2 \frac{1}{1-\rho^2} \\ \mathrm{cov}(\epsilon_t, \epsilon_{t-k}) &= \rho^{|k|}\, \sigma_a^2 \frac{1}{1-\rho^2} \end{aligned}$$

Entonces los ϵ_t tiene media cero pero están correlacionados a menos que $\rho = 0$.

−ebz

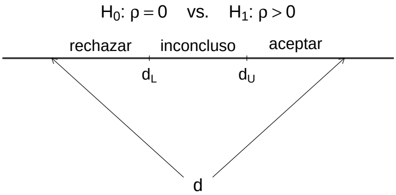
Estadístico *d* de Durbin-Watson

Durbin y Watson mostraron que existen cotas o límites independientes de los datos X tales que, con

$$d = \frac{\sum_{t=2}^{n}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{n}\hat{e}_t^2}$$

si

$$\begin{array}{rcl} d < d_L & \Longrightarrow & \text{Rechazar } \mathrm{H}_0 : \rho = 0 \\ d > d_U & \Longrightarrow & \text{Aceptar } \mathrm{H}_0 : \rho = 0 \\ d_L < d < d_U & \Longrightarrow & \text{Prueba inconclusa} \end{array}$$



Los límites d_L y d_U dependen del número de observaciones n, el número de regresores p y la significancia de la prueba α.

Corrección de Cochran–Orcutt

Considere el modelo de regresión lineal simple

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \quad \text{donde} \quad \epsilon_t = \rho\epsilon_{t-1} + a_t$$

y suponga que rechaza la *hipótesis nula* $\mathrm{H}_0 : \rho = 0$ en favor de la *hipótesis alternativa* $\mathrm{H}_1 : \rho > 0$. Entonces, considere

$$\begin{aligned}
y'_t &= y_t - \rho y_{t-1} \\
&= (\beta_0 + \beta_1 x_t + \epsilon_t) - \rho\,(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) \\
&= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + (\epsilon_t - \rho\epsilon_{t-1}) \\
Y_t &= \alpha_0 + \alpha_1 X_t + a_t
\end{aligned}$$

que sí satisface los supuestos usuales de un modelo de regresión lineal. El problema ahora es que y'_t y x'_t dependen de ρ, que en general es desconocido.

–ebz

Violaión de supuestos: colinealidad

Colinealidad

Potencialmente, las consecuencias de la presencia de colinealidad son muchas. E. g., si se considera el modelo

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

entonces las ecuaciones normales pueden escribirse como

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix}$$

con $r_{12} = \mathrm{corr}(x_1, x_2)$, $r_{yi} = \mathrm{corr}(y, x_i)$. Entonces,

$$C = (X'X)^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

Por lo que

$$\hat{\beta}_i = \frac{r_{yi} - r_{12} r_{yj}}{1 - r_{12}^2}, \quad i = 1, 2 \neq j = 1, 2$$

Por lo que si x_1 y x_2 están correlacionados, $r_{12}^2 \nearrow 1 \Longrightarrow \hat{\beta}_i \nearrow \infty$.

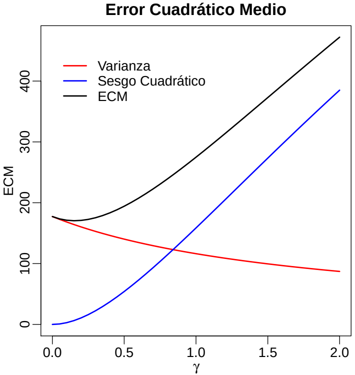
–ebz

Si se cumplen las condiciones de *Gauss-Markov*, el estimador $\hat{\beta}$ es el de varianza mínima entre la clase de estimadores insesgados. Pero si se amplia la clase (dejando entrar estimadores no insesgados) se pueden obtener estimadores de menor varianza que el de *mínimos cuadrados ordinarios (MCO)*.

Existe un compromiso entre insesgamiento y varianza:

$$\text{ECM}(\hat{\beta}) = \text{var}(\hat{\beta}) + \text{Sesgo}^2(\hat{\beta})$$

Hay una vecindad donde el estimador *ridge* es más eficiente que el de MCO.



Las ecuaciones normales se modifican de manera que para $\gamma \geq 0$

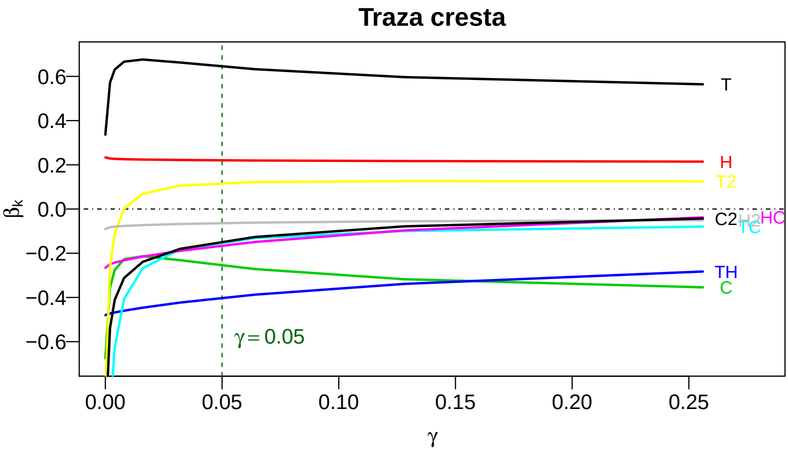
$$\begin{aligned}\hat{\beta}(\gamma) &= (X'X + \gamma I)^{-1}X'y \\ &= (X'X + \gamma I)^{-1}(X'X)\hat{\beta} \\ &= Z(\gamma)\hat{\beta}\end{aligned}$$

Esto es, el estimador cordillera es una transformación lineal del estimador de mínimos cuadrados ordinarios (MCO) $\hat{\beta}$. Se tiene además que para $\hat{\beta}_\gamma = \hat{\beta}(\gamma)$,

$$\begin{aligned}\mathrm{var}(\hat{\beta}_\gamma) &= \sigma^2(X'X + \gamma I)^{-1}X'X(X'X + \gamma I)^{-1} \\ \mathrm{ECM}(\hat{\beta}_\gamma) &= \sigma^2\sum_{j=1}^{q}\frac{\lambda_j}{(\lambda_j + \gamma)^2} + \gamma^2\beta'(X'X + \gamma I)^{-2}\beta\end{aligned}$$

Note que conforme crece γ el estimador $\hat{\beta}_\gamma$ se hace más estable pero también más sesgado.

¿Qué γ usar?


Traza cresta
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
β_k
T
H
T2
C2
H2
HC
TC
TH
C
$\gamma = 0.05$
0.00
0.05
0.10
0.15
0.20
0.25
γ

−ebz

Modelos lineales generalizados

Familia Exponencial de Distribuciones

La variable aleatoria Y se dice que es miembro de la *familia exponencial de distribuciones* si su función de densidad de probabilidad, $f(y;\theta)$, puede expresarse como

$$f(y;\theta) = \exp\left\{a(y)b(\theta) + c(\theta) + d(y)\right\} \qquad (1)$$

si $a(y) = y$, la distribución anterior (1) se dice estar en su *forma canónica* y a $b(\theta)$ se le llama el *parámetro natural* de la distribución.

Ejemplos:

$$\text{Binomial}: \quad f(y;n,p) = \begin{cases} \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ \exp\left\{y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right\} \end{cases}$$

$$\text{Poisson}: \quad f(y;\lambda) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!} \\ \exp\left\{y\log\lambda - \lambda - \log\lambda\right\} \end{cases}$$

$$\text{Normal}: \quad f(y;\mu,\sigma^2) = \begin{cases} \frac{1}{2\pi\sigma^2}\, e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma^2}\right)^2} \\ \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \end{cases}$$

También incluye otras distribuciones como la *gamma*, la *lognormal*, la *gaussiana inversa*, etc.

−ebz

Componentes de un modelo lineal generalizado

1. Se supone que las *variables respuesta*, $y_1, \ldots, y_n$, siguen una distribución común miembro de la *familia exponencial*.
2. Un conjunto de *variables explicativas*, $x_1, \ldots, x_p$, y de parámetros $\beta_0, \beta_1, \ldots, \beta_p$. Así,

$$y_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \quad \beta_{q\times 1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}; \quad X_{n\times q} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

3. Una *función liga* monótona g tal que

$$g(\mu_i) = x_i'\beta$$

donde $\mu_i = \mathbb{E}[y_i]$.

Estimación por máxima verosimilitud

Dada la muestra $y_1, \ldots, y_n$ de $y \sim f(y; \boldsymbol{\theta})$, con $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_m)$, el estimador $\hat{\boldsymbol{\theta}}$ obtenido por el *método de máxima verosimilitud* es aquel tal que

$$L(\hat{\boldsymbol{\theta}}; y) \geq L(\boldsymbol{\theta}; y), \qquad \text{para todo } \boldsymbol{\theta} \in \Theta$$

Equivalentemente, puesto *log* es una función monótona creciente

$$\ell(\hat{\boldsymbol{\theta}}; y) \geq \ell(\boldsymbol{\theta}; y), \qquad \text{para todo } \boldsymbol{\theta} \in \Theta$$

Generalmente el EMV $\hat{\boldsymbol{\theta}}$ se obtiene por diferenciación de la función log-de-verosimilitud $\ell(\theta; y)$ y resolviendo el sistema

$$\frac{\partial^2 \ell(\boldsymbol{\theta}; y)}{\partial \theta_j} = 0, \qquad \text{para todo } j = 1, \ldots, m$$

Es necesario confirmar que la solución $\hat{\boldsymbol{\theta}}$ corresponde a un máximo verificando que la matriz de segundas derivadas

$$\left. \frac{\partial^2 \ell(\boldsymbol{\theta}; y)}{\partial \theta_j \partial \theta_k} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$$

es definida negativa.

Propiedades de los estimadores de máxima verosimilitud: $\left\{ \begin{array}{l} \text{invarianza} \\ \text{consistencia} \\ \text{suficiencia} \\ \text{eficiencia asintótica} \end{array} \right.$

Resultados asintóticos de EMV $\hat{\theta}$

La idea básica es que el EMV $\hat{\theta}$ es un estimador consistente del parámetro θ y que si var$(\hat{\theta})$ es su varianza, entonces:

1. El estimador $\hat{\theta}$ es *asintóticamente insesgado*.

$$\mathbb{E}[\hat{\theta}_n] \to \theta$$

2. El estadístico $\hat{\theta}$ tiene distribución asintótica normal.

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathrm{var}(\hat{\theta})}} \dot{\sim} \mathrm{N}(0, 1) \quad \Longrightarrow \quad \frac{(\hat{\theta} - \theta)^2}{\mathrm{var}(\hat{\theta})} \dot{\sim} \chi^2_1$$

3. En el caso multivariado el estadístico $\hat{\boldsymbol{\theta}}$ tiene distribución asintótica normal.

$$\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}, V) \quad \Longrightarrow \quad (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})V^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2_\nu$$

Devianza *D*

Nelder y Wedderburn (1972) definieron el estadístico *log del cociente de la verosimilitud* como la *devianza D* (escalada)

$$D = 2\log\Lambda = 2\left[\ell(\hat{\beta}_{\text{máx}}\,; y) - \ell(\hat{\beta}; y)\right]$$

D puede descomponerse como

$$D = \left\{ \underbrace{\left[\ell(\hat{\beta}_{\text{máx}}\,; y) - \ell(\beta_{\text{máx}}\,; y)\right]}_{\chi^2_n} + \underbrace{\left[\ell(\hat{\beta}; y) - \ell(\beta; y)\right]}_{\chi^2_p} + \underbrace{\left[\ell(\beta_{\text{máx}}\,; y) - \ell(\beta; y)\right]}_{\geq 0} \right\}$$

En grandes rasgos, si los primeros dos sumandos son independientes y el tercero es cercano a cero, entonces

$$D \;\dot{\sim}\; \chi^2_{n-p}$$

si el modelo es adecuado. Si por el contrario el modelo no es bueno el tercer término será grande y D será mucho mayor que lo esperado por una distribución χ^2_{n-p}.
En la práctica uno tiende a comparar el D calculado de los datos con $(n-p)$, el valor medio de la distribución.

Nota: en general la descomposición anterior no es una aproximación para la distribución muestral del estadístico aunque para el caso normal el resultado es exacto.

Regresión Logística

Respuesta binaria

Considere una variable que puede tomar dos valores solamente: *éxito* ó *fracaso*; *sí* ó *no*; 1 ó 0. Así,

$$Z = \begin{cases} 1 & \text{éxito, sí} \\ 0 & \text{fracaso, no} \end{cases}$$

y tal que

$$P(Z=1) = \pi, \qquad P(Z=0) = 1-\pi, \qquad f(z) = \pi^{z}(1-\pi)^{1-z}$$

Z se dice que sigue una *distribución Bernoulli parámetro* π. Sea $Z_1, \ldots, Z_r$, independientes con $Z_j \sim \mathrm{Bernoulli}(\pi_j)$, entonces

$$f(\boldsymbol{z};\pi) = \prod_{j=1}^{r} f(z_j;\pi) = \prod \pi_j^{z_j}(1-\pi_j)^{1-z_j}$$

que se puede reescribir como

$$\begin{aligned} f(\boldsymbol{z};\pi) &= \exp\left\{\sum_{j=1}^{r} z_j \log(\pi_j) + \sum_{j=1}^{r}(1-z_j)\log(1-\pi_j)\right\} \\ &= \exp\left\{\sum z_j \log\frac{\pi_j}{1-\pi_j} + \sum \log(1-\pi_j)\right\} \\ &= \exp\{a(z)b(\theta) + c(\theta) + d(z)\} \end{aligned}$$

Esto es, *la distribución Bernoulli es miembro de la familia exponencial.*

Los primeros modelos tipo regresión lineal usados para ajustar datos binomiales fue en bioensayos. Respuestas como proporción de animales que sobreviven determinada dosis de toxinas. Tales respuestas son llamadas *respuestas cuantiles*.

Modelo probit: Si la distribución de tolerancia es la normal,

$$\pi = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt$$

donde Φ es la *f. p. a.* de la normal estándar.Luego,

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$$

donde $\beta_0 = -\mu/\sigma$ y $\beta_1 = 1/\sigma$, y la función liga es la inversa de la *f. p. a.* Φ^{-1}.
Los *modelos probit* se usan en áreas de las ciencias biológicas y sociales donde se dan interpretaciones naturales del modelo. Por ejemplo, $x = \mu$ es llamada al *dosis letal mediana (LD(50))*.

Modelo logístico o *logit*

Modelo que permite resultados similares al modelo *probit* pero computacionalmente más sencillo.
Para este caso la distribución de tolerancia es

$$f(t) = \frac{\beta_0 \exp(\beta_0 + \beta_1 t)}{[1 + \exp(\beta_0 + \beta_1 t)]^2}$$

Lo que implica que las probabilidades π quedan determinadas por

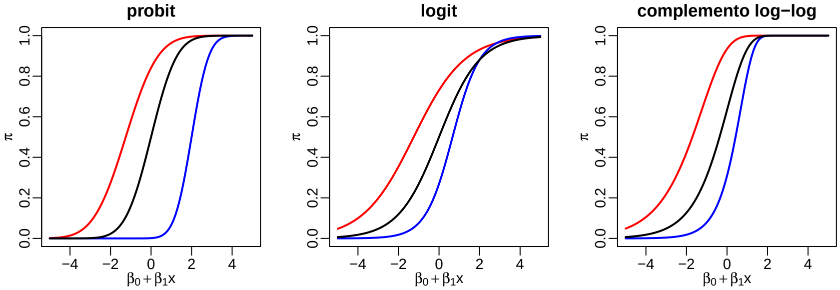
$$\pi = \int_{-\infty}^{x} f(t)dt = \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} = g^{-1}(x'\beta)$$

O bien,

$$\eta = g(\mu) = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$$

La función liga $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ se conoce como *función logística*.
El comportamiento de $f(t)$ y $\pi(x)$ es muy parecido al de la función *probit* excepto en las colas de las distribuciones.

Funciones liga para respuestas binarias



Funciones liga

Probit: $\pi = \Phi^{-1}(x'\beta)$

Logit: $\pi = \dfrac{e^{x'\beta}}{1 + e^{x'\beta}}$

Complemento log-log: $\pi = 1 - \exp\{-\exp(x'\beta)\}$

Coeficientes:

$\beta_0 = 0;\quad \beta_1 = 1$

$\beta_0 = 1;\quad \beta_1 = 4/5$

$\beta_0 = -3;\quad \beta_1 = 3/2$

−ebz

El modelo logístico sería, para $i = 1, \dots, n$

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

o bien,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

En este caso, el estadístico *log-razón-de-verosimilitud* es

$$D = \sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n - y_i}{n - \hat{y}_i}\right)\right]$$

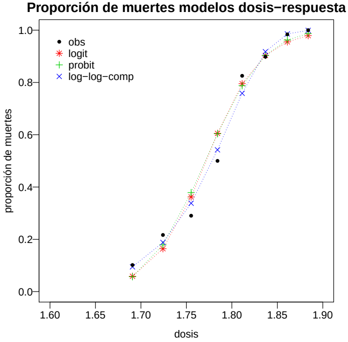
Modelo *logit:*

```
Call:
glm(formula = y ~ x, family = binomial("logit"), weights = dat$n)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -60.717      5.181  -11.72   <2e-16
 x             34.270      2.912   11.77   <2e-16

     Null deviance: 284.202  on 7  degrees of freedom
 Residual deviance:  11.232  on 6  degrees of freedom (*)

 AIC: 41.43

 Number of Fisher Scoring iterations: 4
```

(*) valor-p=0.0815.

Problema de mortalidad de insectos

Proporción de muertes en modelos dosis-respuesta



dosis x_i	número de insectos n_i	número de muertes y_i	predicciones de modelos *logístico*	*probit*	*valor extremo*
1.6907	59	6	3.46	3.36	5.59
1.7242	60	13	9.84	10.72	11.28
1.7552	62	18	22.45	23.48	20.95
1.7842	56	28	33.90	33.82	30.37
1.8113	63	52	50.10	49.62	47.78
1.8369	59	53	53.29	53.32	54.14
1.8610	62	61	59.22	59.66	61.11
1.8839	60	60	58.74	59.23	59.95
Devianza *D*			11.23	10.12	3.45

−ebz