Bioestadística con R

Roberto Bustillos, MVZ, M.Sc - Basado en un material de C. Armero

Universidad Central del Ecuador

Facultad de Medicina Veterinaria y Zootecnia

Maestría en Epidemiología y Salud Pública Veterinaria, 2016-2018

Contenido

Modelo de regresión.

Historia

- Los modelos de regresión fueron usados inicialmente en Astronomía y Física por Laplace y Gauss, sin embargo su nombre genérico proviene de trabajos de Galton en Biología a finales del siglo XIX.
- Galton, estudió la dependencia de la estatura de los hijos (Y) respecto a la de sus padres (X), encontrando lo que denominó una "regresión" a la media.
- Por tanto, los modelos estadísticos que explican la dependencia de una variable Y respecto de una o varias variables cuantitativas X se denominan modelos de regresión.

Examples

La edad es uno de los factores determinantes en la pérdida de masa muscular. Con objeto de explotar dicha relación en mujeres, un nutricionista selecciona aleatoriamente 15 mujeres en cada uno de los grupos de edad: [40,49], [50,59], [60,69] y [70,79], y calcula, a través de diferentes medidas, un indicador de su masa muscular. Una parte de los resultados obtenidos se muestra en la siguiente tabla:

Mujer	1	2	3	...	58	59	60
Edad	43	41	47	...	76	72	76
Mmuscular	106	106	97	...	56	70	74

Examples

mmuscular	edad
Min. : 52.00	Min. : 41.00
1st Qu.: 73.00	1st Qu.: 50.25
Median : 84.00	Median : 60.00
Mean : 84.97	Mean : 59.98
3rd Qu.: 97.00	3rd Qu.: 70.00
Max. : 119.00	Max. :78.00

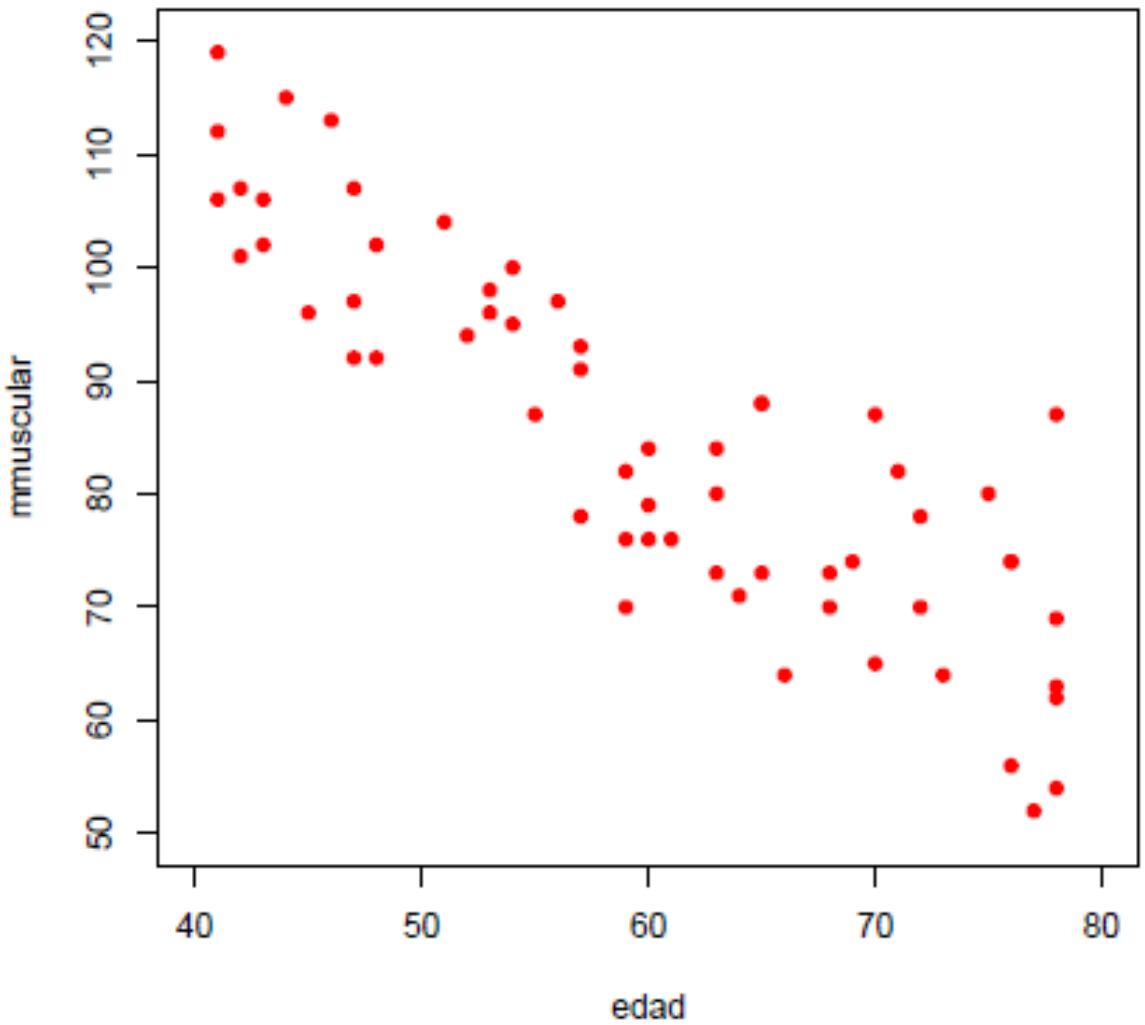
$s^2_{\text{mmuscular}} = 262.7446$

$s_{\text{mmuscular}} = 16.2094$

$s^2_{\text{edad}} = 139.1692$

$s_{\text{edad}} = 11.797$

Examples



Examples

Se pretende analizar el porcentaje de ácido ascórbico (ascórbico) retenido por los manojos de espinacas frescas después de un tratamiento de secado a 900 °C en relación a su porcentaje de materia seca (mateseca). Los 24 manojos de espinacas tratados pueden considerarse una muestra aleatoria de la correspondiente población de espinacas.

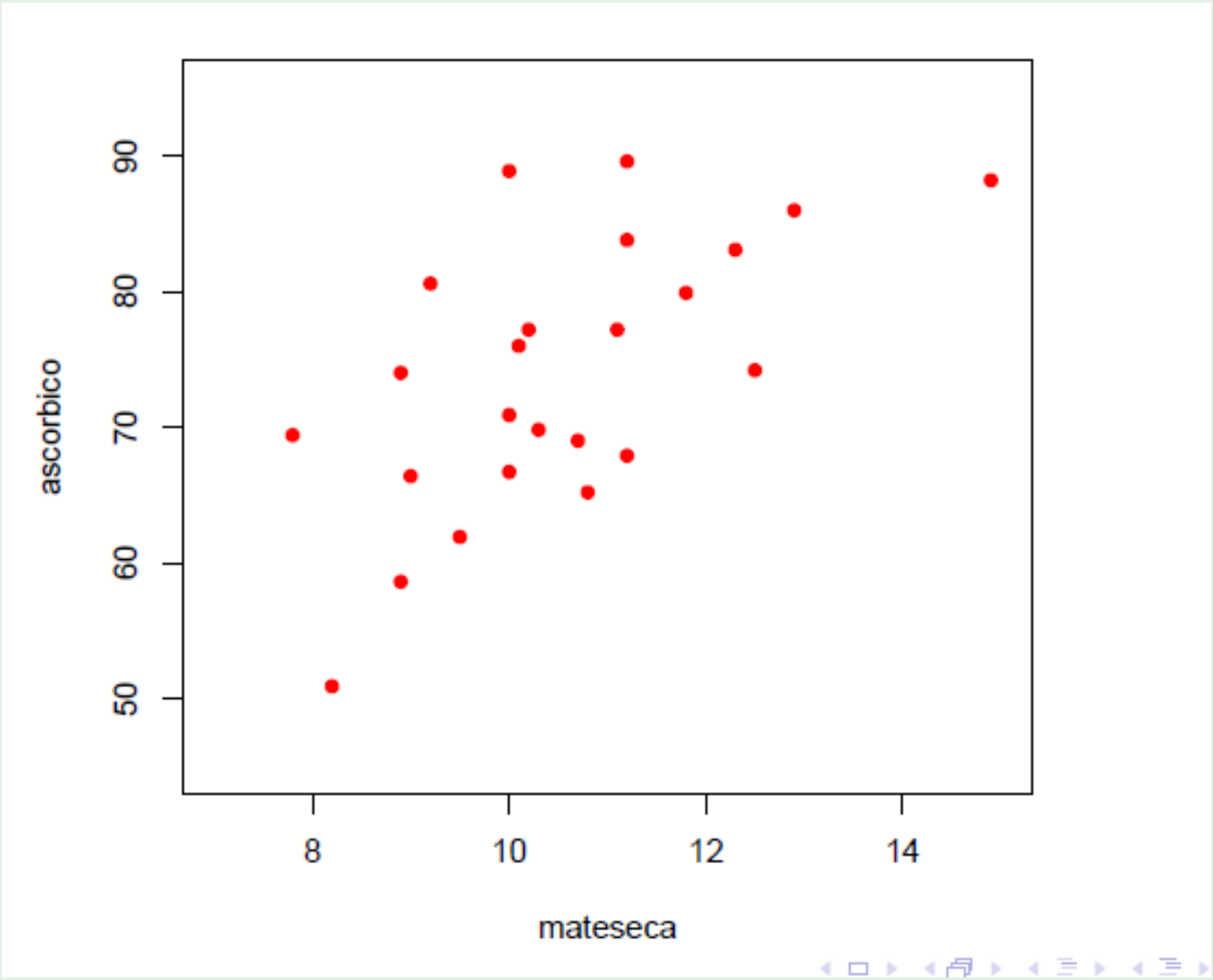
manojo	mateseca	ascórbico	manojo	mateseca	ascórbico
01	10.00	70.90	13	08.90	74.00
02	08.90	58.60	14	09.20	80.60
03	07.80	69.40	15	10.10	76.00
04	09.00	66.40	16	08.20	50.90
05	09.50	61.90	17	10.80	65.20
06	11.10	77.20	18	11.20	89.60
07	12.50	74.20	19	12.30	83.10
08	10.00	66.70	20	10.20	77.20
09	11.20	83.80	21	11.20	67.90
10	10.00	88.90	22	10.70	69.00
11	10.30	69.80	23	12.90	86.00
12	11.80	79.90	24	14.90	88.20

Examples

mateseca	ascorbico
Min. : 7.800	Min. : 50.90
1st Qu.: 9.425	1st Qu.:67.60
Median : 10.250	Median : 74.10
Mean : 10.529	Mean :73.97
3rd Qu.: 11.200	3rd Qu.: 81.22
Max. : 14.900	Max. : 89.60

$$s^2_{\text{mateseca}} = 2.5943$$
$$s_{\text{mateseca}} = 1.6107$$
$$s^2_{\text{ascorbico}} = 100.3924$$
$$s_{\text{ascorbico}} = 10.0196$$

Examples



Conceptos.

- El objetivo es estudiar el comportamiento de una variable cuantitativa (masa muscular en mujeres, porcentaje de ácido ascórbico en manojos de espinacas frescas) en función de otra variable (edad, porcentaje de materia seca), también cuantitativa.
- Variable de interés se le llama variable respuesta o dependiente (Y).
- Variable que no es de interés, pero que se utiliza para estudiar la Y se le llama variable predictora, explicativa o independiente (X) y suele ser cuantitativa.
- Existen muchas relaciones posibles entre dos variables:
  - $Y = f(X) \Longrightarrow$ relación exacta.
  - $Y = f(X) +$ error $\Longrightarrow$ relación no exacta, estadística cuando el error se modeliza en términos probabilísticos.
  - $Y = \beta_0 + \beta_1 X +$ error $\Longrightarrow$ la más sencilla de las relaciones estadísticas (modelo de regresión lineal simple).

El modelo de regresión lineal simple.

El modelo de regresión lineal simple para una muestra.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, ..., n$$

siendo:

- Y_i la variable respuesta correspondiente al elemento i de la muestra (Y_i es una variable aleatoria).
- X_i el valor de la variable predictora correspondiente al elemento i de la muestra (a X_i no se le considera variable aleatoria).
- β_0 y β_1 los coeficientes de la recta de regresión (parámetros desconocidos) y $\beta_0 + \beta_1 X$ la recta o función de regresión.
- $\epsilon_i = 1,...,n$, los errores aleatorios que son variables aleatorias i.i.d. según $N(0,\sigma^2)$, siendo σ^2 la varianza del modelo (parámetro desconocido).

Estimación de parámetros.

- Existen dos métodos: el más popular en estadística es el método de máxima verosimilitud, pero en el contexto de regresión el más conocido y utilizado es el método de los mínimos cuadrados.
- Como el modelo de regresión lineal simple,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

depende de tres parámetros, nuestro objetivo será estimarlos.

Estimación: método de mínimos cuadrados.

- A partir de la información proporcionada por n observaciones $\{(X_i, Y_i), i = 1, ..., n\}$ el método de los mínimos cuadrados considera la diferencia entre cada observación Y_i de la variable respuesta y su correspondiente media, $\beta_0 + \beta_1 X_i$, a través del estadístico:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

y elige como estimadores de β_0 y β_1 aquellos valores que minimizan el valor de $Q(\beta_0, \beta_1)$.

Examples

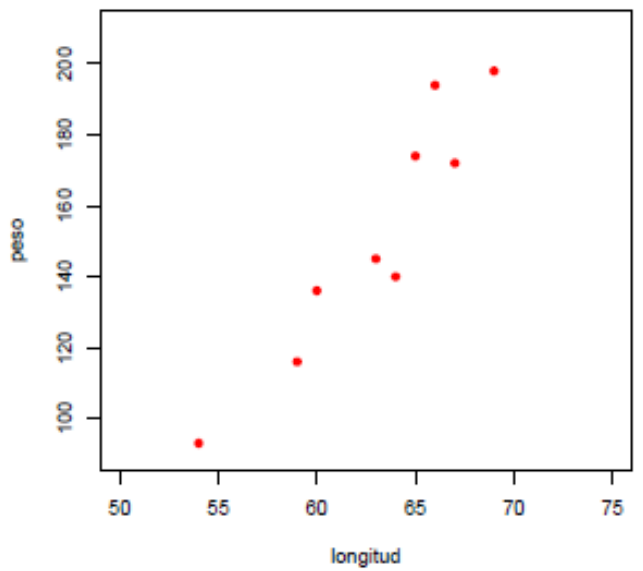
En un estudio sobre una población salvaje de la serpiente *Vipera bertis* se capturaron y midieron nueve ejemplares de hembra adultos. En la tabla se muestra su peso, en gramos, y longitud, en cm.

serpiente	longitud, X	peso, Y
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145

Examples

Descripción básica de los datos:

	n	mínimo	media	máximo	varianza	desv. típica
longitud	9	54	63	69	21.500	4.637
peso	9	93	152	198	1248.750	35.338



- Modelo:

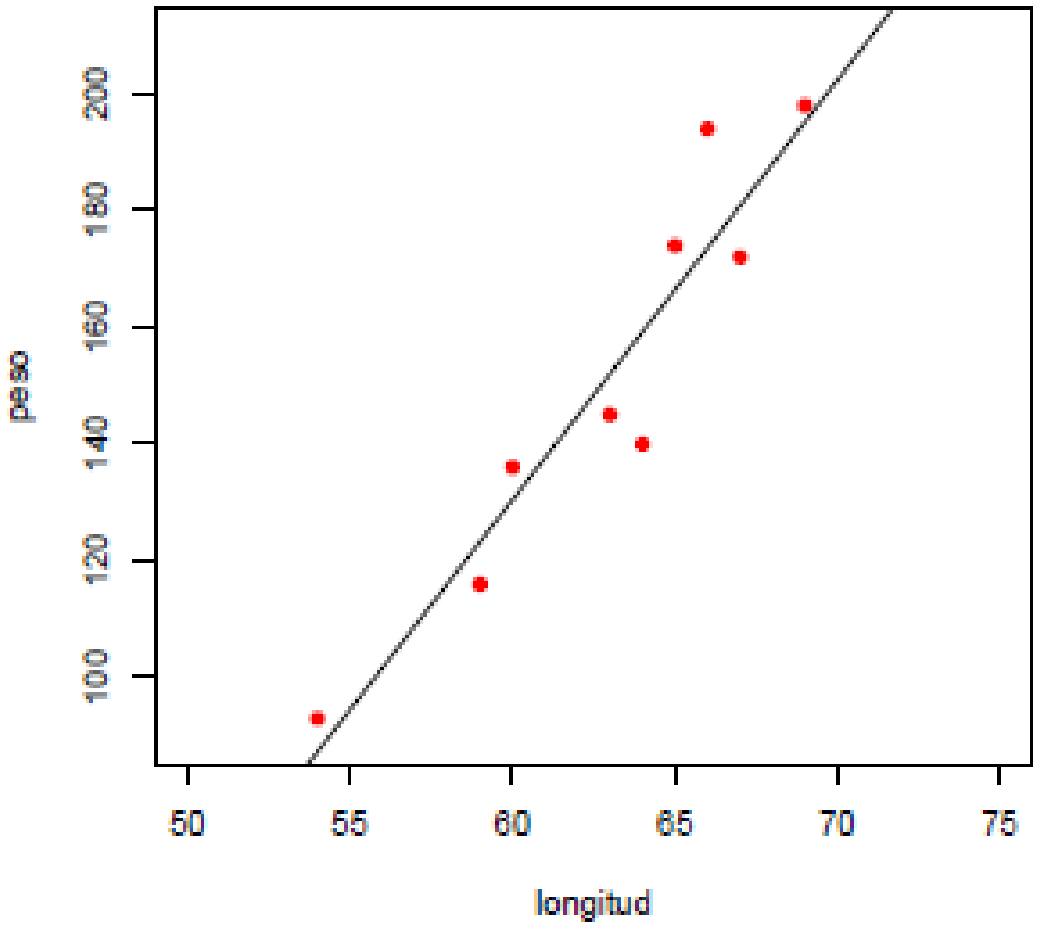
$$\underbrace{Peso}_{gramos} = \underbrace{\beta_0}_{gramos} + \underbrace{\beta_1}_{\frac{gramos}{cm}} \underbrace{Longitud}_{cm} + \epsilon,$$

con $\epsilon \sim N(0, \sigma^2)$ y siendo β_0, β_1 y σ^2 los parámetros desconocidos del modelo.
El modelo también puede expresarse como:

$$(Peso|Longitud) \sim N(\beta_0 + \beta_1 Longitud, \sigma^2)$$

- Parámetros estimados:
$\hat{\beta}_1 =$ 7.192 gramos por cm: estimación de β_1
$\hat{\beta}_0 =$ -301.087 gramos: estimación de β_0

- Recta de regresión ajustada:
  $\hat{Peso} = -301.087 + 7.192 Longitud$



Propiedades de la recta de regresión ajustada.

- La suma de todos los residuos es cero, $\sum_{i=1}^{n} e_i = 0$.
- El residuo i-ésimo es la diferencia entre el valor observado de la variable respuesta Y_i y el ajustado $\hat{Y}_i$. Lo representaremos por e_i:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- La suma de todas las observaciones de la variable respuesta es igual a la suma de los valores ajustados, $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$, y por lo tanto, $\bar{Y} = \hat{Y}$.
- La recta de regresión ajustada siempre pasa por el punto $(\bar{X}, \bar{Y})$.

- Recta de regresión ajustada:

$$\hat{Peso} = -301.087 + 7.102 Longitud$$

- Valores observados, ajustados y residuos.

	serp	longitud, X	peso, Y	ajustados, $\hat{Y}$	residuos, $Y - \hat{Y}$
	1	60	136	130.424	5.576
	2	69	198	195.151	2.849
	3	66	194	173.576	20.424
	4	64	140	159.192	-19.192
	5	54	93	87.273	5.727
	6	67	172	180.767	-8.767
	7	59	116	123.233	-7.233
	8	65	174	166.384	7.616
	9	63	145	152.000	-7.000
media		63	152	152	0
varianza		21.500	1248.750	1112.041	136.709

Estimación de la varianza del modelo, σ^2

- Una estimación de la varianza del modelo de regresión lineal simple es

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- Al término $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$ se le denomina suma de cuadrados residual o también suma de cuadrados debido al error y se la representa habitualmente como SSE.
- s^2 es un estimador insesgado de σ^2

- Recta de regresión ajustada:

$$\hat{Peso} = -301.087 + 7.192 Longitud$$

- Estimación de la varianza del modelo: $s^2 = 156.238 \quad gramos^2$
- Introducimos notación:
  Estimación de la varianza del modelo $\Longrightarrow s^2 = 156.238$
  Varianza muestral de los errores $\Longrightarrow s_e^2 = 136.709$
  Varianza muestral de los valores observados de Y $\Longrightarrow$
  $s_y^2 = 1248.750$
  Varianza muestral de los valores ajustados, $\hat{Y} \Longrightarrow$
  $s_{\hat{y}}\hat{y}^2 = 1112.041$
- $1248.750 = 136.709 + 1112.041$
- $156.238 = (7/8)136.709$, $\frac{n-2}{n-1}s_e^2$

Estimación: método de máxima verosimilitud.

- El método de máxima verosimilitud (MLE) selecciona como estimadores de los parámetros del modelo los que son más compatibles con los datos. Esto se expresa a través de la función de verosimilitud.
- Estimadores:

Parámetro	Método de mínimos cuadrados	Método de máxima verosimilitud
β_0	$\hat{\beta}_0$	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$	$\hat{\beta}_1$
σ^2	s^2	$s^2\,(n-2)/n$

- Es similar a la estimación de mínimos cuadrados, pero el elemento diferente es la estimación de la varianza del modelo.

Examples

```
serps<-read.table("serps.txt", header=T)
attach(serps); summary(serps)
x <- longitud; y <- peso
# Más descripción de los datos.
length(y); var(y); sd(y)
length(x); var(x); sd(x)
# Nube de puntos.
plot(x,y,col="red",xlim=c(50,75),ylim=c(90,210),pch=16)
```

Examples

```
# Recta de regresión estimada.
lm(y~x); model <- lm(y~x); abline(lm(y~x))
coef(model)
# Valores ajustados.
fitted.values(model)
fitted <- fitted.values(model)
mean(fitted); var(fitted)
# Residuos.
residuals(model)
residuals <- residuals(model)
mean(residuals); var(residuals)
# Estimación de la varianza del modelo.
deviance(model)
SSE <- deviance(model); SSE
squadrat <- SSE/(length(x)-2); squadrat
```