Práctica 3: Regresión lineal simple

Módulo de Modelos Lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

1. Para el banco de datos `Auto` de la librería `ISLR`, queremos estudiar el consumo de distintos vehículos, variable `mpg` (miles per galon). Queremos estudiar esta variable en función de su potencia, variable `horsepower`. Para ello llevaremos a cabo un modelo de regresión lineal simple.

- Evalua la existencia de relación lineal entre ambas variables ¿Encuentras evidencia de que pudiera tener sentido resumir la relación entre ambas variables de forma lineal?

```
data(Auto, package = "ISLR")

with(Auto, cor.test(mpg, horsepower))
```

```
##
##  Pearson's product-moment correlation
##
## data:  mpg and horsepower
## t = -24.489, df = 390, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8146631 -0.7361359
## sample estimates:
##        cor
## -0.7784268
```

```
# Encontramos una evidente (significativa) relación lineal,
# descendente, entre ambas variables.
```

- Ajusta la recta de regresión necesaria para resumir la relación entra ambas variables.

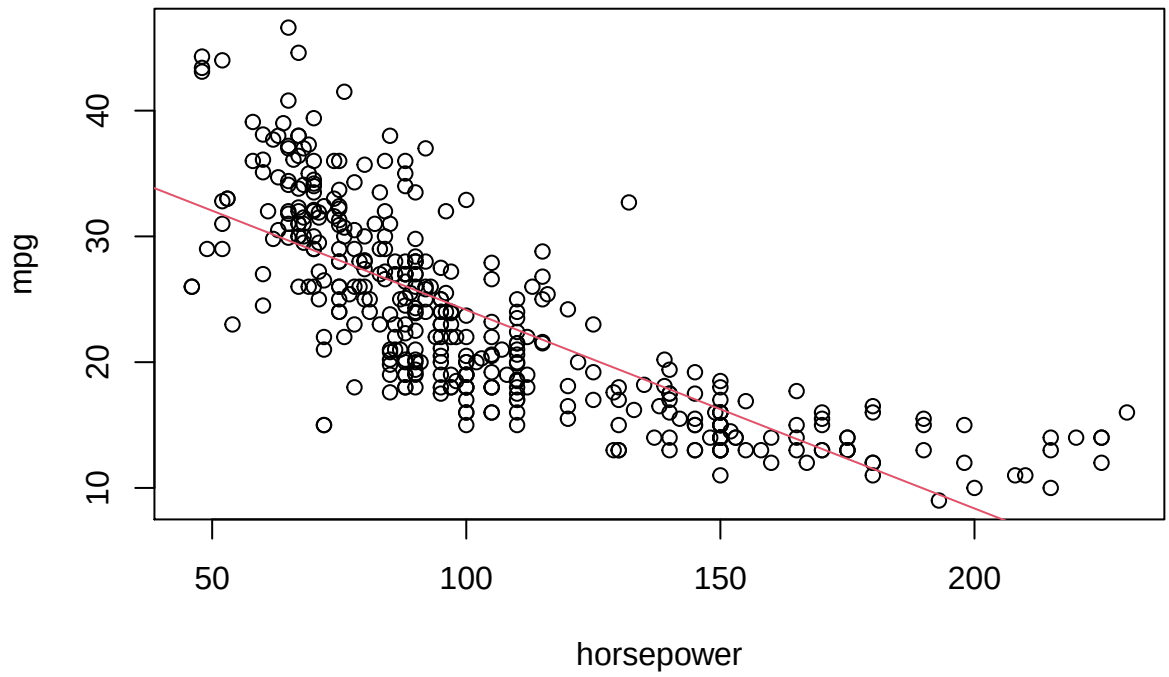
```
modelo <- lm(mpg ~ horsepower, data = Auto)
summary(modelo)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Representa gráficamente la relación entre ambas variables y la recta de regresión que has ajustado.
- ¿Es la relación entre la variable y la respuesta positiva o negativa? Interpreta dicha relación.

```
# Representación de la relación:
with(Auto, plot(horsepower, mpg))
abline(modelo$coef, col = 2)
```




```
# Evidentemente la relación es negativa, cuando aumenta una variable
# disminuye la otra, tal y como se puede ver en la gráfica, en el
# coeficiente de correlación y la pendiente de la recta en el modelo de
# regresión lineal.
```

- ¿Encuentras que la relación entre ambas variables es significativa(mente distinta de 0)? Halla un intervalo de confianza al 95% para el coeficiente asociado a la potencia de los vehículos.

```
confint(modelo)
```

```
##                 2.5 %     97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
```

```r
# La relación entre ambas variables es significativa, tal y como
# podíamos deducir del correspondiente p-valor del modelo de regresión
# y del intervalo de confianza que acabamos de obtener.
```

- ¿Qué `mpg` predecirías para un `horsepower` de 98? Halla un intervalo de confianza para $E(mpg|horsepower = 98)$ y un intervalo de predicción para el valor de `mpg` correspondiente a una potencia de 98 caballos.

```r
# Predicción
predict(modelo, newdata = data.frame(horsepower = c(98)))
```

```
##        1
## 24.46708
```

```r
# IC para la recta
predict(modelo, newdata = data.frame(horsepower = c(98)), interval = c("confidence"))
```

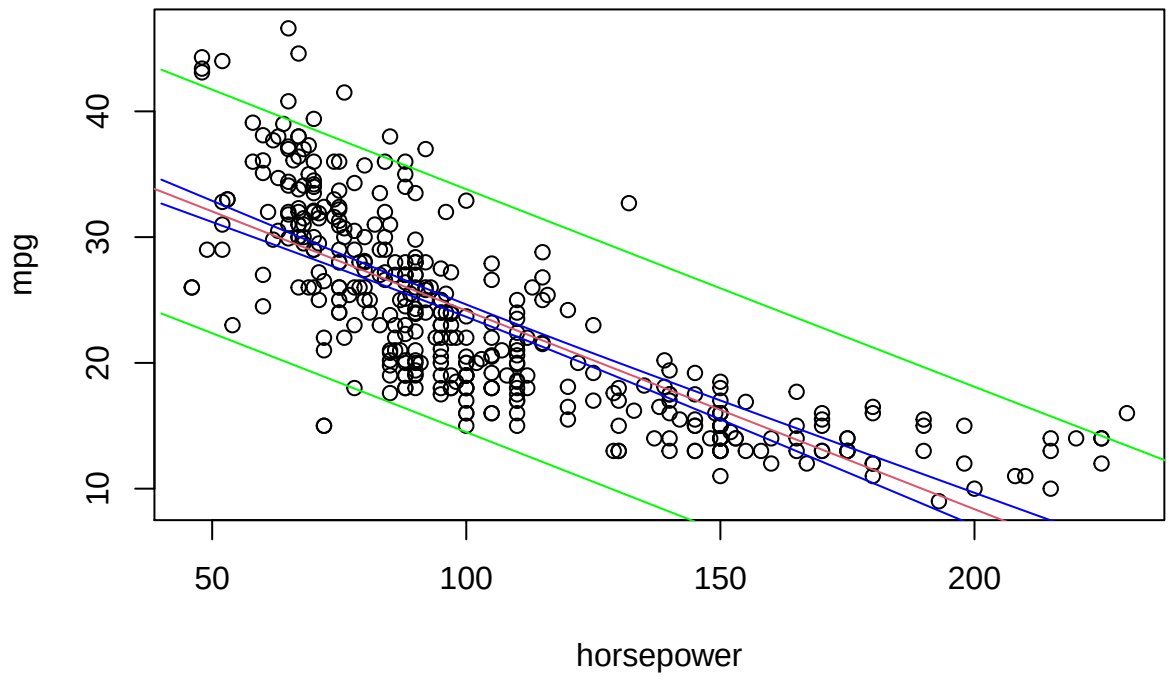
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```r
# Intervalo de predicción para las observaciones
predict(modelo, newdata = data.frame(horsepower = c(98)), interval = c("prediction"))
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

- Representa la nube de puntos junto a la recta de regresión que has ajustado, así como un intervalo de confianza para dicha recta y un intervalo de predicción para el rango de valores de la variable `horsepower`.

```r
# Nube de puntos:
with(Auto, plot(horsepower, mpg))
# Recta de regresión:
abline(modelo$coef, col = 2)
x <- 40:250
# IC:
confianza <- predict(modelo, newdata = data.frame(horsepower = x), interval = c("confidence"))
lines(x, confianza[, 2], col = "blue")
lines(x, confianza[, 3], col = "blue")
# Intervalo de predicción:
predi <- predict(modelo, newdata = data.frame(horsepower = x), interval = c("prediction"))
lines(x, predi[, 2], col = "green")
lines(x, predi[, 3], col = "green")
```




```
# Efectivamente, a simple vista, el intervalo (banda) de predicción
# parece contener el 95% de las observaciones del banco de datos.
```

2. Repite la tarea anterior, pero utilizando el año de fabricación del vehículo (`year`) como variable explicativa. Valora las diferencias entre las conclusiones que extraes de ambos análisis ¿Cuál de los dos ajustes te parece más satisfactorio?

```
with(Auto, cor.test(mpg, year))
```

```
##
##  Pearson's product-moment correlation
##
## data:  mpg and year
## t = 14.08, df = 390, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5108684 0.6426366
## sample estimates:
##      cor
## 0.580541
```

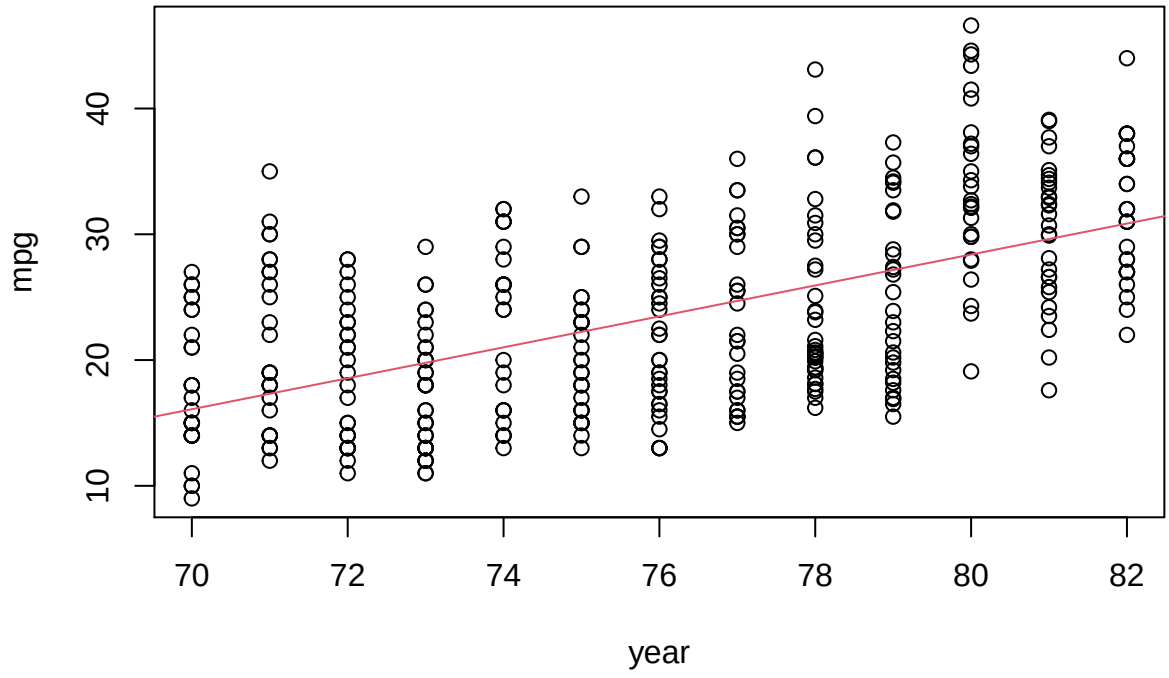
```
# Nuevamente encontramos una clara relación lineal entre ambas
# variables, ahora de tipo ascendente.

modelo2 <- lm(mpg ~ year, data = Auto)
summary(modelo2)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ year, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0212  -5.4411  -0.4412   4.9739  18.2088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70.01167    6.64516  -10.54   <2e-16 ***
## year          1.23004    0.08736   14.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.363 on 390 degrees of freedom
## Multiple R-squared:  0.337,  Adjusted R-squared:  0.3353
## F-statistic: 198.3 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
with(Auto, plot(year, mpg))
abline(modelo2$coef, col = 2)
```




```
# El ajuste anterior parece más satisfactorio en cuanto a la varianza
# explicada ya que la varianza residual ahora aumenta a 6.36 desde 4.91
```

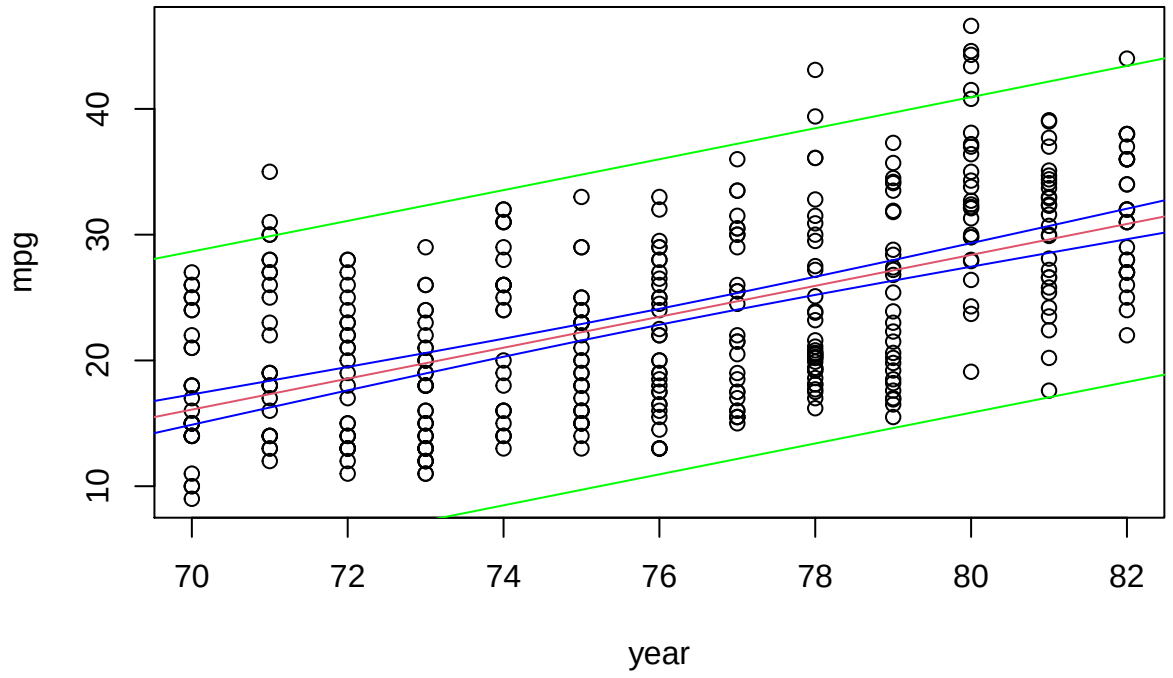
```
confint(modelo2)
```

```
##                   2.5 %     97.5 %
## (Intercept) -83.076498 -56.946851
```

```
## year          1.058285   1.401786
```

```r
# Nuevamente la relación entre variables es significativa como ya
# quedaba claro en la propia salida del modelo lineal.

with(Auto, plot(year, mpg))
abline(modelo2$coef, col = 2)
x <- 69:83
confianza <- predict(modelo2, newdata = data.frame(year = x), interval = c("confidence"))
lines(x, confianza[, 2], col = "blue")
lines(x, confianza[, 3], col = "blue")
predi <- predict(modelo2, newdata = data.frame(year = x), interval = c("prediction"))
lines(x, predi[, 2], col = "green")
lines(x, predi[, 3], col = "green")
```



3. La relación lineal entre `mpg` y `horsepower` ajustada en la Tarea 1 no resulta del todo satisfactoria ya que la nube de puntos se arquea en sus extremos. En ese caso podría parecer más adecuada una relación lineal del tipo `mpg` $\sim$ 1/`horsepower`. Crea la variable `invhorsepower`=1/`horsepower` y ajuste un modelo de regresión lineal simple para `mpg` empleando esta nueva variable como covariable. Representa el ajuste obtenido y valora si la transformación que has hecho de `horsepower` mejora dicho ajuste.

```r
# Variable inversa:
Auto$invhorsepower <- 1/Auto$horsepower

with(Auto, cor.test(mpg, invhorsepower))
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  mpg and invhorsepower
## t = 27.956, df = 390, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7808925 0.8472716
## sample estimates:
##       cor
## 0.8167671
```

```
# La relación lineal entre ambos variables es clara, aunque ahora de
# tipo positivo.

modelo3 <- lm(mpg ~ invhorsepower, data = Auto)
summary(modelo3)
```

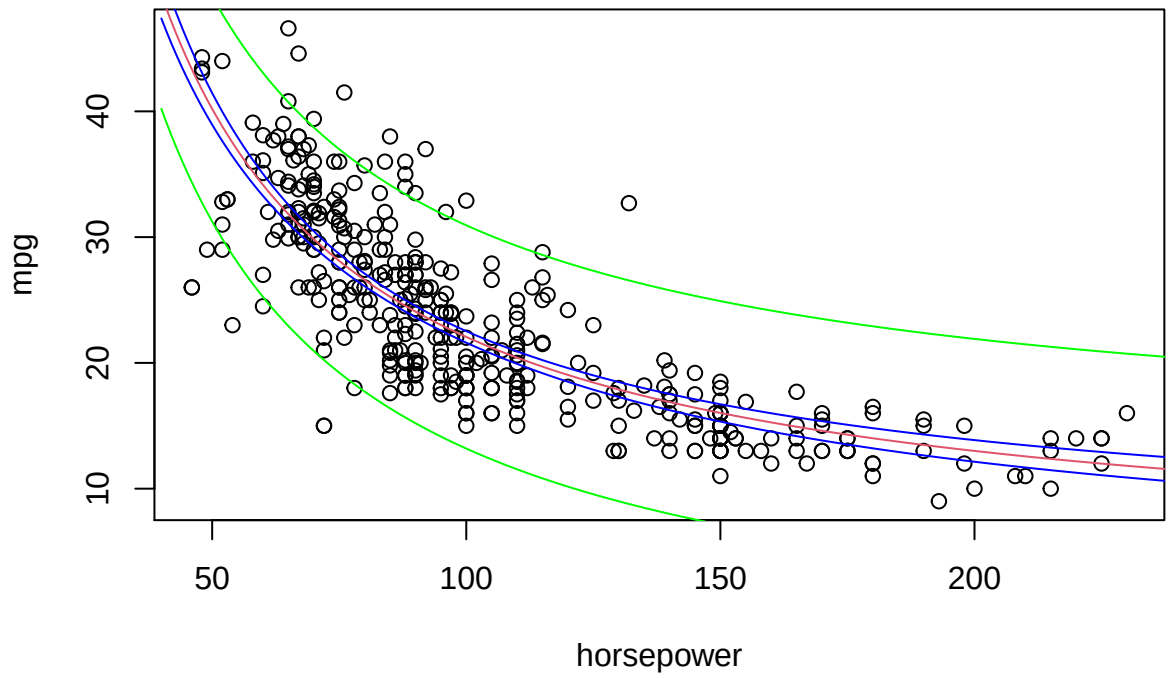
```
##
## Call:
## lm(formula = mpg ~ invhorsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.348  -2.782  -0.106   2.438  15.030
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.9355     0.7341   5.361 1.42e-07 ***
## invhorsepower 1812.9900    64.8509  27.956  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.509 on 390 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.6663
## F-statistic: 781.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
# La varianza residual disminuye ahora a 4.51, por tanto el ajuste con
# esta variable parece más satisfactorio que el de la tarea 1.

# Intervalos de confianza:
confint(modelo3)
```

```
##                    2.5 %      97.5 %
## (Intercept)      2.492211    5.378823
## invhorsepower 1685.488783 1940.491127
```

```
# Representación del ajuste, con intervalos de confianza y predicción:
with(Auto, plot(horsepower, mpg))
x <- 40:250
lines(x, predict(modelo3, newdata = data.frame(invhorsepower = 1/x)), col = 2)
confianza <- predict(modelo3, newdata = data.frame(invhorsepower = 1/x),
    interval = c("confidence"))
lines(x, confianza[, 2], col = "blue")
lines(x, confianza[, 3], col = "blue")
predi <- predict(modelo3, newdata = data.frame(invhorsepower = 1/x), interval = c("prediction"))
lines(x, predi[, 2], col = "green")
lines(x, predi[, 3], col = "green")
```




```
# Este ajuste parece capturar mejor la relación entre ambas variables.
```

4. Plantéate la veracidad o falsedad de las siguientes afirmaciones:

- Siempre que la correlación lineal entre dos variables sea ascendente, el coeficiente de la covariable en el correspondiente modelo de regresión lineal simple será positivo.

```
# Efectivamente, no hay más que ver la última expresión de la
# transparencia de la diapositiva 11 de la sesión 3 de la teoría, donde
# se observa la relación directa entre ambos valores (tener en cuenta
# que el segundo término que multiplica a la correlación es siempre
# positivo).
```

- Podremos reducir la amplitud del intervalo de predicción de un modelo de regresión lineal simple tanto como queramos simplemente elevando el tamaño de la muestra.

```
# Falso, la amplitud del intervalo de predicción viene determinado por
# la varianza de la variable respuesta alrededor de la recta de
# regresión y ésta no disminuirá por más que aumentemos el tamaño de la
# muestra. Caso contrario es del IC de la recta de regresión que sí
# disminuirá su amplitud conforme aumentemos el tamaño de la muestra,
# al disminuir la incertidumbre que tenemos sobre dicha recta.
```

- Para eliminar la correlación entre los coeficientes del modelo de regresión lineal simple es suficiente centrar (restar su media) la covariable del modelo.

```
# Efectivamente, tal y como se muestra en la expresión de la covarianza
# de ambos parámetros del modelo (página 17 de la sesión 3 de teoría),
# dicha covarianza se anulará si la media de la variable explicativa
```

```
# fuera 0, lo que conseguiríamos centrando la variable explicativa
# previamente a aplicar el modelo de regresión.
```

- La suma de los residuos del modelo de regresión valdrá necesariamente 0.

```
# Efectivamente, resulta fácil de demostrar tal y como se muestra a
# continuación:
```

$$\sum e_i = \sum y_i - n * \hat{\beta}_0 - \hat{\beta}_1 \sum x_i = \sum y_i - (\sum y_i - \hat{\beta}_1 \sum x_i) - \hat{\beta}_1 \sum x_i = 0$$