Regresión lineal simple y múltiple
Soluciones a los ejercicios opcionales

Francesc Carmona

14 de marzo de 2018

Ejercicios del libro de Carmona

Ejercicio 6.9

Comparar las rectas de regresión de hombres y mujeres con los logaritmos de los datos del ejercicio 1.4.

Los datos del ejercicio 1.4 son:

```
> TPO_H <- c(9.84,19.32,43.19,102.58,215.78,787.96,1627.34,7956)
> TPO_M <- c(10.94,22.12,48.25,117.73,240.83,899.88,1861.63,8765)
> distancia <- c(100,200,400,800,1500,5000,10000,42192)
> lTPO_H <- log(TPO_H)
> lTPO_M <- log(TPO_M)
> ldistancia <- log(distancia)
```

Como se explica en la sección 6.7.1 vamos a construir un modelo conjunto con los datos de hombres y mujeres. Necesitamos 4 columnas como en la matriz de la página 105:

```
> n <- length(distancia)
> uno.h <- c(rep(1,n),rep(0,n))
> uno.m <- c(rep(0,n),rep(1,n))
> x.h <- c(ldistancia,rep(0,n))
> x.m <- c(rep(0,n),ldistancia)
> y <- c(lTPO_H,lTPO_M)
> modc <- lm(y ~ 0 + uno.h + uno.m + x.h + x.m)
```

Observemos la necesidad de eliminar la constante o intercepción del modelo. En caso contrario el modelo no sería de rango máximo.

Ahora escribimos el modelo donde las pendientes de las dos rectas son iguales, es decir, rectas paralelas:

```
> x <- c(ldistancia,ldistancia)
> modp <- lm(y ~ 0 + uno.h + uno.m + x)
```

y los contrastamos:

```
> anova(modp,modc)

Analysis of Variance Table

Model 1: y ~ 0 + uno.h + uno.m + x
```

```
Model 2: y ~ 0 + uno.h + uno.m + x.h + x.m
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1     13 0.042548
2     12 0.042536  1 1.1479e-05 0.0032 0.9556
```

Aceptamos que las rectas son paralelas y vamos a ver si son coincidentes.

```
> mod0 <- lm(y ~ x)
> anova(mod0,modp)

Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ 0 + uno.h + uno.m + x
  Res.Df      RSS Df Sum of Sq     F   Pr(>F)
1     14 0.100610
2     13 0.042548  1  0.058062 17.74 0.001017 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rechazamos la coincidencia y sólo aceptamos que son paralelas.

Este mismo ejercicio se puede resolver como un modelo ANCOVA.

Ejercicio 8.5

Dado el modelo

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

y los siguientes datos

Y_t	X_{1t}	X_{2t}
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

obtener:

(a) *La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los valores originales.*

Los datos son:

```
> y <- c(10,25,32,43,58,62,67,71)
> x1 <- c(1,3,4,5,7,8,10,10)
> x2 <- c(0,-1,0,1,-1,0,-1,2)
```

y la estimación del modelo

```
> g <- lm(y ~ x1 + x2)
> coef(g)

(Intercept)          x1          x2
  6.4699828   6.5883362   0.2572899
```

(b) *La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los datos expresados en desviaciones respecto de la media.*

Los datos transformados son:

```
> ys <- scale(y, center=TRUE, scale=FALSE)
> x1s <- scale(x1, center=TRUE, scale=FALSE)
> x2s <- scale(x2, center=TRUE, scale=FALSE)
> gs <- lm(ys ~ 0 + x1s + x2s)
> coef(gs)

      x1s       x2s
6.5883362 0.2572899
```

Vemos que las estimaciones de los coeficientes asociados a las variables regresoras son iguales.

Observemos que con las variables centradas no hay que poner el coeficiente de intercepción ya que el hiperplano de regresión debe pasar por el origen de coordenadas. Es decir, cuando las variables regresoras centradas valen cero, la respuesta es cero.

(c) *La estimación insesgada de σ^2.*

```
> sg <- summary(g)
> sgs <- summary(gs)
> c(sg$sigma^2, sgs$sigma^2)

[1] 18.33002 15.27501
```

Las estimaciones de σ^2 son distintas.

(d) *El coeficiente de determinación.*

```
> c(sg$r.squared,sgs$r.squared)

[1] 0.9731074 0.9731074
```

La definición del coeficiente de determinación es distinta para los modelos con y sin intercepcción. En este caso particular coinciden.

(e) *El coeficiente de determinación corregido.*

```
> c(sg$adj.r.squared,sgs$adj.r.squared)

[1] 0.9623503 0.9641432
```

Aquí sí son distintos.

(f) *El contraste de la hipótesis nula $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.*

```
> g0 <- lm(y ~ 0)
> anova(g0,g)

Analysis of Variance Table

Model 1: y ~ 0
```

```
Model 2: y ~ x1 + x2
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1      8 20336.0
2      5    91.7  3     20244 368.15 2.773e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se rechaza.

(g) *El contraste de la hipótesis nula* $H_0 : \beta_1 = \beta_2 = 0$ *utilizando los datos originales.*

```
> g1 <- lm(y ~ 1)
> anova(g1,g)

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1      7 3408.0
2      5   91.7  2    3316.3 90.462 0.0001186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

También se rechaza

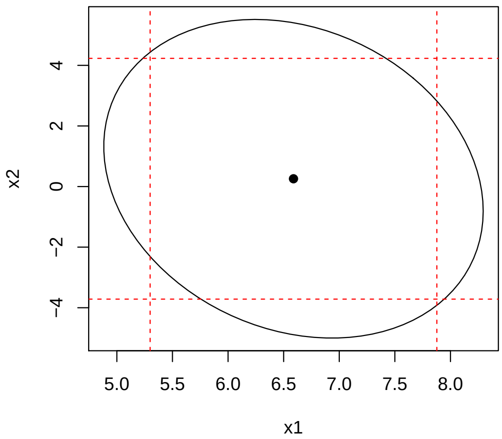
(h) *El contraste de la hipótesis nula* $H_0 : \beta_1 = \beta_2 = 0$ *utilizando los datos en desviaciones respecto a la media.*

```
> g1s <- lm(ys ~ 1)
> anova(g1s,gs)

Analysis of Variance Table

Model 1: ys ~ 1
Model 2: ys ~ 0 + x1s + x2s
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      7 3408.0
2      6   91.7  1    3316.3 217.11 6.14e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se rechaza.

(i) *La representación gráfica de una región de confianza del 95 % para* β_1 *y* β_2.

```
> library(ellipse)
> plot(ellipse(g,2:3),type="l")
> points(coef(g)[2], coef(g)[3], pch=19)
> abline(v=confint(g)[2,],lty=2,col=2)
> abline(h=confint(g)[3,],lty=2,col=2)
```



(j) *El contraste individual de los parámetros β_0, β_1 y β_2.*

Se pueden contrastar con los test t que vemos en el resumen.

```
> sg


Call:
lm(formula = y ~ x1 + x2)

Residuals:
      1       2       3       4       5       6       7       8
-3.0583 -0.9777 -0.8233  3.3310  5.6690  2.8233 -5.0961 -1.8679

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4700     3.3684   1.921    0.113
x1            6.5883     0.5015  13.137 4.56e-05 ***
x2            0.2573     1.5458   0.166    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.281 on 5 degrees of freedom
Multiple R-squared:  0.9731,Adjusted R-squared:  0.9624
F-statistic: 90.46 on 2 and 5 DF,  p-value: 0.0001186
```

Sólo β_1 es significativo.

(k) *El contraste de la hipótesis nula $H_0 : \beta_1 = 10\beta_2$.*

```
> g10 <- lm(y ~ I(10*x1 + x2))
> anova(g10,g)

Analysis of Variance Table

Model 1: y ~ I(10 * x1 + x2)
```

```
Model 2: y ~ x1 + x2
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1      6 92.87
2      5 91.65  1    1.2195 0.0665 0.8067
```

Se acepta.

(l) *El contraste de la hipótesis nula* $H_0 : 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

Haremos el contraste con el estadístico t de Student para una función paramétrica:

```
> a <- c(2,2,7)
> betas <- coef(g)
> X <- model.matrix(g)
> numerador <- t(a) %*% betas - 50
> denominador <- sqrt(sg$sigma^2 * t(a) %*% solve(t(X)%*%X) %*% a)
> t.est <- numerador/denominador
> c(t.est, pt(t.est, 8-3)*2)

[1] -1.6768930  0.1544087
```

Se acepta.

(k) *El contraste de la hipótesis nula conjunta* $H_0 : \beta_1 = 10\beta_2,\ 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

Haremos el contraste con el estadístico F para un conjunto de funciones paramétricas. En forma matricial la hipótesis a contrastar es:

$$H_0 : \begin{pmatrix} 0 & 1 & -10 \\ 2 & 2 & 7 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 50 \end{pmatrix}$$

```
> A <- matrix(c(0,1,-10,
+               2,2,7), ncol=3, byrow = T)
> cc <- c(0,50)
> XtXinv <- solve(t(X) %*% X)
> numerador <- t(A %*% betas - cc) %*% solve(A %*% XtXinv %*% t(A)) %*% (A %*% betas - cc)/2
> denominador <- sg$sigma^2
> F.est <- numerador/denominador
> p.valor <- pf(F.est,2,8-3,lower.tail = F)
> c(F.est, p.valor)

[1] 6.09795790 0.04558918
```

Rechazamos la hipótesis.

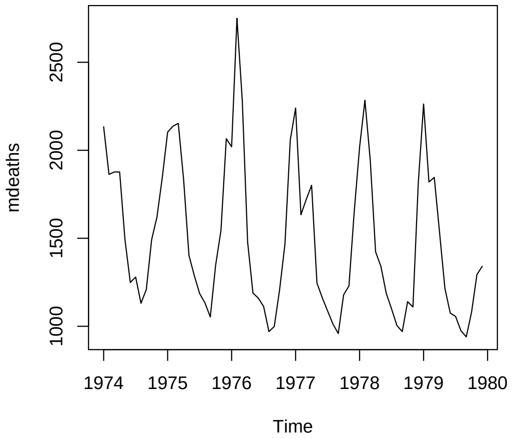
Ejercicios del libro de Faraway. Capítulo 4.

Ejercicio 4

The dataset `mdeaths` *reports the number of deaths from lung diseases for men in the UK from 1974 to 1979.*

(a) *Make an appropriate plot of the data. At what time of year are deaths most likely to occur?*

```
> library(datasets)
> data(UKLungDeaths)
> plot(mdeaths)
```



Se trata de una serie temporal con los fallecimientos mensuales de hombres por bronquitis, emfisema y asma en UK, 1974-1979.

(b) *Fit an autoregressive model of the same form used for the airline data. Are all the predictors statistically significant?*

```
> lagdf <- embed(as.vector(mdeaths),14)
> colnames(lagdf) <- c("y",paste0("lag",1:13))
> lagdf <- data.frame(lagdf)
> armod <- lm(y ~ lag1 + lag12 + lag13, data=lagdf)
> summary(armod)


Call:
lm(formula = y ~ lag1 + lag12 + lag13, data = lagdf)

Residuals:
    Min      1Q  Median      3Q     Max
-762.71  -81.13  -21.12   61.76  724.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.1985   120.7358   0.482   0.6317
lag1          0.2501     0.1327   1.885   0.0647 .
lag12         0.5356     0.1179   4.542 3.09e-05 ***
lag13         0.1512     0.1386   1.091   0.2801
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 238.7 on 55 degrees of freedom
```

```
Multiple R-squared:   0.73,Adjusted R-squared:  0.7153
F-statistic: 49.56 on 3 and 55 DF,  p-value: 1.19e-15
```

El único coeficiente significativo es el que corresponde al mismo mes del año anterior.

(c) *Use the model to predict the number of deaths in January 1980 along with a 95 % prediction interval.*

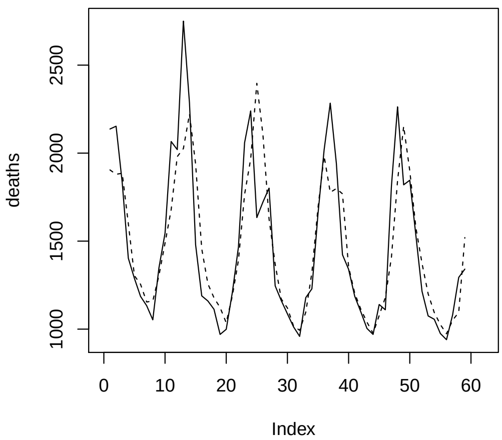
```
> lagdf[nrow(lagdf),]

      y lag1 lag2 lag3 lag4 lag5 lag6 lag7 lag8 lag9 lag10 lag11 lag12 lag13
59 1341 1294 1081  940  975 1056 1075 1215 1531 1846  1820  2263  1812  1110

> predict(armod, data.frame(lag1=1341, lag12=2263, lag13=1812),interval="prediction")

       fit      lwr      upr
1 1879.599 1359.725 2399.474
```

(d) *Use your answer from the previous question to compute a prediction and interval for February 1980.*

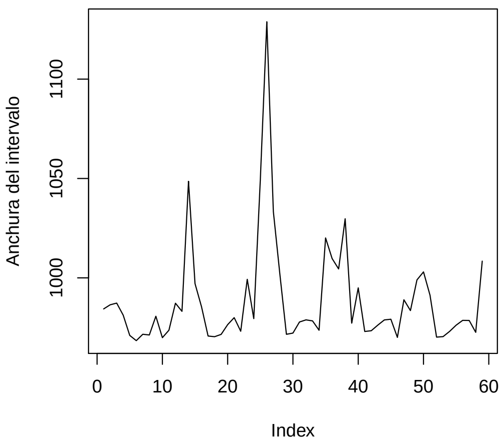
```
> predict(armod, data.frame(lag1=1879.599, lag12=1820, lag13=2263),interval="prediction")

       fit     lwr      upr
1 1845.247 1345.87 2344.625
```

(e) *Compute the fitted values. Plot these against the observed values. Note that you will need to select the appropriate observed values. Do you think the accuracy of predictions will be the same for all months of the year?*

```
> plot(lagdf$y, type="l", xlim=c(0,62), ylab="deaths")
> lines(predict(armod), lty=2)
```



```
> pred.int <- predict(armod, interval = "prediction")
> plot(pred.int[,3]-pred.int[,2], type="l", ylab="Anchura del intervalo")
> which.max(pred.int[,3]-pred.int[,2])

26
26
```



Hay meses con un intervalo de predicción más ancho, en concreto el mes de marzo.

Ejercicio 5

For the `fat` data used in this chapter, a smaller model using only `age`, `weight`, `height` and `abdom` was proposed on the grounds that these predictors are either known by the individual or easily measured.

(a) *Compare this model to the full thirteen-predictor model used earlier in the chapter. Is it justifiable to use the smaller model?*

```
> data(fat,package="faraway")
> lmod <- lm(brozek ~ age + weight + height + neck + chest + abdom +
+              hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
> lmod0 <- lm(brozek ~ age + weight + height + abdom, data=fat)
> anova(lmod0,lmod)

Analysis of Variance Table

Model 1: brozek ~ age + weight + height + abdom
Model 2: brozek ~ age + weight + height + neck + chest + abdom + hip +
    thigh + knee + ankle + biceps + forearm + wrist
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    247 4205.0
2    238 3785.1  9     419.9 2.9336 0.002558 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El contraste de modelos es significativo de modo que no podemos quedarnos con el modelo simple.

```
> summary(lmod)$adj.r.squared

[1] 0.7352688

> summary(lmod0)$adj.r.squared

[1] 0.7166171
```

Sin embargo, el ajuste es bastante parecido.

(b) *Compute a 95 % prediction interval for median predictor values and compare to the results to the interval for the full model. Do the intervals differ by a practically important amount?*

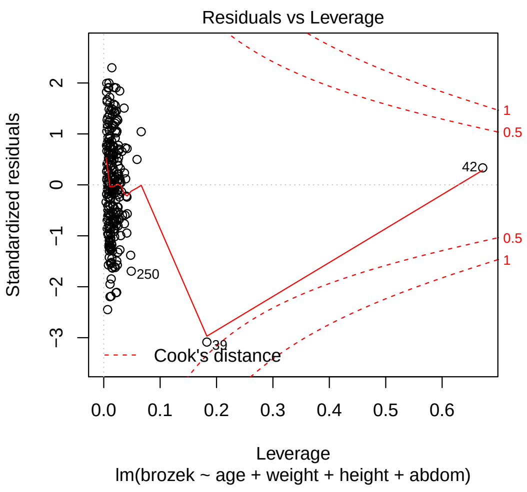
```
> medianas <- apply(fat[,4:18],2,median)
> predict(lmod0, newdata = data.frame(age=medianas[1],
+                                     weight=medianas[2],
+                                     height=medianas[3],
+                                     abdom=medianas[8]), interval="prediction")

         fit      lwr      upr
age 17.84028 9.696631 25.98392

> predict(lmod, newdata = as.data.frame(t(medianas[c(1:3,6:15)])), interval="prediction")

       fit     lwr      upr
1 17.49322 9.61783 25.36861
```

Ciertamente no hay mucha diferencia entre los dos intervalos.

(c) *For the smaller model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?*

Si hacemos un análisis gráfico de los residuos del modelo simple tenemos

```
> plot(lmod0, which=5)
```



En este gráfico vemos claramente dos observaciones con un *leverage* elevado. Son la observación 39 y la 42. Eso significa que para las variables regresoras son puntos muy alejados de la media de todos los datos.

(d) *Recompute the 95 % prediction interval for median predictor values after these two anomalous cases have been excluded from the data. Did this make much difference to the outcome?*

```
> lmod0 <- lm(brozek ~ age + weight + height + abdom, data=fat[-c(39,42),])
> medianas <- apply(fat[-c(39,42),4:18],2,median)
> predict(lmod0, newdata = data.frame(age=medianas[1],
+                                     weight=medianas[2],
+                                     height=medianas[3],
+                                     abdom=medianas[8]), interval="prediction")

        fit      lwr      upr
age 17.9033 9.887851 25.91874
```

Parece que el intervalo ha mejorado pero muy poco.