Ejercicios\_Regresión\_simple\_multiple

David Cáceres

11/4/2022

#### 1. (Ejercicio 6.8 del Capítulo 6 página 118)

#### Hallar la recta de regresión simple de la variable respuesta raíz cuadrada de la velocidad sobre la variable regresora densidad con los datos de la tabla 1.1 del capítulo 1.

require(faraway)

## Loading required package: faraway

dens <- c(12.7,17.0,66.0,50.0,87.8,81.4,75.6,66.2,81.1,62.8,77.0,89.6,  
 18.3,19.1,16.5,22.2,18.6,66.0,60.3,56.0,66.3,61.7,66.6,67.8)  
vel <- c(62.4,50.7,17.1,25.9,12.4,13.4,13.7,17.9,13.8,17.9,15.8,12.6,  
 51.2,50.8,54.7,46.5,46.3,16.9,19.8,21.2,18.3,18.0,16.6,18.3)  
rvel <- sqrt(vel)  
lmod1<-lm(rvel ~ dens)  
summary(lmod1)

##   
## Call:  
## lm(formula = rvel ~ dens)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35337 -0.22722 -0.03566 0.18942 0.53349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.089813 0.130629 61.93 <2e-16 \*\*\*  
## dens -0.056626 0.002177 -26.01 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2689 on 22 degrees of freedom  
## Multiple R-squared: 0.9685, Adjusted R-squared: 0.9671   
## F-statistic: 676.4 on 1 and 22 DF, p-value: < 2.2e-16

#### (i) La suma de los residuos es cero

residuos<- residuals(lmod1)  
sum(residuos)

## [1] 8.881784e-16

# El valor es prácticamente cero por redondeo

#### (ii) Pyi = Pyˆi

sum(rvel)

## [1] 120.1685

sum(fitted(lmod1))

## [1] 120.1685

#### (iii) La suma de los residuos ponderada por los valores de la variable regresora es cero:

sum(residuos\*dens)

## [1] -1.566802e-14

#### (iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero:

sum(fitted(lmod1)\*residuos)

## [1] 2.491063e-15

#### Calcular la estimación de σ² y, a partir de ella, las estimaciones de las desviaciones estándar de los estimadores de los parámetros βˆ0 y βˆ1.

length<-length(dens)  
SCRlmod1<-sum(residuos^2)  
sigma<-SCRlmod1/(length-2)  
sigma

## [1] 0.07232808

s2dens<-(1/length)\*sum((dens-mean(dens))^2)  
Sdens<-length\*s2dens  
ebeta1lmod1<-sqrt(sigma\*((1/length)+(mean(dens)^2/Sdens)))

ebeta0lmod1<-sqrt(sigma\*((1/length)+(mean(dens)^2/Sdens)))  
ebeta0lmod1

## [1] 0.1306294

#### Escribir los intervalos de confianza para los parámetros con un nivel de confianza del 95 %.

tlmod1 <- qt(0.975,lmod1$df)  
# Para Beta1  
lmod1$coef[2]+c(-1,1)\*tlmod1\*ebeta1lmod1

## [1] -0.3275345 0.2142833

# Para beta0  
lmod1$coef[1]+c(-1,1)\*tlmod1\*ebeta0lmod1

## [1] 7.818904 8.360722

#### Construir la tabla para la significación de la regresión y realizar dicho contraste. (H0: ß1=0 (No hay efecto de la relación entre ambas variables)

summary(aov(lmod1))

## Df Sum Sq Mean Sq F value Pr(>F)   
## dens 1 48.92 48.92 676.4 <2e-16 \*\*\*  
## Residuals 22 1.59 0.07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Significativo a p<<alfa=0.005

#### Hallar el intervalo de la predicción de la respuesta media cuando la densidades de 50 vehículos por km. Nivel de confianza: 90 %

y0 <- lmod1$coef[1]+lmod1$coef[2]\*50  
  
y0

## (Intercept)   
## 5.258534

tlmod1 <- qt(0.950,lmod1$df)  
sigma <- sqrt(sigma)  
y0+c(-1,1)\*tlmod1\*sigma\*sqrt(1+(1/length)+((((50-mean(dens))^2)/Sdens)))

## [1] 4.786912 5.730156

predict(lmod1,new=data.frame(dens=50),interval="prediction",level=0.90)

## fit lwr upr  
## 1 5.258534 4.786912 5.730156

predict(lmod1,new=data.frame(dens=50),interval="confidence",level=0.90)

## fit lwr upr  
## 1 5.258534 5.162817 5.354251

### 3. (Ejercicio 6.10 del Capítulo 6 página 118)

####Se admite que una persona es proporcionada si su altura en cm es igual a su peso en kg más 100. En términos estadísticos si la recta de regresión de Y (altura) sobre X (peso) es:

Y = 100 + X

#### Contrastar, con un nivel de significación α = 0.05, si se puede considerar válida esta hipótesis a partir de los siguientes datos que corresponden a una muestra de mujeres jóvenes:

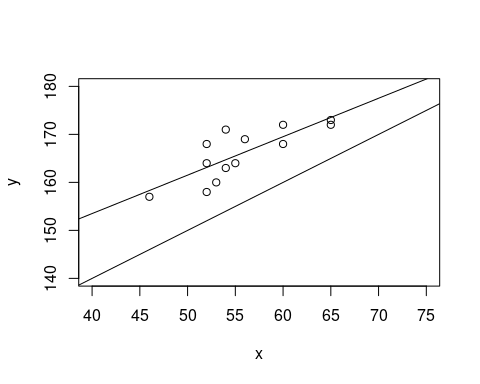
x <- c(55,52,65,54,46,60,54,52,56,65,52,53,60)  
y <- c(164,164,173,163,157,168,171,158,169,172,168,160,172)

#### Razonar la bondad de la regresión y todos los detalles del contraste.

lmod2<-lm(y ~ x)  
summary(lmod2)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.120 -1.531 -1.315 2.473 6.278   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 121.4777 10.1857 11.926 1.24e-07 \*\*\*  
## x 0.8008 0.1821 4.398 0.00107 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.449 on 11 degrees of freedom  
## Multiple R-squared: 0.6375, Adjusted R-squared: 0.6045   
## F-statistic: 19.34 on 1 and 11 DF, p-value: 0.001067

plot(y~x, xlim=c(40,75), ylim=c(140, 180))  
abline(lmod2)  
abline(100,1)



confint(lmod2)

## 2.5 % 97.5 %  
## (Intercept) 99.0591808 143.896222  
## x 0.4000513 1.201578

modelo0<-lm(y ~ 0 + offset(100 + x))  
anova(modelo0, lmod2)

## Analysis of Variance Table  
##   
## Model 1: y ~ 0 + offset(100 + x)  
## Model 2: y ~ x  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 13 1547.00   
## 2 11 130.84 2 1416.2 59.528 1.259e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Rechazamos la hipótesis nula

#### 4. (Ejercicio 6.11 del Capítulo 6 página 119)

El período de oscilación de un péndulo es 2π| l/g donde l es la longitud y g es la constante de gravitación. En un experimento observamos t ij (j = 1, . . . , n i ) períodos correspondientes a l i (i = 1, . . . , k) longitudes.

#### (a) Proponer un modelo, con las hipótesis que se necesiten, para estimar la constante método de los mínimos cuadrados.

longitud<- c(rep(18.3,4),rep(20,2),rep(21.5,3),rep(15,2))  
x<-sqrt(longitud)  
y<-c(8.58,7.9,8.2,7.8,8.4,9.2,9.7,8.95,9.2,7.5,8)  
lmod3<-lm(y ~ 0 + sqrt(x))  
summary(lmod3)

##   
## Call:  
## lm(formula = y ~ 0 + sqrt(x))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.64760 -0.39276 -0.03791 0.28012 0.90512   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## sqrt(x) 4.08433 0.07194 56.77 6.97e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4969 on 10 degrees of freedom  
## Multiple R-squared: 0.9969, Adjusted R-squared: 0.9966   
## F-statistic: 3223 on 1 and 10 DF, p-value: 6.973e-14

#### (b) Contrastar H0:2πg√

# Miramos si el intervalo de confianza contiene el 2  
  
confint(lmod3)

## 2.5 % 97.5 %  
## sqrt(x) 3.924036 4.244618

No lo contiene así que rechazamos la hipótesis nula

#### 5. (Ejercicio 8.4 del Capítulo 8 página 157)

Se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

# Datos  
  
v <- c(60,48,42,36,78,36,72,42,54,90)  
ip <- c(100,110,130,100,80,80,90,120,120,90)  
pu <- c(1.8,2.4,3.6,0.6,1.8,0.6,3.6,1.2,2.4,4.2)

lmod4<-lm(v ~ ip +pu)  
summary(lmod4)

##   
## Call:  
## lm(formula = v ~ ip + pu)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.865 -5.944 0.869 7.516 13.051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.2462 21.7077 4.388 0.00320 \*\*  
## ip -0.6240 0.2112 -2.954 0.02128 \*   
## pu 10.9038 2.9208 3.733 0.00733 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.94 on 7 degrees of freedom  
## Multiple R-squared: 0.7359, Adjusted R-squared: 0.6604   
## F-statistic: 9.752 on 2 and 7 DF, p-value: 0.009469

#### 1) Estimar el vector de coeficientes β = (β 0 , β 1 , β 2 ) 0 del modelo

lmod4$coefficients

## (Intercept) ip pu   
## 95.2462340 -0.6240469 10.9038497

#### 2) Estimar la matriz de varianzas-covarianzas del vector β.

summary(lmod4)$sigma^2 \* summary(lmod4)$cov.unscaled

## (Intercept) ip pu  
## (Intercept) 471.224820 -4.32088690 -8.3457769  
## ip -4.320887 0.04462209 -0.1038586  
## pu -8.345777 -0.10385856 8.5312386

#### 3) Calcular el coeficiente de determinación.

sum4<-summary(lmod4)  
names(sum4)

## [1] "call" "terms" "residuals" "coefficients"   
## [5] "aliased" "sigma" "df" "r.squared"   
## [9] "adj.r.squared" "fstatistic" "cov.unscaled"

sum4$r.squared

## [1] 0.7358837

## Ejercicios del libro de Faraway

#### 1. (Ejercicio 1 cap. 4 pág. 56)

For the prostate data, fit a model with lpsa as the response and the other variables as predictors:

require(faraway)  
data(prostate)

lmod5<- lm(lpsa~lcavol + lweight + age + lbph + svi +lcp+ gleason+pgg45, data = prostate)  
summary(lmod5)

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +   
## gleason + pgg45, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7331 -0.3713 -0.0170 0.4141 1.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.669337 1.296387 0.516 0.60693   
## lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*  
## lweight 0.454467 0.170012 2.673 0.00896 \*\*   
## age -0.019637 0.011173 -1.758 0.08229 .   
## lbph 0.107054 0.058449 1.832 0.07040 .   
## svi 0.766157 0.244309 3.136 0.00233 \*\*   
## lcp -0.105474 0.091013 -1.159 0.24964   
## gleason 0.045142 0.157465 0.287 0.77503   
## pgg45 0.004525 0.004421 1.024 0.30886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7084 on 88 degrees of freedom  
## Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234   
## F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

#### (a) Suppose a new patient with the following values arrives:

Predict the lpsa for this patient along with an appropriate 95% CI.

head(x0pros <- data.frame(lcavol=1.44692,  
 lweight=3.62301,  
 age=65,  
 lbph=0.30010,  
 svi=0,  
 lcp=-0.79851,  
 gleason=7,  
 pgg45=15))

## lcavol lweight age lbph svi lcp gleason pgg45  
## 1 1.44692 3.62301 65 0.3001 0 -0.79851 7 15

predict(lmod5, x0pros, interval="prediction", level=0.95)

## fit lwr upr  
## 1 2.389053 0.9646584 3.813447

#### (b) Repeat the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

x1pros <- data.frame(lcavol=1.44692,lweight=3.62301,age=20,lbph=0.30010,svi=0,lcp=-0.79851,gleason=7,pgg45=15)  
predict(lmod5,x1pros,interval="prediction")

## fit lwr upr  
## 1 3.272726 1.538744 5.006707

E el apartado B, el valor para la edad, está fuera del rango de los valores del dataset, por eso el intervalo de confianza es mayor.

#### (c) For the model of the previous question, remove all the predictors that are not significant at the 5% level. Now recompute the predictions of the previous question. Are the CIs wider or narrower? Which predictions would you prefer? Explain.

summary(lmod5)

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +   
## gleason + pgg45, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7331 -0.3713 -0.0170 0.4141 1.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.669337 1.296387 0.516 0.60693   
## lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*  
## lweight 0.454467 0.170012 2.673 0.00896 \*\*   
## age -0.019637 0.011173 -1.758 0.08229 .   
## lbph 0.107054 0.058449 1.832 0.07040 .   
## svi 0.766157 0.244309 3.136 0.00233 \*\*   
## lcp -0.105474 0.091013 -1.159 0.24964   
## gleason 0.045142 0.157465 0.287 0.77503   
## pgg45 0.004525 0.004421 1.024 0.30886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7084 on 88 degrees of freedom  
## Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234   
## F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

Las variables con significación al 5% son “lcavol”, “lweight”, “svi”. Creamos un nuevo modelo con ellas.

lmod6<-lm(lpsa ~ lcavol + lweight + svi, data=prostate)  
summary(lmod6)

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.72964 -0.45764 0.02812 0.46403 1.57013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.26809 0.54350 -0.493 0.62298   
## lcavol 0.55164 0.07467 7.388 6.3e-11 \*\*\*  
## lweight 0.50854 0.15017 3.386 0.00104 \*\*   
## svi 0.66616 0.20978 3.176 0.00203 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7168 on 93 degrees of freedom  
## Multiple R-squared: 0.6264, Adjusted R-squared: 0.6144   
## F-statistic: 51.99 on 3 and 93 DF, p-value: < 2.2e-16

predict(lmod6, x0pros, interval="prediction")

## fit lwr upr  
## 1 2.372534 0.9383436 3.806724

predict(lmod6, x1pros, interval="prediction")

## fit lwr upr  
## 1 2.372534 0.9383436 3.806724

#### 2. (Ejercicio 2 cap. 4 pág. 57)

Using the teengamb data, fit a model with gamble as the response and the other variables as predictors.

data(teengamb)  
lmod7<-lm(gamble ~ sex + status + income +verbal, data=teengamb)  
summary(lmod7)

##   
## Call:  
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.082 -11.320 -1.451 9.452 94.252   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.55565 17.19680 1.312 0.1968   
## sex -22.11833 8.21111 -2.694 0.0101 \*   
## status 0.05223 0.28111 0.186 0.8535   
## income 4.96198 1.02539 4.839 1.79e-05 \*\*\*  
## verbal -2.95949 2.17215 -1.362 0.1803   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.69 on 42 degrees of freedom  
## Multiple R-squared: 0.5267, Adjusted R-squared: 0.4816   
## F-statistic: 11.69 on 4 and 42 DF, p-value: 1.815e-06

#### (a) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI.

attach(teengamb)  
x0gamb<-data.frame( sex=0, status=median(status),   
 income=median(income), verbal=median(verbal))  
  
predict(lmod7,x0gamb,interval="prediction")

## fit lwr upr  
## 1 20.21168 -26.98701 67.41037

#### (b) Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

x1gamb <- data.frame(sex=0, status=max(status),  
 income=max(income),verbal=max(verbal))  
  
predict(lmod7,x1gamb,interval="prediction")

## fit lwr upr  
## 1 71.30794 17.06588 125.55

#### (c) Fit a model with sqrt(gamble) as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

lmod8<-lm(sqrt(gamble) ~ sex + status + income +verbal, data=teengamb)  
predict(lmod8,x0gamb,interval="prediction")

## fit lwr upr  
## 1 3.155673 -1.179373 7.490718

predict(lmod8,x0gamb,interval="prediction")^2

## fit lwr upr  
## 1 9.958271 1.39092 56.11086

#### (d) Repeat the prediction for the model in (c) for a female with status = 20, income = 1, verbal = 10. Comment on the credibility of the result

x2gamb <- data.frame(sex=1,status=20,income=1,verbal=10)  
 predict(lmod8,x2gamb,interval="prediction")^2

## fit lwr upr  
## 1 4.353398 47.73238 7.485167

## 3. (Ejercicio 3 cap. 4 pág. 57)

The snail dataset contains percentage water content of the tissues of snails grown under three different levels of relative humidity and two different temperatures.

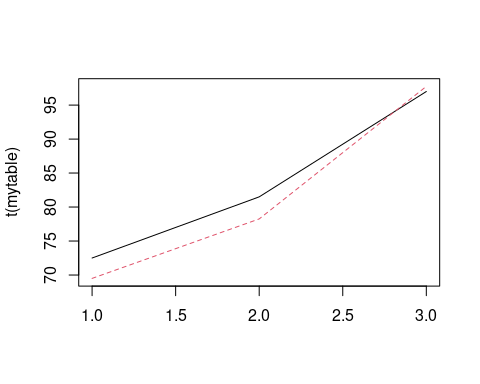
#### (a) Use the command xtabs(water ~ temp + humid, snail)/4 to produce a table of mean water content for each combination of temperature and humidity. Can you use this table to predict the water content for a temperature of 25 ◦ C and a humidity of 60%? Explain.

require(faraway)  
data(snail)

xtabs(water ~ temp + humid, snail)/4

## humid  
## temp 45 75 100  
## 20 72.50 81.50 97.00  
## 30 69.50 78.25 97.75

mytable <- xtabs(water~temp+humid,snail)/4  
colnames(mytable) <- c(45,75,100)  
rownames(mytable) <- c(20, 30)  
matplot (t(mytable), type="l")



xtabs(mean(water[humid<100])~mean(temp[humid<100])+mean(humid[humid<100]),snail)

## mean(humid[humid < 100])  
## mean(temp[humid < 100]) 60  
## 25 75.4375

#### (b) Fit a regression model with the water content as the response and the temperature and humidity as predictors. Use this model to predict the water content for a temperature of 25 ◦ C and a humidity of 60%?

lmod9<-(lm(water ~ temp + humid, data=snail))  
summary(lmod9)

##   
## Call:  
## lm(formula = water ~ temp + humid, data = snail)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.456 -2.915 1.461 3.613 8.749   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52.61081 6.85346 7.677 1.59e-07 \*\*\*  
## temp -0.18333 0.22645 -0.810 0.427   
## humid 0.47349 0.05036 9.403 5.63e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.547 on 21 degrees of freedom  
## Multiple R-squared: 0.8092, Adjusted R-squared: 0.791   
## F-statistic: 44.53 on 2 and 21 DF, p-value: 2.793e-08

x0wat<-data.frame(temp=25,humid=60)  
predict(lmod9, x0wat, interval="prediction")

## fit lwr upr  
## 1 76.43681 64.58094 88.29269

#### (c) Use this model to predict water content for a temperature of 30 ◦ C and a humidity of 75%?. Compare your prediction to the prediction from (a). Discuss the relative merits of these two predictions.

x1wat<-data.frame(temp=30,humid=75)  
predict(lmod9, x1wat, interval="prediction")

## fit lwr upr  
## 1 82.62248 70.6147 94.63027

#### (d) The intercept in your model is 52.6%. Give two values of the predictors for which this represents the predicted response. Is your answer unique? Do you think this represents a reasonable prediction?

x3wat<-data.frame(temp=0,humid=0)  
predict(lmod9,x3wat,interval="prediction")

## fit lwr upr  
## 1 52.61081 34.27498 70.94663

#### (e) For a temperature of 25 ◦ C, what value of humidity would give a predicted response of 80% water content.

x6wat<-data.frame(temp=25,humid=seq(67.1,68.0,0.1))  
predict(lmod9,x6wat,interval="prediction")

## fit lwr upr  
## 1 79.79859 68.00714 91.59003  
## 2 79.84593 68.05507 91.63680  
## 3 79.89328 68.10298 91.68358  
## 4 79.94063 68.15089 91.73038  
## 5 79.98798 68.19878 91.77718  
## 6 80.03533 68.24667 91.82399  
## 7 80.08268 68.29455 91.87081  
## 8 80.13003 68.34242 91.91764  
## 9 80.17738 68.39028 91.96448  
## 10 80.22473 68.43813 92.01133

## Ejercicios Opcionales

## Ejercicios del libro de Carmona

#### 2. (∗) (Ejercicio 6.9 del Capítulo 6 página 118)

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

# Datos  
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)

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)

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

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

#### 6. (∗) (Ejercicio 8.5 del Capítulo 8 página 157)

Dado el modelo Y t = β 0 + β 1 X 1t + β 2 X 2t + u t y los siguientes datos:

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)

#### (a) La estimación MC de β 0 , β 1 , β 2 utilizando los valores originales.ç

lmod10<-lm(y ~ x1 + x2)  
coef(lmod10)

## (Intercept) x1 x2   
## 6.4699828 6.5883362 0.2572899

#### (b) La estimación MC de β 0 , β 1 , β 2 utilizando los datos expresados en desviaciones respecto de la media.

ys <- scale(y, center=TRUE, scale=FALSE)  
x1s <- scale(x1, center=TRUE, scale=FALSE)  
x2s <- scale(x2, center=TRUE, scale=FALSE)  
lmod11 <- lm(ys ~ 0 + x1s + x2s)  
coef(lmod11)

## x1s x2s   
## 6.5883362 0.2572899

#### (c) La estimación insesgada de σ2

sg <- summary(lmod10)  
sgs <- summary(lmod11)  
c(sg$sigma^2, sgs$sigma^2)

## [1] 18.33002 15.27501

#### (d) El coeficiente de determinación

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

## [1] 0.9731074 0.9731074

#### (e) El coeficiente de determinación corregido

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

## [1] 0.9623503 0.9641432

#### (f) El contraste de la hipótesis nula H 0 : β0=β1=β2=0

g0 <- lm(y ~ 0)  
anova(g0,lmod10)

## 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 : β 1 = β 2 = 0 utilizando datos originales.

g1<-lm(y ~ 1)  
anova(g1,lmod10)

## 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

Se rechaza

#### (h) El contraste de la hipótesis nula H 0 : β 1 = β 2 = 0 utilizando datos en desviaciones respecto a la media.

g1s<-lm(ys ~ 1)  
anova(g1s, lmod11)

## 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

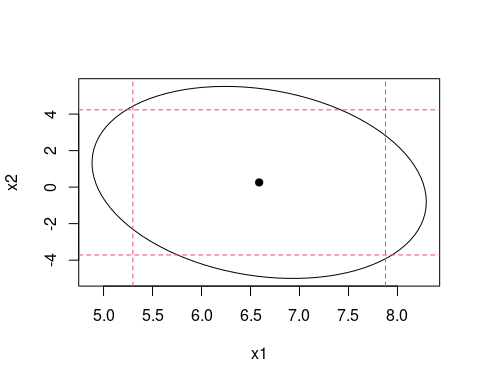
#### (i) La representación gráfica de una región de confianza del 95 % para β 1 y β 2 .

library(ellipse)

##   
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':  
##   
## pairs

plot(ellipse(lmod10,2:3),type="l")  
points(coef(lmod10)[2], coef(lmod10)[3], pch=19)  
abline(v=confint(lmod10)[2,],lty=2,col=2)  
abline(h=confint(lmod10)[3,],lty=2,col=2)



#### (j) El contraste individual de los parámetros β 0 , β 1 y β 2 .

summary(lmod10)

##   
## 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

Beta 1 es significativo

#### (k) El contraste de la hipótesis nula H0: β1 = 10 β2 .

lmod12<-lm(y ~ I(10\*x1 + x2))  
anova(lmod12,lmod10)

## 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

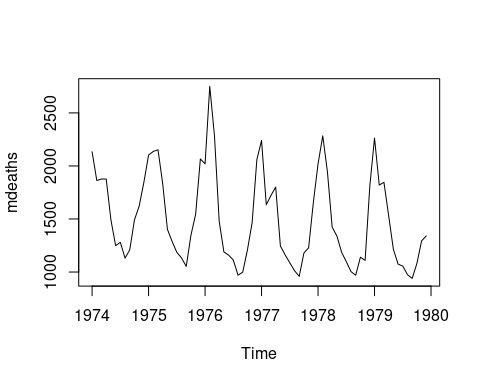
## Ejercicios del libro de Faraway

#### 4. (∗) (Ejercicio 4 cap. 4 pág. 57)

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)

 Hay más muertes a primeros de año

#### (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

#### 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

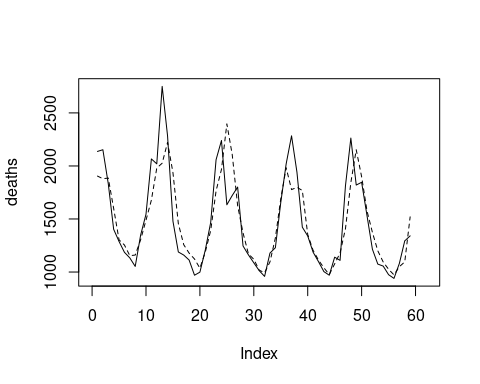
#### (d) Use your answer from the previous question to compute a prediction and interval for February 1980.

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?

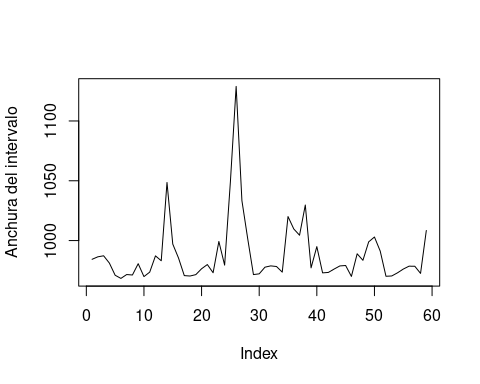
plot(lagdf$y, type="l", xlim=c(0,62), ylab="deaths")  
lines(predict(armod), lty=2)



pred.int <- predict(armod, interval = "prediction")

## Warning in predict.lm(armod, interval = "prediction"): predictions on current data refer to \_future\_ responses

plot(pred.int[,3]-pred.int[,2], type="l", ylab="Anchura del intervalo")



which.max(pred.int[,3]-pred.int[,2])

## 26   
## 26

#### 5. (∗) (Ejercicio 5 cap. 4 pág. 58)

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")  
  
lmod13<-lm(brozek ~ age + weight + height + neck + chest + abdom +  
 hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)  
  
lmod14<-lm(brozek ~ age + weight + height + abdom, data=fat)  
  
anova(lmod13, lmod14)

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

Es significativo, así que no podemos usar el modelo simple.

#### (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(lmod14, newdata = data.frame(age=medianas[1],weight=medianas[2],height=medianas[3],abdom=medianas[8]), interval="prediction")

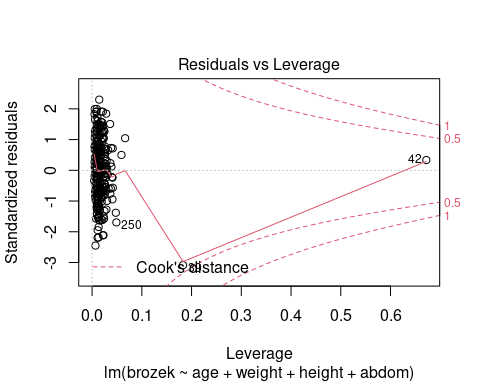
## fit lwr upr  
## age 17.84028 9.696631 25.98392

predict(lmod13, newdata = as.data.frame(t(medianas[c(1:3,6:15)])), interval="prediction")

## fit lwr upr  
## 1 17.49322 9.61783 25.36861

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

plot(lmod14, which=5)



#### (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?

lmod15 <- lm(brozek ~ age + weight + height + abdom, data=fat[-c(39,42),])  
medianas <- apply(fat[-c(39,42),4:18],2,median)  
  
predict(lmod15, 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