Regressió lineal simple i múltiple amb R

Disseny experimental i anàlisi de dades

Objectius de la sessió:

- Calcular i interpretar el coeficient de correlació lineal.
- Definir un model de regressió lineal simple i múltiple i interpretar els coeficients de regressió.
- Estudiar la significació global del model de regressió i la variabilitat explicada per cada variable independent o explicativa.
- Mesurar la bondat de l'ajust del model de regressió múltiple.

En el fitxer **colesterol.txt** tenim informació d'una mostra de 50 homes amb problemes amb el colesterol. Se'ls hi ha valorat:

- colesterol: Nivells de colesterol. (mg/dL)
- pes: Mesurat en kg
- cintura: Diàmetre de la cintura (cm)
- hemoglobina: Nivells d'hemoglobina (g/dL)

El nostre objectiu és estudiar de quines variables depén el colesterol.

1. Representa gràficament la relació entre colesterol i el pes, colesterol i cintura i finalment colesterol i hemoglobina

```
dades<-read.table("colesterol.txt",sep="\t",header=TRUE)
head(dades)
```

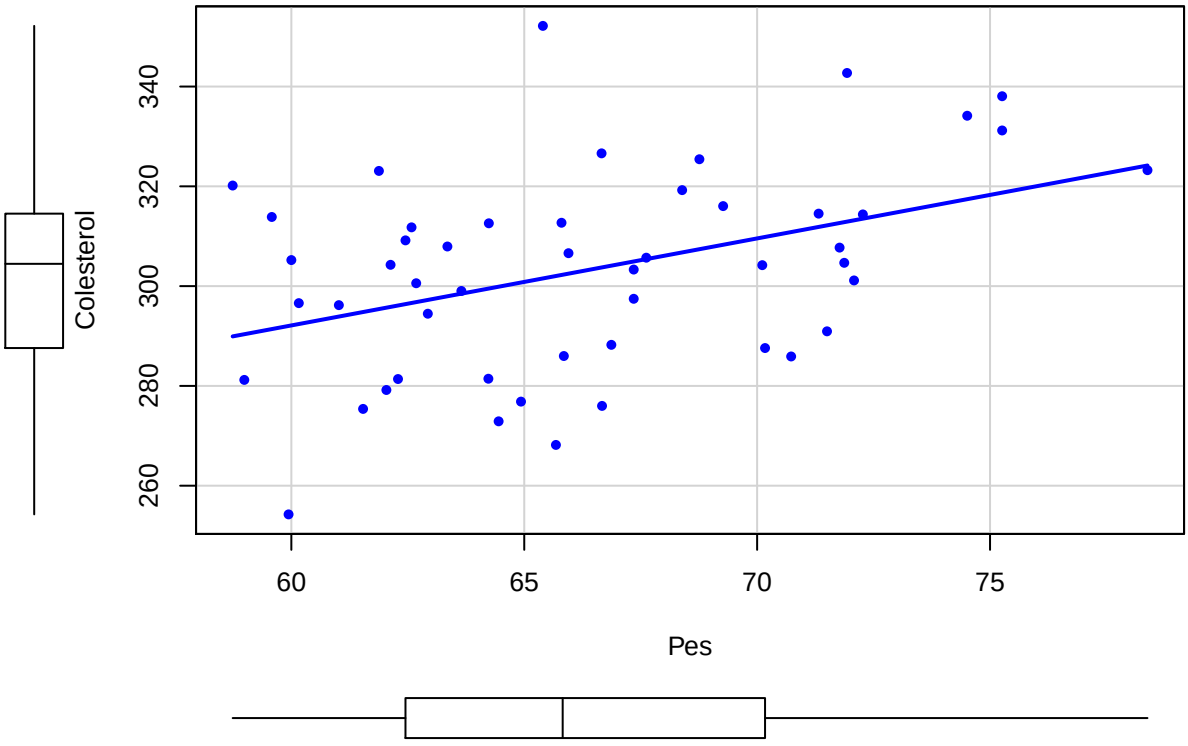
```
##   colesterol   pes cintura hemoglobina
## 1     276.85 64.93   85.17       14.31
## 2     301.14 72.08   83.32       14.19
## 3     305.70 67.62   85.58       14.08
```

```
## 4     287.59 70.17   81.37       13.05
## 5     304.67 71.87   84.61       11.96
## 6     290.94 71.50   84.55       14.54
```

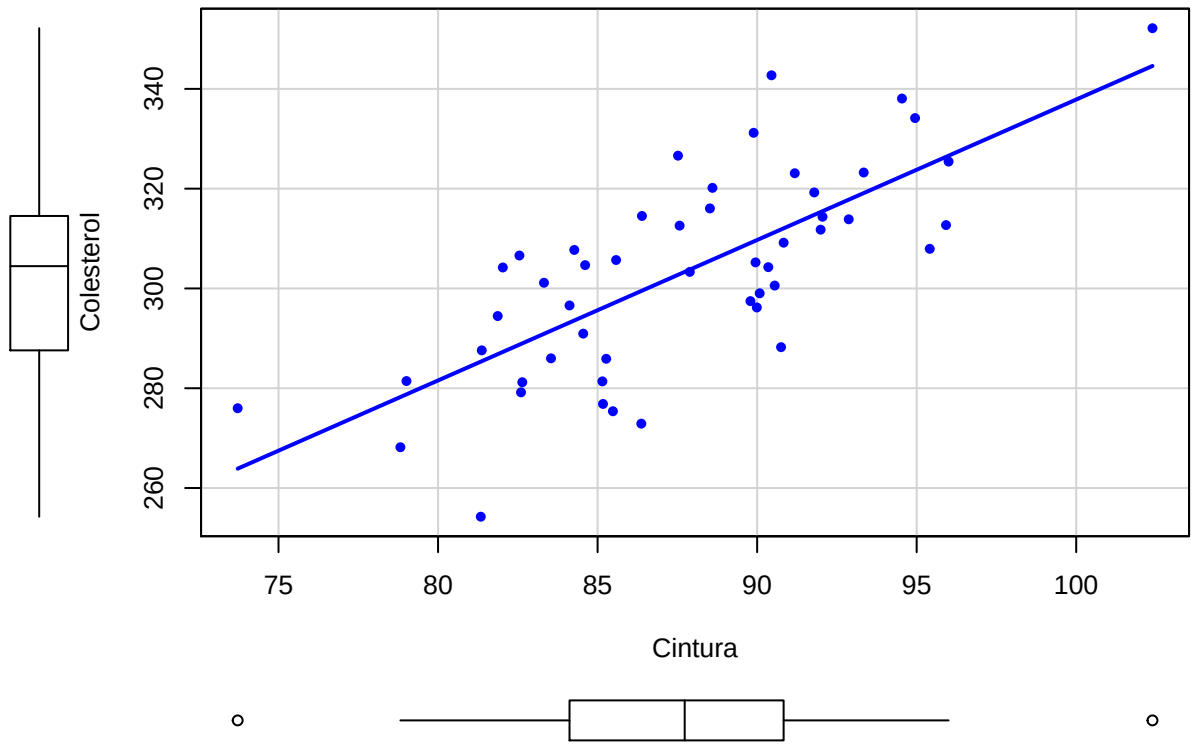
```r
library(car)
```

```
## Loading required package: carData
```

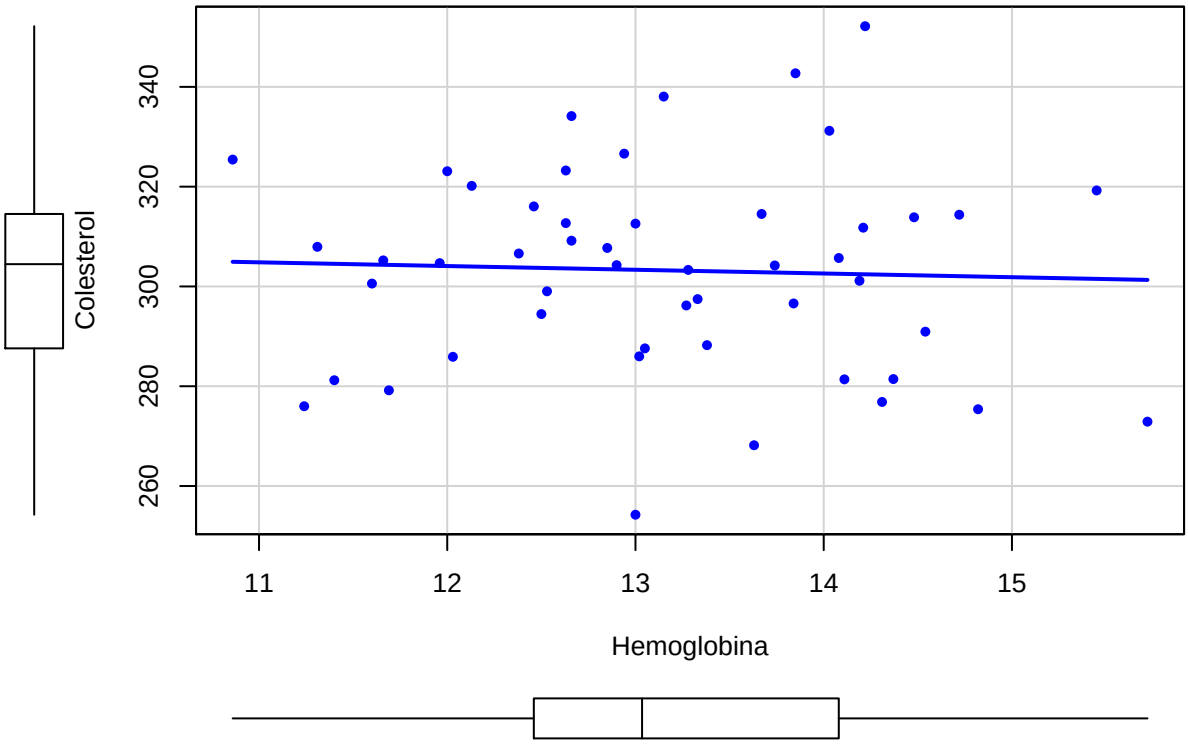
```r
scatterplot(colesterol ~ pes , regLine = TRUE,smooth=FALSE,
    data = dades,pch=16,col="blue",
    ylab="Colesterol", xlab=c("Pes"))
```




```r
scatterplot(colesterol ~ cintura , regLine = TRUE,smooth=FALSE,
    data = dades,pch=16,col="blue",
    ylab="Colesterol", xlab=c("Cintura"))
```




```
scatterplot(colesterol ~ hemoglobina , regLine = TRUE,smooth=FALSE,
    data = dades,pch=16,col="blue",
    ylab="Colesterol", xlab=c("Hemoglobina"))
```



2. Existeix relació entre el colesterol i el pes? i amb quina variable de les recollides (cintura i hemoglobina) també està associat el colesterol?

Per poder estudiar si existeix correlació, utilitzarem el coeficient de correlació i realitzarem el següent contrast:

$$Ho: \ Independència, \ \rightarrow \rho = 0$$

$$Ha: \ Dependència, \ \rightarrow \rho \neq 0$$

```
cor.test(dades$colesterol,dades$pes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dades$colesterol and dades$pes
## t = 3.1358, df = 48, p-value = 0.002924
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.1513665 0.6195774
## sample estimates:
##       cor
## 0.4123412
```

```
cor.test(dades$colesterol,dades$cintura)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dades$colesterol and dades$cintura
## t = 7.5159, df = 48, p-value = 1.196e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5745005 0.8414137
## sample estimates:
##       cor
## 0.7352705
```

```
cor.test(dades$colesterol,dades$hemoglobina)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dades$colesterol and dades$hemoglobina
## t = -0.28467, df = 48, p-value = 0.7771
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3157925  0.2400373
## sample estimates:
##         cor
## -0.04105335
```

Podem observar que tant el pes com el diàmetre de la cintura, rebutgem la hipòtesis nul.la amb un estadístic de contrast igual a 3.1358 (pvalor=0.0029) i 7.5159 (pvalor<0.001), respectivament. En canvi, l'hemoglobina

no és rebutja la hipòtesis nul.la per tant, no existeix una relació estadísticament significativa amb el colesterol.

3. Ajusta un model de regressió simple entre colesterol i les tres variables, interpreta els valors dels coeficients del model de regressió i calcula la taula ANOVA per cada model

PES:

```
reg.sim.pes<-lm(formula = colesterol ~ pes, data = dades)
summary(reg.sim.pes)
```

```
##
## Call:
## lm(formula = colesterol ~ pes, data = dades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.766 -15.587   0.698  12.705  50.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 187.4906    37.0054   5.067 6.42e-06 ***
## pes           1.7438     0.5561   3.136  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.78 on 48 degrees of freedom
## Multiple R-squared:   0.17,  Adjusted R-squared:  0.1527
## F-statistic: 9.833 on 1 and 48 DF,  p-value: 0.002924
```

L'estimació del coeficient de pes és igual a 1.7438 i la recta seria:

$$Y = a + b_1 \times pes$$

$$Y = 187.4906 + 1.7438 \times pes$$

El coeficient, a, és l'ordenada en l'origen, per tant, quan el pes és igual a 0, el colesterol és igual a 187.49.

(interpretació purament matemàtica, no té sentit biològic).

El coeficient de pes, s'interpreta per cada kilogram de pes el colesterol incrementa en promig 1.78 mg/dL.

En les dues últimes columnes, podem veure el contrast sobre els coeficients del model:

$$Ho : \beta_1 = 0$$

$$Ha : \beta_1 \neq 0$$

Tant per la constant, a, com la pendent β_1, rebutgem la hipòtesis nul.la, per tant existeix relació entre pes i colesterol.

```
anova(reg.sim.pes)
```

```
## Analysis of Variance Table
##
## Response: colesterol
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## pes        1  3467.3  3467.3  9.8331 0.002924 **
## Residuals 48 16925.5   352.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En relació a la taula aNOVA, observem que la suma de quadrats de la regressió és igual a 3467 i la dels residus 16925.5. Per tant, la suma de quadrats de la variable Y, és igual a 20392.8 i obtenim una $R^2 = \frac{3467.3}{20392.8} = 0.17$. Aquest valor també el podeu trobar en el summary del model de regressió lineal.

CINTURA

```
reg.sim.cintura<-lm(formula = colesterol ~ cintura, data = dades)
summary(reg.sim.cintura)
```

```
##
## Call:
## lm(formula = colesterol ~ cintura, data = dades)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -31.082 -10.388  -0.781  10.209  31.742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.3467    32.9076   1.712   0.0933 .
## cintura       2.8152     0.3746   7.516  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.97 on 48 degrees of freedom
## Multiple R-squared:  0.5406, Adjusted R-squared:  0.5311
## F-statistic: 56.49 on 1 and 48 DF,  p-value: 1.196e-09
```

L'estimació del coeficient de cintura és igual a 2.8152 i la recta seria:

$$Y = a + b_1 \times cintura$$

$$Y = 56.3467 + 2.8152 \times cintura$$

El coeficient, a, és l'ordenada en l'origen, per tant, quan el diàmetre de la cintura és igual a 0, el colesterol és igual a 56.3467. (interpretació purament matemàtica, no té sentit biològic)

El coeficient de cintura, s'interpreta per cada centímetre de cintura el colesterol incrementa en promig 2.8152 mg/dL.

En relació al contrast sobre els coeficients del model:

$$Ho : \beta_1 = 0$$

$$Ha : \beta_1 \neq 0$$

Tant per la constant, a, com la pendent β_1, rebutgem la hipòtesis nul.la, per tant existeix relació entre cintura i colesterol.

```r
anova(reg.sim.cintura)
```

```
## Analysis of Variance Table
##
## Response: colesterol
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cintura    1  11025 11024.8  56.489 1.196e-09 ***
## Residuals 48   9368   195.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En relació a la taula ANOVA, observem que la suma de quadrats de la regressió és igual a 11025 i dels residus 9368 Per tant, la suma de quadrats de la variable Y, és igual a 20392.8 i obtenim una $R^2 = \frac{11025}{20392.8} = 0.54$. Per tant, podem observar que en aquest cas la recta de regressió amb cintura presenta un millor ajust que la recta amb el pes.

HEMOGLOBINA

```r
reg.sim.hemoglobina<-lm(formula = colesterol ~ hemoglobina, data = dades)
summary(reg.sim.hemoglobina)
```

```
##
## Call:
## lm(formula = colesterol ~ hemoglobina, data = dades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.094 -15.495   0.865  12.142  49.726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 313.0367    34.5659   9.056 5.86e-12 ***
## hemoglobina  -0.7456     2.6193  -0.285    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 20.59 on 48 degrees of freedom
## Multiple R-squared:  0.001685,   Adjusted R-squared:  -0.01911
## F-statistic: 0.08103 on 1 and 48 DF,  p-value: 0.7771
```

L'estimació del coeficient de hemoglobina és igual a 2.8152 i la recta seria:

$$Y = a + b_1 \times hemoglobina$$

$$Y = 313.0367 - 0.7456 \times hemoglobina$$

El coeficient, a, és l'ordenada en l'origen, per tant, quan el diàmetre de la cintura és igual a 0, el colesterol és igual a 313.0367.

El coeficient d'hemoglobina, s'interpreta per cada g/dL de hemoglobina el colesterol disminueix en promig 0.7456 mg/dL

En relació al contrast sobre els coeficients del model:

$$Ho: \ \beta_1 = 0$$

$$Ha: \ \beta_1 \neq 0$$

En relació al contrast de la pendent β_1, no rebutgem la hipòtesis nul.la, per tant no existeix relació entre hemoglobina i colesterol.

```
anova(reg.sim.hemoglobina)
```

```
## Analysis of Variance Table
##
## Response: colesterol
##             Df  Sum Sq Mean Sq F value Pr(>F)
## hemoglobina  1    34.4   34.37   0.081 0.7771
## Residuals   48 20358.4  424.13
```

En relació a la taula ANOVA, observem que la suma de quadrats de la regressió és igual a 34.4 i dels residus 20358.4. Com es pot observar la suma de quadrats de la regressió presenta un valor baix, i això fa que el

$R^2 = \frac{34.4}{20392.8} = 0.001685$, també presenti un valor molt dolent.

4. Quina és la variable que explica un percentatge més alt de la variabilitat del colesterol

Es la que tingui un coeficient de determinació major. En aquest cas seria la variable cintura.

Recordar que el coeficient de determinació en un model de regressió lineal simple és igual al quadtrat del coeficient de correlació de Pearson i en el output de Rstudio seria el *Multiple R-squared*

5. Construeix un model de regressió múltiple del colesterol en funció del pes, diàmetre de cintura i els nivells d'hemoglobina.

El model que volem estimar és:

Model:

$$Y = a + b_1 \times pes + b_2 \times diametre + b_3 \times hemoglobina + e$$

Per estimar els coeficients del model amb Rstudio:

```
res.mul.1<-lm(formula = colesterol ~ pes+cintura+hemoglobina, data = dades)
summary(res.mul.1)
```

```
##
## Call:
## lm(formula = colesterol ~ pes + cintura + hemoglobina, data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4505  -7.6457  -0.3034   8.3758  25.3503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.4463    40.0290  -0.086 0.931764
## pes           1.4394     0.3720   3.870 0.000341 ***
## cintura       2.6615     0.3339   7.971 3.26e-10 ***
## hemoglobina  -1.6936     1.5858  -1.068 0.291097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.36 on 46 degrees of freedom
## Multiple R-squared:  0.6553, Adjusted R-squared:  0.6328
## F-statistic: 29.15 on 3 and 46 DF,  p-value: 1.034e-10
```

Segons l'estimació dels coeficients la recta seria:

$$Y = -3.4463 + 1.4394 \times pes + 2.6615 \times diametre - 1.6936 \times hemoglobina + e$$

6. En general, el model de regressió considerat explica part de la variabilitat del colesterol?

Prova de significació global: En el resultat del RStudio només ens dona directament l'estadístic F=29.15 i el p=1.034e-10. Aquest valor el podríem extreure de la taula ANOVA.

```
anova(res.mul.1)
```

```
## Analysis of Variance Table
##
## Response: colesterol
##             Df Sum Sq Mean Sq F value    Pr(>F)
## pes          1 3467.3  3467.3 22.6891 1.936e-05 ***
## cintura      1 9721.6  9721.6 63.6158 3.209e-10 ***
## hemoglobina  1  174.3   174.3  1.1406    0.2911
## Residuals   46 7029.6   152.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La suma de quadrats de la regressió és igual a 3467.3+9721.6+174.3=13363.2 i la suma de quadrats dels residus és igual a 7029.6. La variabilitat total de la Y era de 20392.8, per tant l'estadístic F és igual a

$$F = \frac{\frac{SQRegressió}{p}}{\frac{SQResidual}{n-p-1}} = \frac{\frac{13363.2}{3}}{\frac{7029.6.}{46}} = 29.15$$

Aquest resultat està en el summary. Podem veure que el pvalor associat és inferior al 0.001, per tant, podem afirmar que aquestes variables expliquen una part significativa de la variabilitat de Y.

7. Quin percentatge de la variabilitat de colesterol es explicat per aquest model?

En aquest cas donat que hi ha més d'una variable explicativa utilitzarem el Ajusted R-squared=0.6328. Per tant, el model de regressió explicat el 63.28% de la variabilitat del colesterol.

8. Tots els coeficients de regressió són estadísticament diferent de zero?

Els coeficients de cintura i pes són estadísticament diferents a zero, en canvi tal i com esperàvem no ho és el de l'hemoglobina.

9. Quina part de la variabilitat de colesterol explica cada variable quan estan la resta de variables en el model?

Per poder estudiar, haurem de calcular la variabilitat de cada variable calculada a partir d'ANOVA tipus II.

```
library(car)
Anova(res.mul.1,type=2)
```

```
## Anova Table (Type II tests)
##
## Response: colesterol
##             Sum Sq Df F value    Pr(>F)
## pes         2288.6  1 14.9763 0.0003415 ***
## cintura     9709.7  1 63.5379 3.263e-10 ***
## hemoglobina  174.3  1  1.1406 0.2910967
## Residuals   7029.6 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les sumes de quadrats que tenim en aquesta taula són:

- La suma de quadrats de pes en un model que prèviament hi ha hemoglobina i cintura:

$$SQ(pes|constant,\ cintura\ i\ hemoglobina) = 2288.6$$

- La suma de quadrats de cintura en un model que prèviament hi ha hemoglobina i pes:

$$SQ(cintura|constant,\ hemoglobina\ i\ pes) = 9709.7$$

- La suma de quadrats de hemoglobina en un model que prèviament hi ha cintura i pes:

$$SQ(hemoglobina|\ constant,\ cintura\ i\ pes) = 174.3$$

Tal i com es pot observar només l'hemoglobina no explica significativament part de la variabilitat de colesterol en un model que conté el diàmetre de la cintura i el pes. Les altres dues en canvi si que aporten informació al model.

10. Hi ha alguna variable que no faria falta tenir-la en compte? En cas afirmatiu elimina-la del model de regressió lineal i estima un altre cop el model. Si la variable hemoglobina.

```
res.mul.2<- lm(colesterol~cintura+pes, data=dades)
summary(res.mul.2)
```

```
##
## Call:
## lm(formula = colesterol ~ cintura + pes, data = dades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.127  -9.209   1.665   7.586  25.119
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.4348    35.9173  -0.625 0.535239
## cintura       2.6631     0.3344   7.964 2.88e-10 ***
## pes           1.3879     0.3694   3.758 0.000473 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.38 on 47 degrees of freedom
## Multiple R-squared:  0.6467, Adjusted R-squared:  0.6317
## F-statistic: 43.02 on 2 and 47 DF,  p-value: 2.4e-11
```

Podem comprovar que les estimacions dels coeficients de regressió de les variables cintura i pes han canviat

poc. També el coeficient de determinació, donat que hem eliminat una variable que no estava associada a colesterol.