Activitat 3: Modelització predictiva

Semestre 2019.2

Índice general

En aquesta activitat usarem el conjunt de dades sobre el pes en néixer que es preprocesó en l'activitat anterior.

Us proporcionem el fitxer BWprocessed.csv perquè tots treballeu amb el mateix fitxer de dades, independentment del resultat obtingut en l'activitat 1 i activitat 2. Recordem que el fitxer de dades conté el pes i grandària dels nens i nenes nascuts a Espanya durant l'any 2019. L'arxiu conté 300 registres i 11 variables. Aquestes variables són: ID, HP, City, Time, Day, BW, BPD, AD, Sex, Ge y Sm.

En aquesta activitat comencem aplicant models de regressió lineal i posteriorment, s'aplicaran models de regressió logística binària.

---

1 Model de regressió lineal

1.1 Model de regressió lineal univariant

a) Estimar per mínims quadrats ordinaris un model lineal que expliqui la variable pes del bebè en néixer en funció del diàmetre abdominal abans de néixer.

S'avaluarà la bondat de l'ajust, a partir del coeficient de determinació.

```
#Estimacion del modelo

attach (datA3)
Model.1.1<- lm(BW~AD, data=datA3 )
summary(Model.1.1)
```

```
##
## Call:
## lm(formula = BW ~ AD, data = datA3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1012.90  -215.20   -30.48   165.55  1508.64
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3105.728    185.848  -16.71   <2e-16 ***
## AD             57.113      1.829   31.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.7 on 298 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.7652
## F-statistic: 975.5 on 1 and 298 DF,  p-value: < 2.2e-16
```

```
cor(x = BW, y = AD, method = "pearson")
```

```
## [1] 0.8752187
```

A la vista dels resultats, hi ha una relació lineal positiva alta, entre les dues variables. S'observa que el coeficient de determinació ajustat és: 0.765. Si es calcula el coeficient de correlació obtenim un valor de 0.875.

b) Posteriorment, es dividirà la mostra en dues, segons el sexe del bebè i es repetirà l'estudi per a cada mostra per separat.Raonar els resultats.

NOTA: En tenir un model amb una sola variable, es podrà prendre el coeficient de determinació sense ajustar, ja que el seu valor no s'altera.

```
#Estimacion del modelo

selected_Female <- which(datA3$Sex=="F" )
data1=datA3[selected_Female,]
selected_Male <- which(datA3$Sex=="M" )
data0=datA3[selected_Male,]
dim(data1)
```

```
## [1] 168  11
```

```
dim(data0)
```

```
## [1] 132  11
```

```
Model.1.1.1<- lm(BW~AD, data=data1 )
summary(Model.1.1.1)
```

```
##
## Call:
## lm(formula = BW ~ AD, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -999.82 -245.31  -42.18  155.75 1522.40
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3103.114    256.213  -12.11   <2e-16 ***
## AD             56.946      2.499   22.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 385.7 on 166 degrees of freedom
## Multiple R-squared:  0.7577, Adjusted R-squared:  0.7563
## F-statistic: 519.2 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
Model.1.1.2<- lm(BW~AD, data=data0)
summary(Model.1.1.2)
```

```
##
## Call:
## lm(formula = BW ~ AD, data = data0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1028.14  -223.10    -8.69   152.41  1491.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3139.756    271.672  -11.56   <2e-16 ***
## AD             57.637      2.703   21.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 347.8 on 130 degrees of freedom
## Multiple R-squared:  0.7776, Adjusted R-squared:  0.7759
## F-statistic: 454.7 on 1 and 130 DF,  p-value: < 2.2e-16
```

A la vista dels resultats, com es podria esperar, podem concloure que el sexe del nadó no varia la relació lineal entre les dues variables. Els coeficients de determinació són de 0.756 i 0.776.

1.2 Model de regressió lineal múltiple (regresors quantitatius)

a) Estimar per mínims quadrats ordinaris un model lineal que expliqui la variable pes del bebè en funció del diàmetre abdominal i el diàmetre biparietal. Es procedirà a avaluar la bondat d'ajust a través del coeficient de determinació ajustat. Discutir si es produeix una millora del model.

```
# Ajustamos el modelo de regresión múltiple:
Model.1.2<- lm(BW~AD+BPD, data= datA3)
summary( Model.1.2)
```

```
##
## Call:
## lm(formula = BW ~ AD + BPD, data = datA3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -714.53 -225.87  -52.55  164.57 1260.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4767.601    279.220 -17.075  < 2e-16 ***
```

```
## AD             43.401      2.479  17.510  < 2e-16 ***
## BPD            34.250      4.555   7.519 6.57e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 338.5 on 297 degrees of freedom
## Multiple R-squared:  0.8034, Adjusted R-squared:  0.8021
## F-statistic: 606.9 on 2 and 297 DF,  p-value: < 2.2e-16
```

En el model 1.1.a, el coeficient de bondat d'ajust és de 0.765, i en aquest últim de 0.802, per la qual cosa el valor és molt semblant. No hi ha evidència de millora del model.

b) Estudiar l'existència o no de multicolinealidad entre les covariables del model anterior, AD i BPD. Podeu usar la llibreria (faraway).

```
# Se calculará el coeficiente de correlación entre las variables AD y BPD
cor(x = AD, y = BPD, method = "pearson")
```

```
## [1] 0.735707
```

```
# Vegem com difereixen les estimacions de el model global amb 'AD' i 'BPD', dels models de regressió li
model.AD <- lm(BW~AD, data=datA3 )
model.BPD<- lm(BW~BPD, data=datA3 )
summary(model.AD)
```

```
##
## Call:
## lm(formula = BW ~ AD, data = datA3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1012.90  -215.20   -30.48   165.55  1508.64
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3105.728    185.848  -16.71   <2e-16 ***
## AD             57.113      1.829   31.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.7 on 298 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.7652
## F-statistic: 975.5 on 1 and 298 DF,  p-value: < 2.2e-16
```

```
summary(model.BPD)
```

```
##
## Call:
## lm(formula = BW ~ BPD, data = datA3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1138.13  -344.84   -22.13   223.71  1626.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5604.587    391.517  -14.31   <2e-16 ***
```

```
## BPD            92.929      4.391   21.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 481.7 on 298 degrees of freedom
## Multiple R-squared:  0.6005, Adjusted R-squared:  0.5992
## F-statistic: 447.9 on 1 and 298 DF,  p-value: < 2.2e-16
```

```
# Cargamos la librería faraway
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.6.3
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
# Cálculo de FIV
vif(Model.1.2)
```

```
##       AD      BPD
## 2.179907 2.179907
```

```
# Se compara con 1/(1-R2)
1/(1-summary(Model.1.2)$r.squared)
```

```
## [1] 5.087178
```

D'una banda el coeficient de correlació entre les dues variables és de 0.7, de manera que la relació lineal entre les dues variables és alta. D'altra banda, si comparem el model global, amb cada un dels models simples, els coeficients estimats per a AD i BPD difereixen dels estimats amb la regressió múltiple, especialment pel diàmetre biparietal (BPD). Això pot ser indicatiu d'un problema de colinealitat.

Es procedirà a detectar possibles efectes de multicolinealitat. Ja que un dels efectes principals de la multicolinealitat és la inflació de la variància i covariància de les estimacions, es calcularà el FIV (factor d'inflació de la variància). A la vista dels últims resultats, no es troba indicis de multicolinealitat entre els regressors 'AD' i 'BPD', respecte als criteris de diagnòstic proposats. El FIV = 2,179 resulta menor que el seu equivalent en el model global, $1/(1-R^2) = 5{,}087$. A mé el valor de FIV, és molt baix.

NOTA: Generalment, valors d'un FIV superiors a 10 donen indicis d'un problema de multicolinealitat, si bé la seva magnitud depèn del model ajustat. (Altres autors consideren valors per sobre de 4). És millor comparar-lo amb el seu equivalent en el model ajustat, és a dir, $1/(1-R^2)$, on R^2 és el coeficient de determinació del model. Els valors FIV superior a aquesta quantitat impliquen que la relació entre les variables explicatives és més gran que la que hi ha entre la resposta i els predictors, i per tant donen indicis de multicolinealitat.

1.3 Model de regressió lineal múltiple (regresors quantitatius i qualitatius)

a) Volem conèixer en quina mesura es relaciona el pes, en funció del diàmetre abdominal, diàmetre biparietal i les setmanes de gestació. Es recodificará la variable Ge,en menor i major o igual de 35 setmanes. Aplicar un model de regressió lineal múltiple i explicar el resultat.

```r
#Estimacion del modelo múltiple
low.Ge <- (datA3$Ge < 35)
Ge_RE <- ifelse(low.Ge==TRUE, 1, 0)
table(Ge_RE)
```

```
## Ge_RE
##   0   1
## 161 139
```

```r
Model.1.3.a = lm(BW~AD+BPD+Ge_RE, data=datA3 )
summary(Model.1.3.a)
```

```
##
## Call:
## lm(formula = BW ~ AD + BPD + Ge_RE, data = datA3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -791.57 -195.67   -3.27  152.75 1262.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3175.879    325.887  -9.745  < 2e-16 ***
## AD             32.557      2.651  12.280  < 2e-16 ***
## BPD            30.754      4.177   7.362 1.80e-12 ***
## Ge_RE        -400.994     51.248  -7.825 9.08e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 308.6 on 296 degrees of freedom
## Multiple R-squared:  0.8371, Adjusted R-squared:  0.8355
## F-statistic: 507.1 on 3 and 296 DF,  p-value: < 2.2e-16
```

Es pot observar que les tres variables explicatives són significatives al model. D'altra banda, s'obté que el coeficient de determinació ajustat és 0,835, de manera que l'ajust és bo.

Recta de regressió: y=-3175.879 + 32.55* AD + 30.75* BPD - 400.99* Ge_RE

b) Ara es calcularà el model de regressió lineal que relacioni el pes amb diàmetre abdominal i diàmetre biparietal, per a la mostra els bebès de la qual han nascut abans de les 35 setmanes. Posteriorment, es calcularà el mateix model, per als bebès nascuts en la setmana 35 i posteriors. En vista als resultats obtinguts, existeix relació amb l'apartat a)? Raonar la resposta.

```r
#Se divide la muestra
lower.Ge <- (datA3$Ge < 35)
Ge_RE_2 <- ifelse(lower.Ge==TRUE, "Si", "No")
table(Ge_RE_2)
```

```
## Ge_RE_2
##  No  Si
## 161 139
```

```r
selected_lower.Ge <- which(Ge_RE_2=="Si" )
data_lower.Ge=datA3[selected_lower.Ge,]
dim(data_lower.Ge)
```

```
## [1] 139  11
```

```r
selected_high.Ge <- which(Ge_RE_2=="No" )
data_high.Ge=datA3[selected_high.Ge,]
dim(data_high.Ge)
```

```
## [1] 161  11
```

```r
#Estimacion del modelo
Model.1.3.1.b = lm(BW~AD+BPD, data=data_lower.Ge )
summary(Model.1.3.1.b)
```

```
##
## Call:
## lm(formula = BW ~ AD + BPD, data = data_lower.Ge)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -594.6 -117.5   -8.0  146.7  419.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3037.307    219.603 -13.831  < 2e-16 ***
## AD             29.115      2.883  10.100  < 2e-16 ***
## BPD            28.133      3.624   7.764 1.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.8 on 136 degrees of freedom
## Multiple R-squared:  0.8067, Adjusted R-squared:  0.8038
## F-statistic: 283.7 on 2 and 136 DF,  p-value: < 2.2e-16
```

```r
Model.1.3.2.b = lm(BW~AD+BPD, data=data_high.Ge )
summary(Model.1.3.2.b)
```

```
##
## Call:
## lm(formula = BW ~ AD + BPD, data = data_high.Ge)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -783.04 -234.92  -56.89  199.48 1185.13
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5392.809    861.680  -6.258 3.50e-09 ***
## AD             34.716      3.971   8.742 3.19e-15 ***
## BPD            52.223      9.855   5.299 3.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368.3 on 158 degrees of freedom
## Multiple R-squared:  0.4958, Adjusted R-squared:  0.4894
## F-statistic: 77.68 on 2 and 158 DF,  p-value: < 2.2e-16
```

Amb l'estudi proposat hem obtingut els següents coeficients de determinació ajustats:

- setmanes de gestació menor de 35 $R^2 = 0{,}803$

- setmanes de gestació igual o major de 35 $R^2 = 0{,}489$

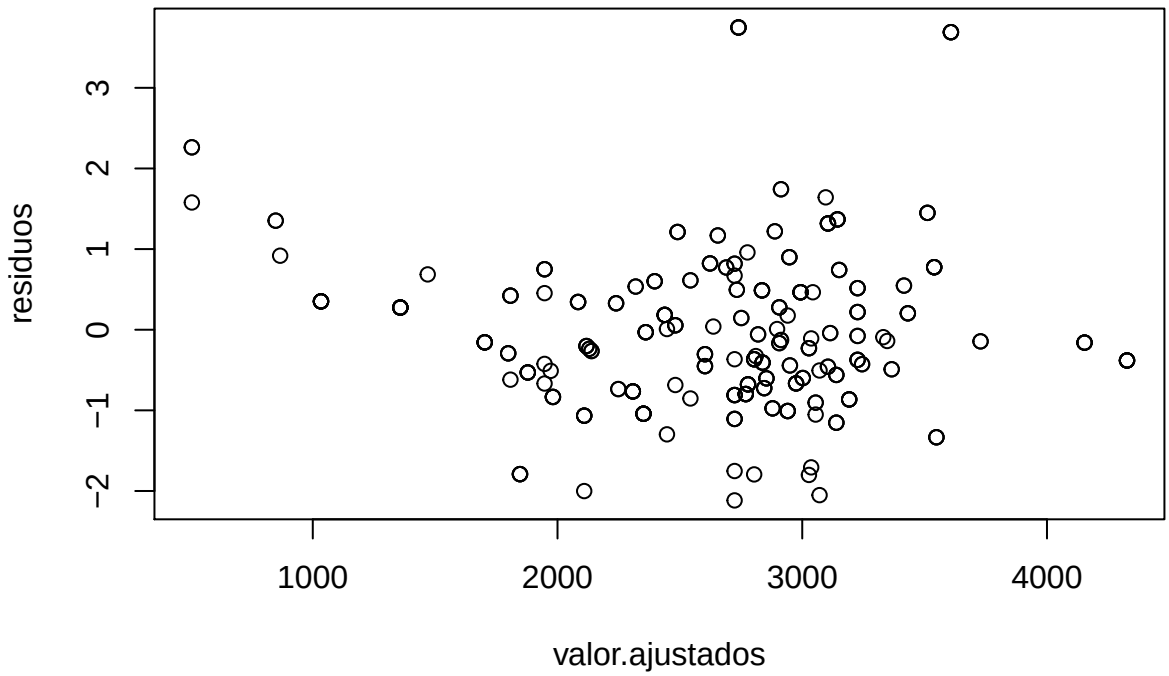
Comparant els coeficients de determinació ajustat per AD i BPD, s'aprecien diferències entre els nadons nascuts abans de la setmana 35 i després de la mateixa. En vista als resultats, escolliria el model de l'apartat a), per predir els resultats.

1.4 Diagnosi del model

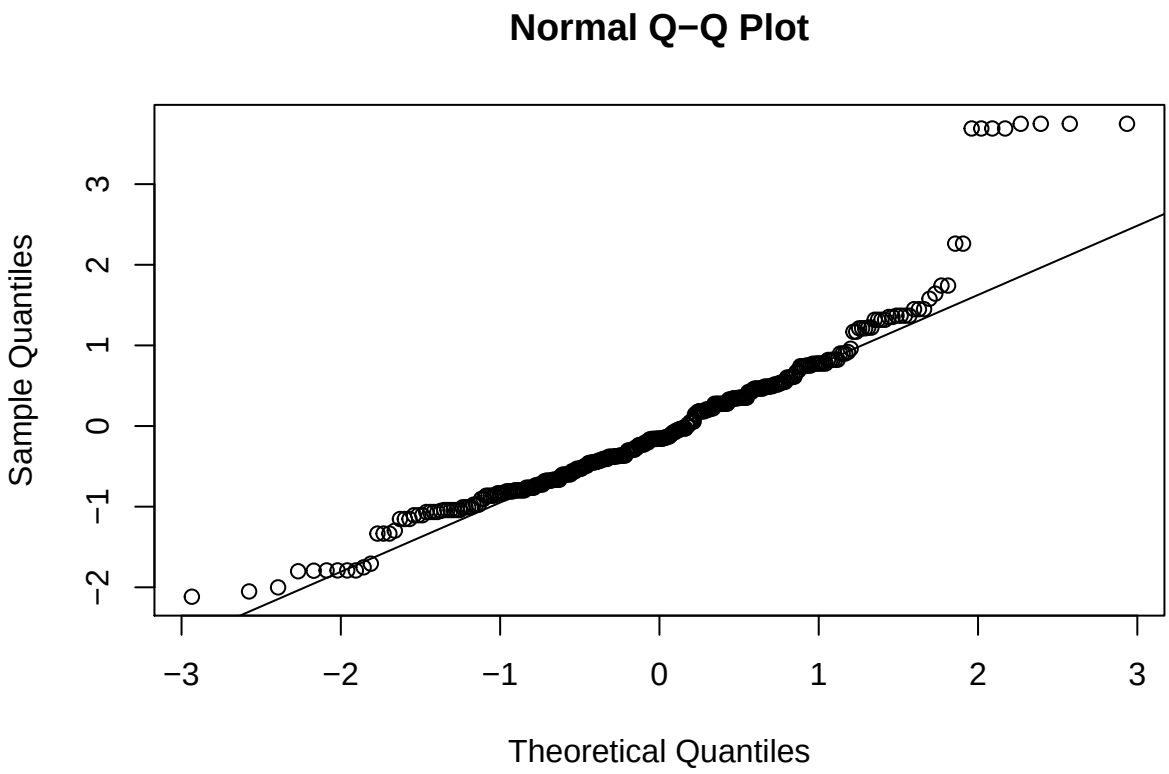
Es prendrà el model de l'apartat 1.2, que relaciona el pes del bebè en funció del diàmetre abdominal i el diàmetre biparietal.

Per a la diagnosi d'aquest model es faran dos gràfics: un amb els valors ajustats enfront dels residus (que ens permetrà veure si la variància és constant) i el gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment(QQ plot). Explicar conclusions a partir dels gràfics obtinguts.

```
residuos <- rstandard(Model.1.2)
valor.ajustados <- fitted(Model.1.2)
plot(valor.ajustados, residuos)
```




```
qqnorm(residuos)
qqline(residuos)
```



A la vista del gràfic, no s'observa cap patró especial, de manera que tant la homocedasticitat com la linealitat resulten hipòtesis raonables.

D'altra banda, el Q_Q plot mostra que les dades no s'ajusten bé a una normal.

1.5 Predicció del model

Segons el model de l'apartat 1.3, calcular el pes d'un bebè amb AD de 94 i BPD de 75, nascut en la setmana 34 de gestació.

```
newdata = data.frame(AD = 94, BPD=75, Ge_RE=1)
predict(Model.1.3.a, newdata)
```

```
##        1
## 1789.974
```

S'obté un valor de 1789.97, de manera que estem davant d'un cas d'un pes molt baix, per a un nadó.

2 Model de regressió logística

2.1 Estimació de OR (Odds Ràtio)

Es vol estudiar la probabilitat de tenir un bebè amb baix pes (aquells que són inferiors a 2.5kg).

Per a avaluar-la, primer es realitzarà una anàlisi crua dels possibles factors de risc.

a) Estudiar la relació entre el baix pes i cadascuna de les variables següents:

- a.1) si la mare és fumadora (Sm)

- a.2) sexe
- a.3) setmanes de gestació (Ge), (Es dividirà entre menor i major o igual de 35 setmanes)

Es demana:

Estimar i interpretar les OR en cada cas. Per a comprovar si existeix associació entre la variable dependent i cadascuna de les variables explicatives, s'aplicarà el test Chi-quadrat de Pearson. Un resultat significatiu ens dirà que existeix associació. Posteriorment, per a conèixer el grau d'aquesta associació, es calcularà les OR.

NOTA: Per al càlcul de les OR, es partirà de la taula de contingència i es calcularà a partir de la seva fórmula.

Selecció dels casos amb baix pes al neixer

```
lower.BW <- (datA3$BW < 2500)
table(lower.BW)
```

```
## lower.BW
## FALSE  TRUE
##   193   107
```

Recodifica la variable BW

```
BW_RE <- ifelse(lower.BW==TRUE, 1, 0)
table(BW_RE)
```

```
## BW_RE
##   0   1
## 193 107
```

a.1) Baix pes i fumar

```
table(datA3$Sm)
```

```
##
##   N   S
## 238  62
```

```
fumar_RE<-recode(datA3$Sm,'"S"=1;"N"=0')
table(fumar_RE)
```

```
## fumar_RE
##   0   1
## 238  62
```

```
fumar.tab = table(BW_RE,fumar_RE)
fumar.tab
```

```
##      fumar_RE
## BW_RE   0   1
##     0 191   2
##     1  47  60
```

```
chi.test<-chisq.test(fumar.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  fumar.tab
```

```
## X-squared = 123.85, df = 1, p-value < 2.2e-16
```

```
fisher.test(fumar.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fumar.tab
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    29.86902 1033.61939
## sample estimates:
## odds ratio
##   119.5901
```

```
# Cáculo de las OR
fumar.tab_M<- addmargins(fumar.tab, FUN = list(Total = sum), quiet = TRUE)
fumar.tab_M
```

```
##        fumar_RE
## BW_RE     0   1 Total
##   0     191   2   193
##   1      47  60   107
##   Total 238  62   300
```

```
p1_f=((fumar.tab_M[2,2]/fumar.tab_M[3,2])/(1-(fumar.tab_M[2,2]/fumar.tab_M[3,2])))
p2_f=((fumar.tab_M[2,1]/fumar.tab_M[3,1])/(1-(fumar.tab_M[2,1]/fumar.tab_M[3,1])))
OR_f=p1_f/p2_f
OR_f
```

```
## [1] 121.9149
```

a.2) Baix pes i sexe

```
table(datA3$Sex)
```

```
##
##   F   M
## 168 132
```

```
Sex_RE<-recode(datA3$Sex,'"F"=1;"M"=0')
table(Sex_RE)
```

```
## Sex_RE
##   0   1
## 132 168
```

```
Sex.tab = table(BW_RE,Sex_RE)
chi.test<-chisq.test(Sex.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Sex.tab
## X-squared = 0.68959, df = 1, p-value = 0.4063
```

```r
fisher.test(Sex.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Sex.tab
## p-value = 0.3955
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.480573 1.314131
## sample estimates:
## odds ratio
##  0.7947397
```

```r
# Cáculo de las OR. No procede
```

a.3) Baix pes i setmanes de gestació

```r
lower.Ge <- (datA3$Ge < 35)
table(lower.Ge)
```

```
## lower.Ge
## FALSE  TRUE
##   161   139
```

```r
Ge_RE <- ifelse(lower.Ge==TRUE, 1, 0)
table(Ge_RE)
```

```
## Ge_RE
##   0   1
## 161 139
```

```r
Ge.tab = table(BW_RE,Ge_RE)
chi.test<-chisq.test(Ge.tab)
print(chi.test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Ge.tab
## X-squared = 157.51, df = 1, p-value < 2.2e-16
```

```r
fisher.test(Ge.tab,simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Ge.tab
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   31.81035 281.26236
## sample estimates:
## odds ratio
##    84.2438
```

```r
# Cáculo de las OR
Ge.tab_M<- addmargins(Ge.tab, FUN = list(Total = sum), quiet = TRUE)
```

```
Ge.tab_M
```

```
##        Ge_RE
## BW_RE     0   1 Total
##   0     156  37   193
##   1       5 102   107
##   Total 161 139   300
```

```
p1_Ge=((Ge.tab_M[2,2]/Ge.tab_M[3,2])/(1-(Ge.tab_M[2,2]/Ge.tab_M[3,2])))
p2_Ge=((Ge.tab_M[2,1]/Ge.tab_M[3,1])/(1-(Ge.tab_M[2,1]/Ge.tab_M[3,1])))
OR_Ge=p1_Ge/p2_Ge
OR_Ge
```

```
## [1] 86.01081
```

Tant per fumar i setmanes de gestació, s'obté un p-valor inferior a 0.05, de manera que podem concloure que hi ha relació entre la variable baix pes i aquests factors. Per a la variable sexe el p-valor és de 0.4, de manera que el pes d'un nen al néixer és independent del sexe.

Passem a calcular les OR, per aquelles variables significatives:

-Fumar: OR de 121,9 per la qual cosa ens indica que una dona fumadora, té una probabilitat 122 vegades més gran de tenir un nadó amb baix pes.

-Setmanes de gestació: OR de 86,01, de manera que donar a llum en les setmanes 33 i 34, comporta que es tingui 86 vegades més possibilitats de tenir un nadó amb baix pes.

b) Si no s'hi hagués recodificado la variable setmanes de gestació, podríem seguir el procediment anterior per al càlcul de la OR?. Explicar la resposta.

No podem seguir el procediment anterior per al càlcul de l'OR ja que és una variable continua. En aquest cas per calcular les OR, hauríem construir un model de regressió logística.

c) Si volem veure la relació entre baix pes i lloc de procedència, podríem seguir el procediment anterior per al càlcul de la OR? En el cas que la resposta sigui negativa, quina seria una solució?.

No podem seguir el procediment anterior per al càlcul de l'OR ja que ciutat és una variable amb més de dues categories. En aquest cas per calcular les OR, hauríem construir un model de regressió logística.

2.2 Model de regressió logística

a) Estimar el model de regressió logística prenent com a variable dependent, tenir baix pes en néixer o no i sent la variable explicativa, fumar o no. Podem considerar que el fet de fumar és un factor de risc de baix pes? Justifica la teva resposta. Té relació amb l'obtingut en l'apartat anterior?

```
logit_model_1 <- glm(formula=BW_RE~fumar_RE, data=datA3, family=binomial)
summary(logit_model_1)
```

```
##
## Call:
## glm(formula = BW_RE ~ fumar_RE, family = binomial, data = datA3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6207  -0.6633  -0.6633   0.2561   1.8012
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4021     0.1628  -8.611  < 2e-16 ***
```

```
## fumar_RE1     4.8033     0.7370   6.517 7.16e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 390.89  on 299  degrees of freedom
## Residual deviance: 254.19  on 298  degrees of freedom
## AIC: 258.19
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coefficients(logit_model_1))
```

```
## (Intercept)   fumar_RE1
##   0.2460733 121.9148936
```

Com podem observar, el valor de l'OR coincideix amb l'apartat anterior, ja que en el model hem introduït només la mateixa variable explicativa. La interpretació dels resultats és la mateixa.

b) Afegim al model anterior les variable contínua diàmetre abdominal (AD). S'observa una millora del model? Explicar. Realitzeu també el càlcul de les OR i el seu interval de confiança.

```
logit_model_2 <- glm(formula=BW_RE~fumar_RE+AD, data=datA3, family=binomial)
summary(logit_model_2)
```

```
##
## Call:
## glm(formula = BW_RE ~ fumar_RE + AD, family = binomial, data = datA3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.37085  -0.15867  -0.04501   0.00807   1.99501
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 48.67005    7.31254   6.656 2.82e-11 ***
## fumar_RE1    5.60171    1.19523   4.687 2.78e-06 ***
## AD          -0.50513    0.07516  -6.721 1.80e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 390.89  on 299  degrees of freedom
## Residual deviance: 102.03  on 297  degrees of freedom
## AIC: 108.03
##
## Number of Fisher Scoring iterations: 8
```

```
exp(coefficients(logit_model_2))
```

```
##  (Intercept)    fumar_RE1           AD
## 1.371299e+21 2.708884e+02 6.034266e-01
```

```
exp(confint(logit_model_2))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %       97.5 %
## (Intercept) 4.054746e+15 1.530786e+28
## fumar_RE1   3.180081e+01 3.631726e+03
## AD          5.106254e-01 6.876337e-01
```

Basant-se l'indicador AIC, s'observa que és més petit que en el model anterior, de manera que hi ha una millora en l'ajust. Tant fumar, com AD, són factors significatius per a baix pes. Com el coeficient per a AD és negatiu, indica que a mesura que disminueix el diàmetre abdominal, augmenta el risc de baix pes.

Si es calculen les OR ajustades es té: La OR d'AD ajustada per fumar és de 0,60, de manera que per cada unitat que augmenti el diàmetre abdominal, el odds de tenir un nadó amb baix pes és 0,60 vegades menor. La OR de fumar ajustada per AD és de 270,8 pel que ens indica que fumar suposa 270 vegades més risc de tenir un nadó amb baix pes, que no tenir-lo.

El valor tan diferent entre l'OR estimada de fumar amb el primer model i el model ajustat, podria indicar-nos que AD, podria ser una variable de confusió.

c) Ara afegim al model anterior les variable City. Es prendrà com a ciutat de referència Barcelona. Càlcul de les OR. S'observa una millora del model? Explicar.

```
City_Rel=relevel(datA3$City, ref = 'Barcelona')
logit_model_3 <- glm(formula=BW_RE~fumar_RE+factor(City_Rel), data=datA3, family=binomial)
summary(logit_model_3)
```

```
##
## Call:
## glm(formula = BW_RE ~ fumar_RE + factor(City_Rel), family = binomial,
##     data = datA3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6683  -0.6780  -0.6467   0.2434   1.9261
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.353238   0.243124  -5.566 2.61e-08 ***
## fumar_RE1                  4.884267   0.744500   6.560 5.36e-11 ***
## factor(City_Rel)Granada   -0.105337   0.543199  -0.194    0.846
## factor(City_Rel)Madrid    -0.331582   0.442521  -0.749    0.454
## factor(City_Rel)Pamplona  -0.103511   0.584736  -0.177    0.859
## factor(City_Rel)Sevilla    0.548309   0.645776   0.849    0.396
## factor(City_Rel)Valencia   0.008722   0.540947   0.016    0.987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 390.89  on 299  degrees of freedom
## Residual deviance: 252.59  on 293  degrees of freedom
## AIC: 266.59
##
## Number of Fisher Scoring iterations: 6
```

La variable City, no és significativa, pel que no procedirem al càlcul de les ORs. No millora el model.

2.3 Predicció

Segons el model de l'apartat 2.2 b), quina seria la probabilitat de baix pes en néixer, si la mare és fumadora i AD és de 90?

```
pred<-predict(logit_model_2, data.frame(fumar_RE="1",AD=90),type = "response")
pred
```

```
##         1
## 0.9998508
```

El model de l'apartat anterior ens prediu una probabilitat de baix pes de 0.99, és a dir, pràcticament la probabilitat és total.

2.4 Bondat de l'ajust

Usa el test de Hosman-Lemeshow per a veure la bondat d'ajust del model final triat. En la llibreria (ResourceSelection) hi ha una funció que ajusta el test de Hosmer-Lemeshow.

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.6.3
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(BW_RE,fitted(logit_model_2))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  BW_RE, fitted(logit_model_2)
## X-squared = 5.5917, df = 8, p-value = 0.6929
```

La probabilitat és de 0,6929, el que indicaria que el model 2, s'ajusta correctament a les dades.

2.5 Corba ROC

Dibuixar la corba ROC, i calcular l'àrea sota la corba. Discutir el resultat.

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.6.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

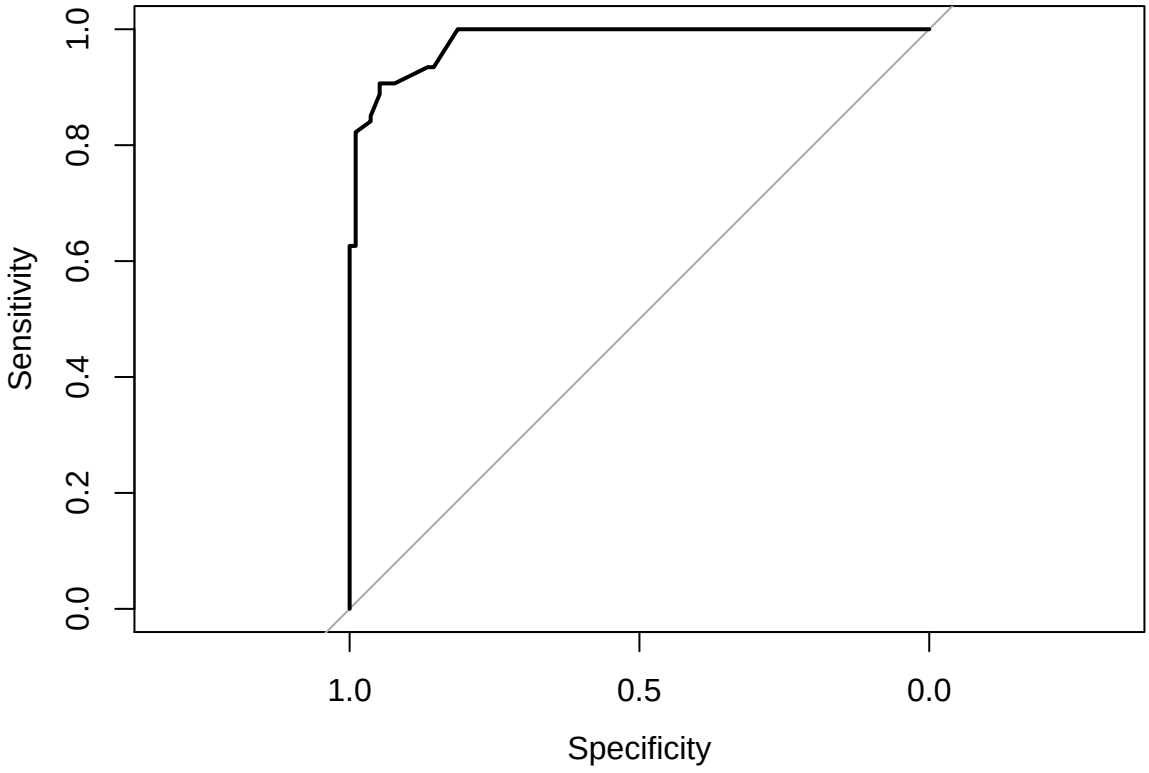
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
prob_low=predict(logit_model_2, datA3, type="response")
r=roc(BW_RE,prob_low, data=datA3)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(r)
```




```
auc(r)
```

```
## Area under the curve: 0.9808
```

L'àrea per sota d'aquesta corba pren el valor de 0,98, de manera que l'habilitat de el model per a discriminar entre aquells nadons que presenten baix pes enfront dels que no, és molt bona.

3 Conclusions de l'anàlisi

En aquest apartat s'hauran d'exposar les conclusions sobre la base dels resultats obtinguts en tot l'estudi. Regressió lineal i logística.

A la primera part d'aquest estudi, s'han estudiat les possibles associacions lineals entre la variable pes d'un nadó i diferents covariables. Segons els resultats obtinguts, podem concloure que hi ha una relació lineal positiva alta entre aquesta variable i el diàmetre abdominal (AD), diàmetre biparietal (BPD) i setmanes de gestació (GE), sense trobar-relació significativa entre el pes del nadó i el seu sexe. D'altra banda, donada l'associació lineal entre les covariables AD i BPD, s'ha procedit a estudiar la col · linealitat entre les dues, sense haver-se trobat evidències significatives. Podria ser degut a un problema de la mostra obtinguda.

Un cop ajustat el model de regressió lineal, s'ha obtingut la recta: y = -3175.879 + 32.55 * AD + 30.75 * BPD - 400.99 * Ge_RE, amb la qual podem predir el pes d'un nadó, en funció de les variables esmentades anteriorment (AD, BPD i Ge) D'altra banda el coeficient de determinació ajustat és 0,835, de manera que l'ajust és bo. Podem concloure que el model de regressió lineal és un bon procediment per predir el pes d'un nadó.

A la segona part s'han analitzat els possibles factors de risc de tenir un nadó amb baix pes a l'néixer. Segons els resultats, podem concloure que el fet de fumar, és un factor de risc per al baix pes, així com AD i Ge serien factors protectors. És a dir, a mesura que augmenta l'AD i Ge, la probabilitat de baix pes disminueix.

Amb referència als valors de les OR ajustades per fumar i AD, es té: La OR de AD ajustada per fumar és de 0,60, de manera que per cada unitat que augmenti el diàmetre abdominal, el odds de tenir un nadó amb baix pes és 0,60 vegades menor. La OR de fumar ajustada per AD és de 270,8 pel que ens indica que fumar suposa 270 vegades més risc de tenir un nadó amb baix pes, que no tenir-lo. També s'han estudiat els possibles efectes de les variables sexe i ciutat de procedència, sense trobar-se relacions significatives.

A partir dels coeficients del model de regressió logística ajustat, s'ha obtingut la recta: I = 48.67 + 5.60 * fumar_RE - 0.50 * AD, amb la qual podem predir la probabilitat de baix pes al néixer, en funció de fumar i AD. D'altra banda, de l'estudi de la corba ROC, es pot deduir que el model és molt bo a l'hora de discriminar entre aquells nadons que presenten baix pes enfront dels que no. Podem concloure que el model de regressió logística és un bon procediment per predir la probabilitat de néixer amb baix pes.