EX10

roi hezkiyahu

15 5 2022

```
library(glue)
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.9
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::collapse() masks glue::collapse()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.1.1 --
```

```
## v broom     0.7.0      v recipes   0.1.13
## v dials     0.0.8      v rsample   0.0.7
## v infer     0.5.3      v tune      0.1.1
## v modeldata 0.0.2      v workflows 0.1.2
## v parsnip   0.1.2      v yardstick 0.0.7
```

```
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x dplyr::collapse() masks glue::collapse()
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ResourceSelection)
```

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

Q1

שאלה 1

בשאלה זו נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.

אלו נתונים בנוגע לקמפיין של רשת בנקים בפורטוגל, שמטרתו הייתה לשכנע אנשים לקחת הלוואה. כל שורה מייצגת חשבון לקוח/ה והמשתנה אותו נרצה לחזות הוא האם הלקוח/ה לקח/ה הלוואה. המשתנים הנתונים הם:

- y - האם לקח/ה הלוואה (כן/לא)
- age - גיל (נומרי)
- housing - האם ללקוח/ה יש משכנתא (כן/לא)
- university - האם הלקוח/ה עשה/עשתה תואר באוניברסיטה (כן/לא)
- contact - המספר אליו נערכה השיחה (סלולרי/קווי)
- duration - אורך השיחה בשניות (נומרי)
- previous - מספר הפניות הקודמות שנעשו ללקוח/ה לפני קמפיין זה (נומרי)
- marital - מצב משפחתי (רווק/ה, נשוי/אה, גרוש/ה)

נרצה לבנות מודל רגרסיה לוגיסטית על מנת לחזות הלוואה. תוכלו להשתמש בשבעת המשתנים המסבירים הנתונים: גיל, האם ללקוח/ה יש משכנתא, האם עשה/עשתה תואר באוניברסיטה, המספר אליו נערכה השיחה, אורך השיחה בשניות, מספר הפניות הקודמות ומצב משפחתי (ללא אינטראקציות).

א. כמה מודלים אפשריים קיימים (תוכלו להשתמש בשבעת המשתנים המסבירים הנתונים, ללא אינטראקציות, עם אפקטים לינאריים בלבד למשתנים הרציפים)?

ב. השתמשו ב Backward elimination על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

ג. השתמשו ב Forward selection על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

ד. השתמשו ב Backward elimination וב- Forward selection יחד על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

a

we have 7 features to choose from each feature can either be in the model or not thus we have $2^7 = 128$ possible models

b

```
bank <- read_csv("bank.csv") %>% dplyr::select(-1,-2) %>% mutate(y = ifelse(y=="yes",1,0))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   ...1 = col_double(),
##   age = col_double(),
##   university = col_character(),
##   housing = col_character(),
##   duration = col_double(),
##   contact = col_character(),
##   marital = col_character(),
##   previous = col_double(),
##   y = col_character()
## )
```

```
full_model <- glm(y~.,data = bank,family = "binomial")
back <- MASS::stepAIC(full_model,method = "backward")
```

```
## Start:  AIC=21596.98
## y ~ age + university + housing + duration + contact + marital +
##     previous
##
##              Df Deviance   AIC
## - housing     1    21579 21595
## <none>             21579 21597
## - university  1    21649 21665
## - age         1    21692 21708
## - marital     2    21698 21712
## - contact     1    22111 22127
## - previous    1    22762 22778
## - duration    1    26741 26757
##
## Step:  AIC=21595.34
## y ~ age + university + duration + contact + marital + previous
##
##              Df Deviance   AIC
## <none>             21579 21595
## - university  1    21650 21664
## - age         1    21693 21707
## - marital     2    21699 21711
## - contact     1    22116 22130
## - previous    1    22762 22776
## - duration    1    26741 26755
```

```
summary(back)
```

```
##
## Call:
## glm(formula = y ~ age + university + duration + contact + marital +
##     previous, family = "binomial", data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5107  -0.4145  -0.2984  -0.1961   3.0612
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.463e+00  1.070e-01 -41.695  < 2e-16 ***
## age               1.918e-02  1.787e-03  10.732  < 2e-16 ***
## universityyes     3.243e-01  3.834e-02   8.459  < 2e-16 ***
## duration          3.973e-03  6.312e-05  62.943  < 2e-16 ***
## contacttelephone -1.032e+00  4.766e-02 -21.646  < 2e-16 ***
## maritalmarried    3.889e-02  6.106e-02   0.637    0.524
## maritalsingle     5.077e-01  6.832e-02   7.431 1.08e-13 ***
## previous          9.177e-01  2.709e-02  33.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28999  on 41187  degrees of freedom
## Residual deviance: 21579  on 41180  degrees of freedom
## AIC: 21595
##
## Number of Fisher Scoring iterations: 6
```

the model that was chosen is a model containing age,university,duration,contact,marital and precious

each estimate is the log OR for a 1 unit change

C

```
forward <- MASS::stepAIC(full_model,method = "forward")
```

```
## Start:  AIC=21596.98
## y ~ age + university + housing + duration + contact + marital +
##     previous
##
##              Df Deviance   AIC
## - housing     1    21579 21595
## <none>             21579 21597
## - university  1    21649 21665
## - age         1    21692 21708
## - marital     2    21698 21712
## - contact     1    22111 22127
## - previous    1    22762 22778
## - duration    1    26741 26757
##
## Step:  AIC=21595.34
## y ~ age + university + duration + contact + marital + previous
##
##              Df Deviance   AIC
## <none>             21579 21595
## - university  1    21650 21664
## - age         1    21693 21707
## - marital     2    21699 21711
## - contact     1    22116 22130
## - previous    1    22762 22776
## - duration    1    26741 26755
```

```
summary(forward)
```

```
##
## Call:
## glm(formula = y ~ age + university + duration + contact + marital +
##     previous, family = "binomial", data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5107  -0.4145  -0.2984  -0.1961   3.0612
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.463e+00  1.070e-01 -41.695  < 2e-16 ***
## age               1.918e-02  1.787e-03  10.732  < 2e-16 ***
## universityyes     3.243e-01  3.834e-02   8.459  < 2e-16 ***
## duration          3.973e-03  6.312e-05  62.943  < 2e-16 ***
## contacttelephone -1.032e+00  4.766e-02 -21.646  < 2e-16 ***
## maritalmarried    3.889e-02  6.106e-02   0.637    0.524
## maritalsingle     5.077e-01  6.832e-02   7.431 1.08e-13 ***
## previous          9.177e-01  2.709e-02  33.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28999  on 41187  degrees of freedom
## Residual deviance: 21579  on 41180  degrees of freedom
## AIC: 21595
##
## Number of Fisher Scoring iterations: 6
```

the model that was chosen is a model containing age,university,duration,contact,marital and precious

each estimate is the log OR for a 1 unit change

d

```
both <- MASS::stepAIC(full_model,method = "both")
```

```
## Start:  AIC=21596.98
## y ~ age + university + housing + duration + contact + marital +
##     previous
##
##              Df Deviance   AIC
## - housing     1    21579 21595
## <none>             21579 21597
## - university  1    21649 21665
## - age         1    21692 21708
## - marital     2    21698 21712
## - contact     1    22111 22127
## - previous    1    22762 22778
## - duration    1    26741 26757
##
## Step:  AIC=21595.34
## y ~ age + university + duration + contact + marital + previous
##
##              Df Deviance   AIC
## <none>             21579 21595
## - university  1    21650 21664
## - age         1    21693 21707
## - marital     2    21699 21711
## - contact     1    22116 22130
## - previous    1    22762 22776
## - duration    1    26741 26755
```

```
summary(both)
```

```
##
## Call:
## glm(formula = y ~ age + university + duration + contact + marital +
##     previous, family = "binomial", data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5107  -0.4145  -0.2984  -0.1961   3.0612
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.463e+00  1.070e-01 -41.695  < 2e-16 ***
## age               1.918e-02  1.787e-03  10.732  < 2e-16 ***
## universityyes     3.243e-01  3.834e-02   8.459  < 2e-16 ***
## duration          3.973e-03  6.312e-05  62.943  < 2e-16 ***
## contacttelephone -1.032e+00  4.766e-02 -21.646  < 2e-16 ***
## maritalmarried    3.889e-02  6.106e-02   0.637    0.524
## maritalsingle     5.077e-01  6.832e-02   7.431 1.08e-13 ***
## previous          9.177e-01  2.709e-02  33.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28999  on 41187  degrees of freedom
## Residual deviance: 21579  on 41180  degrees of freedom
## AIC: 21595
##
## Number of Fisher Scoring iterations: 6
```

the model that was chosen is a model containing age,university,duration,contact,marital and precious

each estimate is the log OR for a 1 unit change

Q2

שאלה 2

נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.
נרצה לבנות מודל רגרסיה לוגיסטית החוזה הלוואה. לשם כך נרצה לבחור אחד משלושת האופציות למשתנים מסבירים:
- גיל, האם עשה/עשתה תואר באוניברסיטה ואינטראקציה שלהם.
- גיל, האם יש משכנתא ואינטראקציה שלהם.
- האם יש משכנתא, האם עשה/עשתה תואר באוניברסיטה ואינטראקציה שלהם.
השתמשו Leave One Out Cross Validation (LOOCV) על מנת לבחור מבין שלושת המודלים את המודל הטוב ביותר מבחינת Deviance.

```
# this takes to long to run so i wont knit it to the pdf.
# ressults were saved to csv so i could reuse them

mod_1 <-  glm(y~age+university+age*university,data = bank,family = "binomial")
mod_2 <-  glm(y~age+housing+age*housing,data = bank,family = "binomial")
mod_3 <-  glm(y~housing+university+housing*university,data = bank,family = "binomial")

CalcValidDev <- function(valid_preds, valid_y)
{
-2*(sum(valid_y*log(valid_preds) + (1-valid_y)*log(1-valid_preds)))
}

dev_1 <- c()
dev_2 <- c()
dev_3 <- c()

for (i in 1:nrow(bank)){
  print(i)
  #data
  bank_dat <- bank[-i,]
  left_dat <- bank[i,]
  #models
  mod_1 <-  glm(y~age+university+age*university,data = bank_dat,family = "binomial")
  mod_2 <-  glm(y~age+housing+age*housing,data = bank_dat,family = "binomial")
  mod_3 <-  glm(y~housing+university+housing*university,data = bank_dat,family = "binomial")
  #predictions
  pred_1 <- as.numeric(predict(mod_1,left_dat,type = "response"))
  pred_2 <- as.numeric(predict(mod_2,left_dat,type = "response"))
  pred_3 <- as.numeric(predict(mod_3,left_dat,type = "response"))
  #deviances
  dev_1[i] <- CalcValidDev(pred_1,left_dat$y)
  dev_2[i] <- CalcValidDev(pred_2,left_dat$y)
  dev_3[i] <- CalcValidDev(pred_3,left_dat$y)
}
#saving results
tbl <- tibble("model 1"= dev_1,
       "model 2"= dev_2,
       "model 3"= dev_3)
```

```
tbl <- read_csv("loocv.csv") %>% dplyr::select(-X1)
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   `model 1` = col_double(),
##   `model 2` = col_double(),
##   `model 3` = col_double()
## )
```

```
tbl%>% summarise_all(mean)
```

```
## # A tibble: 1 x 3
##   `model 1` `model 2` `model 3`
##       <dbl>     <dbl>     <dbl>
## 1     0.700     0.703     0.702
```

Q3

שאלה 3

א. נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.
נרצה לבחון שני מודלים לוגיסטיים. המודל הנבחר בשאלה 1 סעיף ד', והמודל הנבחר בשאלה 2. התאימו את שני המודלים.

ב. השתמשו בפונקציה hoslem.test, מהחבילה ResourceSelection, על מנת לבצע Hosmer-Lemeshow test לכל אחד משני המודלים שהתאמתם בסעיף א.

I. עשו זאת פעם אחת עם חלוקה ל5 קבוצות. מה ניתן ללמוד מהתוצאה?

II. עשו זאת פעם נוספת עם חלוקה ל10 קבוצות. מה ניתן ללמוד מהתוצאה?

a

```
mod_a <- both
mod_b <- glm(y~age+university+age*university,data = bank,family = "binomial")
```

b

```
#g5
ht_a_5 <- ResourceSelection::hoslem.test(bank$y,predict(mod_a,type = "response"),g=5)
ht_b_5 <- ResourceSelection::hoslem.test(bank$y,predict(mod_b,type = "response"),g=5)
#g10
ht_a_10 <- ResourceSelection::hoslem.test(bank$y,predict(mod_a,type = "response"),g=10)
ht_b_10 <- ResourceSelection::hoslem.test(bank$y,predict(mod_b,type = "response"),g=10)

tibble("g"= c(5,10),
       "mod_a" = c(ht_a_5$statistic,ht_a_10$statistic) ,
       "mod_b" = c(ht_b_5$statistic,ht_b_10$statistic) )
```

```
## # A tibble: 2 x 3
##       g mod_a mod_b
##   <dbl> <dbl> <dbl>
## 1     5  248.  383.
## 2    10  357.  560.
```

both models have a good fit according to holsem test