AIED\_Classification\_Assignment2

20201564 김성현

2023-06-21

### Set the working directory and read in the dataset (PISA2018MS\_KOR\_BQ.rdata)

infile <- outfile <- "C:/aied"  
setwd(infile)   
getwd()

## [1] "C:/aied"

load("PISA2018MS\_KOR\_BQ.Rdata")  
class(PISA2018MS\_KOR)

## [1] "data.frame"

### Classification using the logistic regression

#### Make the table of the outcome variable EC154Q02IA

* EC154Q02IA: attending additional instruction in mathematics
* Make sure to to include if there are any missing cases (useNA='always')

table(PISA2018MS\_KOR$EC154Q02IA, useNA='always')

##   
## 1 2 <NA>   
## 2586 3843 221

#### Treatment of outcome variable

* Remove the missing cases of the outcome variable
* Check the dimension
* Recode 1 -> 0 & 2 -> 1
* Create the table again to check if the recoding is done successfully

PISA2018MS\_KOR <- PISA2018MS\_KOR[!is.na(PISA2018MS\_KOR$EC154Q02IA),]  
dim(PISA2018MS\_KOR)

## [1] 6429 863

PISA2018MS\_KOR$EC154Q02IA <- PISA2018MS\_KOR$EC154Q02IA - 1 # recode 1 -> 0 & 2 -> 1  
  
table(PISA2018MS\_KOR$EC154Q02IA, useNA='always')

##   
## 0 1 <NA>   
## 2586 3843 0

contab <- table(math = PISA2018MS\_KOR$EC154Q02IA, gender = PISA2018MS\_KOR$ST004D01T)  
contab\_margins <- addmargins(contab)  
contab\_margins

## gender  
## math 1 2 Sum  
## 0 1226 1360 2586  
## 1 1888 1955 3843  
## Sum 3114 3315 6429

### Three ways to compute proportions: outcome variable by gender

#### Create three types of tables

prop.table(contab)

## gender  
## math 1 2  
## 0 0.1906984 0.2115415  
## 1 0.2936693 0.3040908

prop.table(contab, margin = 1)

## gender  
## math 1 2  
## 0 0.4740913 0.5259087  
## 1 0.4912829 0.5087171

prop.table(contab, margin = 2)

## gender  
## math 1 2  
## 0 0.3937058 0.4102564  
## 1 0.6062942 0.5897436

#### Interpret the proportions of each table

* Out of total observations, 30.4% are male who do not attending additional instruction in mathematics while 21.2% are male who do attending additional instruction in mathematics. 29.3% are female who do not attending additional instruction in mathematics while 19.1% are female who do attending additional instruction in mathematics.
* Among the students who do not attend additional instruction in math (conditional on math == 1), 49.1% are female students and 50.9% are male students
* Among female students (conditional on gender == 1), 39.4% do attending additional instruction in mathematics while 60.6% do not attending additional instruction in mathematics.

### Logistic regression

#### Fit the following three logistic regressions (no interpretation is required)

* M1: EC154Q02IA by ST004D01T (gender)
* M2: EC154Q02IA by explanatory variable: PV1MATH, ST004D01T (gender)
* M3: EC154Q02IA by explanatory variables: PV1MATH, ST004D01T (gender), interaction between PV1MATH and ST004D01T (gender)

M1 <- glm(EC154Q02IA ~ as.factor(ST004D01T), data = PISA2018MS\_KOR, family = "binomial")  
summary(M1)

##   
## Call:  
## glm(formula = EC154Q02IA ~ as.factor(ST004D01T), family = "binomial",   
## data = PISA2018MS\_KOR)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.43176 0.03668 11.771 <2e-16 \*\*\*  
## as.factor(ST004D01T)2 -0.06886 0.05091 -1.352 0.176   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8665.1 on 6428 degrees of freedom  
## Residual deviance: 8663.3 on 6427 degrees of freedom  
## AIC: 8667.3  
##   
## Number of Fisher Scoring iterations: 4

M2 <- glm(EC154Q02IA ~ as.factor(ST004D01T) + PV1MATH, data = PISA2018MS\_KOR, family = "binomial")  
summary(M2)

##   
## Call:  
## glm(formula = EC154Q02IA ~ as.factor(ST004D01T) + PV1MATH, family = "binomial",   
## data = PISA2018MS\_KOR)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.3449389 0.1582351 21.139 <2e-16 \*\*\*  
## as.factor(ST004D01T)2 -0.0266928 0.0525717 -0.508 0.612   
## PV1MATH -0.0054961 0.0002871 -19.145 <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: 8665.1 on 6428 degrees of freedom  
## Residual deviance: 8259.8 on 6426 degrees of freedom  
## AIC: 8265.8  
##   
## Number of Fisher Scoring iterations: 4

M3 <- glm(EC154Q02IA ~ as.factor(ST004D01T) + PV1MATH + as.factor(ST004D01T):PV1MATH, data = PISA2018MS\_KOR, family = "binomial")  
summary(M3)

##   
## Call:  
## glm(formula = EC154Q02IA ~ as.factor(ST004D01T) + PV1MATH + as.factor(ST004D01T):PV1MATH,   
## family = "binomial", data = PISA2018MS\_KOR)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.7949984 0.2382040 15.932 < 2e-16 \*\*\*  
## as.factor(ST004D01T)2 -0.8364192 0.3169486 -2.639 0.00832 \*\*   
## PV1MATH -0.0063358 0.0004382 -14.460 < 2e-16 \*\*\*  
## as.factor(ST004D01T)2:PV1MATH 0.0015034 0.0005802 2.591 0.00957 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8665.1 on 6428 degrees of freedom  
## Residual deviance: 8253.0 on 6425 degrees of freedom  
## AIC: 8261  
##   
## Number of Fisher Scoring iterations: 4

### [BONUS] Model Evaluation (5 points)

#### Predict the probabilities and values (either 0 or 1) from each model

* Keep get\_logistic\_pred function
* You need to generate three sets of predicted values based on each model (M1, M2, M3)

get\_logistic\_pred = function(mod, data, res = "y", pos = 1, neg = 0, cut = 0.5) {  
 probs = predict(mod, newdata = data, type = "response")  
 ifelse(probs >= cut, pos, neg)  
}  
pred\_M1 <- get\_logistic\_pred(M1, data=PISA2018MS\_KOR, res="EC154Q02IA", cut = 0.5)  
tab\_M1 <- table(pred=pred\_M1, actual=PISA2018MS\_KOR$EC154Q02IA)  
pred\_M2 <- get\_logistic\_pred(M2, data=PISA2018MS\_KOR, res="EC154Q02IA", cut = 0.5)  
tab\_M2 <- table(pred=pred\_M2, actual=PISA2018MS\_KOR$EC154Q02IA)  
pred\_M3 <- get\_logistic\_pred(M3, data=PISA2018MS\_KOR, res="EC154Q02IA", cut = 0.5)  
tab\_M3 <- table(pred=pred\_M3, actual=PISA2018MS\_KOR$EC154Q02IA)

#### Evaluate models (M2 & M3) based on the following quantities

* Report Accuracy, Specificity, Subjectivity, F1 score from M2 & M3
* Use confusion matrix

#install.packages("caret")  
library(caret)

## Warning: 패키지 'caret'는 R 버전 4.3.1에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: ggplot2

## 필요한 패키지를 로딩중입니다: lattice

confusionMatrix(tab\_M2, mode='everything', positive="1")

## Confusion Matrix and Statistics  
##   
## actual  
## pred 0 1  
## 0 847 584  
## 1 1739 3259  
##   
## Accuracy : 0.6387   
## 95% CI : (0.6268, 0.6504)  
## No Information Rate : 0.5978   
## P-Value [Acc > NIR] : 9.269e-12   
##   
## Kappa : 0.1894   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8480   
## Specificity : 0.3275   
## Pos Pred Value : 0.6521   
## Neg Pred Value : 0.5919   
## Precision : 0.6521   
## Recall : 0.8480   
## F1 : 0.7372   
## Prevalence : 0.5978   
## Detection Rate : 0.5069   
## Detection Prevalence : 0.7774   
## Balanced Accuracy : 0.5878   
##   
## 'Positive' Class : 1   
##

confusionMatrix(tab\_M3, mode='everything',positive="1")

## Confusion Matrix and Statistics  
##   
## actual  
## pred 0 1  
## 0 860 581  
## 1 1726 3262  
##   
## Accuracy : 0.6412   
## 95% CI : (0.6293, 0.6529)  
## No Information Rate : 0.5978   
## P-Value [Acc > NIR] : 5.034e-13   
##   
## Kappa : 0.1955   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8488   
## Specificity : 0.3326   
## Pos Pred Value : 0.6540   
## Neg Pred Value : 0.5968   
## Precision : 0.6540   
## Recall : 0.8488   
## F1 : 0.7388   
## Prevalence : 0.5978   
## Detection Rate : 0.5074   
## Detection Prevalence : 0.7759   
## Balanced Accuracy : 0.5907   
##   
## 'Positive' Class : 1   
##