MVA Project - Logistics Analysis

Chun-Jung Chen & Akshay Arora

Question 1: Deos Family conditions affect students' final grade in Math?

To fit the logistics regression, I transformed depedent variables from continuous to binary.

Frequency Table

The two by two frequency table for final math and family support, the majority of the students were above mean of the final math and they obtaind their family's support.

```r
xtabs(~FinalMath + famsup, data = FM.data)
```

```
##          famsup
## FinalMath   0   1
##         0  63 112
##         1  76 119
```

The frequency table showed that the majority of students were above mean of final math score but their parents were living apart.

```r
xtabs(~FinalMath + Pstatus, data = FM.data)
```

```
##          Pstatus
## FinalMath   1   0
##         0  15 160
##         1  23 172
```

Simple logistic regression

The Simple logistic regression model for final math is initially formed by

$$FinalMath = \beta_0 + \beta_1\ famsup + \epsilon_j \ \ j = 1, ..., n$$

After fitting, the model is

$$FinalMath = 0.1876 - 0.1270\ famsup \ \ j = 1, ..., n$$

- The result showed that both intercept and family support are insignificant, means that there is not relationship between final math support and students with or without family supports.

```
q1.logreg1 = glm(FinalMath ~ famsup, data = FM.data, family = "binomial")
summary(q1.logreg1)
```

```
##
## Call:
## glm(formula = FinalMath ~ famsup, family = "binomial", data = FM.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.258  -1.203   1.099   1.152   1.152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1876     0.1704   1.101    0.271
## famsup1      -0.1270     0.2153  -0.590    0.555
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.85  on 369  degrees of freedom
## Residual deviance: 511.50  on 368  degrees of freedom
## AIC: 515.5
##
## Number of Fisher Scoring iterations: 3
```

Multi logistic regression

The logistic regression model for final math is initially formed by

$$FinalMath = \beta_0 + \beta_1\ famsize + \beta_2\ Pstatus0 + \beta_3\ famrel + \beta_4\ famsup1 + \epsilon_j \ \ j = 1, ..., n$$

After fitting, the model is

$$FinalMath = 0.71671 - 0.01566\ famsize - 0.33103\ Pstatus0 - 0.3965\ famrel - 0.1139\ famsup1 \ \ j = 1, ..., n$$

- The result showed that all independent variables are insignificant, indiciating that the model should reform and find more variables to better present the final math score.

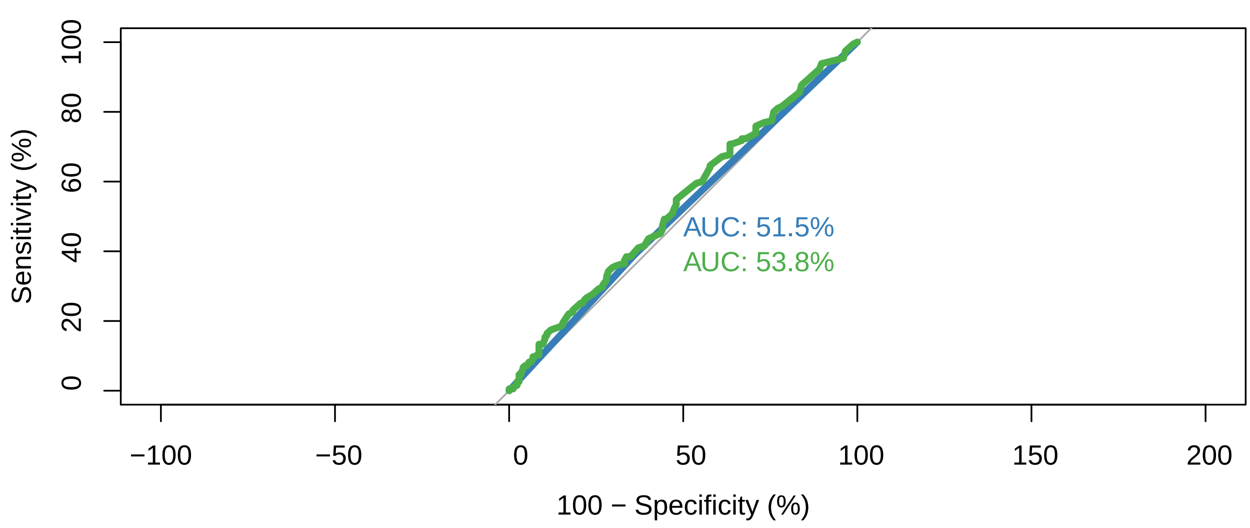
```
q1.logreg2 = glm(FinalMath ~ ., data = FM.data, family = "binomial")
summary(q1.logreg2)
```

```
##
## Call:
## glm(formula = FinalMath ~ ., family = "binomial", data = FM.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4147  -1.2057   0.9984   1.1456   1.2129
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.71671    0.58505   1.225    0.221
## famsize     -0.01566    0.03594  -0.436    0.663
## Pstatus0    -0.33103    0.35218  -0.940    0.347
## famrel      -0.03965    0.11509  -0.344    0.730
## famsup1     -0.11390    0.21732  -0.524    0.600
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.85  on 369  degrees of freedom
## Residual deviance: 510.15  on 365  degrees of freedom
## AIC: 520.15
##
## Number of Fisher Scoring iterations: 4
```

AUC Plot

the plot showed that the multilogistics regression model is better than the simple logistics model, the only reason that the AUC is higher in multilogistics regression model is because it contains more variables. However, they are not significant enough to present the final math score.

```
##
## Call:
## roc.default(response = FM.data$FinalMath, predictor = q1.logreg1$fitted.values,     percent = TRUE, p
##
## Data: q1.logreg1$fitted.values in 175 controls (FM.data$FinalMath 0) < 195 cases (FM.data$FinalMath 1
## Area under the curve: 51.49%
```



Question 2: Does parents' jobs and education level influence students' first period of grade in Math?

To fit the logistics regression, I transformed depedent variables from continuous to binary.

Frequency Table

```
xtabs(~FirstMath + Medu, data = FM.q2data)
```

```
##          Medu
## FirstMath  0  1  2  3  4
##         0  1 35 50 47 50
##         1  2 14 46 46 79
```

```
xtabs(~FirstMath + Fedu, data = FM.q2data)
```

```
##          Fedu
## FirstMath  0  1  2  3  4
##         0  0 52 45 52 34
##         1  2 21 60 45 59
```

```
xtabs(~FirstMath + Mjob, data = FM.q2data)
```

```
##          Mjob
## FirstMath  1  2  3  4  5
##         0 34 11 76 36 26
##         1 19 22 58 57 31
```

```
xtabs(~FirstMath + Fjob, data = FM.q2data)
```

```
##          Fjob
## FirstMath   1   2   3   4   5
##         0   7   8 110  50   8
##         1   9   9  95  53  21
```

Simple Logistic regression

The Simple logistic regression model for first math is initially formed by

$$FirstMath = \beta_0 + \beta_1\ Mjob2 + \beta_2\ Mjob3 + \beta_3\ Mjob4 + \beta_4\ Mjob5 + \epsilon_j \ \ j = 1, ..., n$$

After fitting, the model is

$$FirstMath = -0.5819 + 1.2751\ Mjob2 + 0.3116\ Mjob3 + 1.0415\ Mjob4 + 0.7578\ Mjob5 \ \ j = 1, ..., n$$

- The result showed that except Mjob3, i.e. civil 'services' (e.g. administrative or police) is insignificant, others are all siginificant, indicating that the first math score can be presented by the Mother's job type.

```r
q2.logreg1 = glm(FirstMath ~ Mjob, data = FM.q2data, family = "binomial")
summary(q2.logreg1)
```

```
##
## Call:
## glm(formula = FirstMath ~ Mjob, family = "binomial", data = FM.q2data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4823  -1.0650   0.9005   1.1037   1.4324
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5819     0.2864  -2.032  0.04219 *
## Mjob2         1.2751     0.4673   2.728  0.00637 **
## Mjob3         0.3116     0.3353   0.929  0.35271
## Mjob4         1.0415     0.3569   2.918  0.00352 **
## Mjob5         0.7578     0.3908   1.939  0.05252 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 512.89  on 369  degrees of freedom
## Residual deviance: 497.24  on 365  degrees of freedom
## AIC: 507.24
##
## Number of Fisher Scoring iterations: 4
```

Multi logistic regression

The fitted model shows that only mother job 4 (at home) is significant for first math score.

```r
q2.logreg2 = glm(FirstMath ~ ., data = FM.q2data, family = "binomial")
summary(q2.logreg2)
```

```
##
## Call:
## glm(formula = FirstMath ~ ., family = "binomial", data = FM.q2data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9247  -1.0870   0.3123   1.0788   1.8956
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.8377   615.0531   0.026   0.9795
## Medu1        -1.5189     1.3336  -1.139   0.2547
## Medu2        -1.0482     1.3151  -0.797   0.4254
## Medu3        -1.0661     1.3341  -0.799   0.4242
## Medu4        -0.8303     1.3623  -0.609   0.5422
## Fedu1       -15.6319   615.0516  -0.025   0.9797
## Fedu2       -14.6567   615.0516  -0.024   0.9810
## Fedu3       -15.1897   615.0516  -0.025   0.9803
```

```
## Fedu4       -14.7419   615.0517  -0.024   0.9809
## Mjob2         0.9881     0.5648   1.749   0.0802 .
## Mjob3         0.2682     0.3733   0.718   0.4726
## Mjob4         0.8457     0.4093   2.066   0.0388 *
## Mjob5         0.1688     0.5194   0.325   0.7452
## Fjob2        -0.5915     0.7675  -0.771   0.4409
## Fjob3        -0.3023     0.5639  -0.536   0.5919
## Fjob4        -0.1980     0.5849  -0.339   0.7350
## Fjob5         0.4280     0.7199   0.594   0.5522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 512.89  on 369  degrees of freedom
## Residual deviance: 470.48  on 353  degrees of freedom
## AIC: 504.48
##
## Number of Fisher Scoring iterations: 13
```

AUC Plot

According to the AUC plot, the multi logistics regression model has higher AUC but it does not have significant difference to the simple logistic regression model since most of the varialbes are insignificant in the multi logistics regression model.

```r
roc(FM.q2data$FirstMath, q2.logreg1$fitted.values, plot = TRUE,
    legacy.axes = TRUE, percent = TRUE, col = "#377eb8", lwd = 4,
    print.auc = TRUE)
```

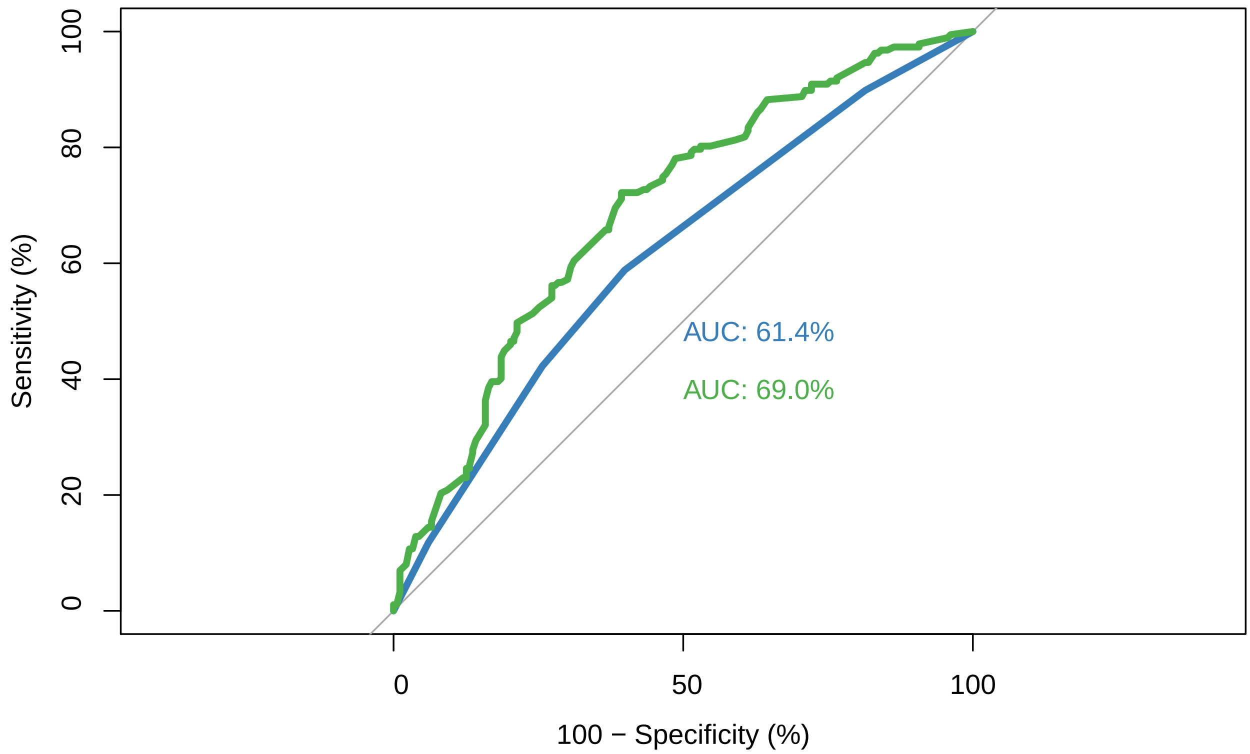
```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

##
## Call:
## roc.default(response = FM.q2data$FirstMath, predictor = q2.logreg1$fitted.values,     percent = TRUE
##
## Data: q2.logreg1$fitted.values in 183 controls (FM.q2data$FirstMath 0) < 187 cases (FM.q2data$FirstMa
## Area under the curve: 61.35%
```

```r
# Lets add the other graph
plot.roc(FM.q2data$FirstMath, q2.logreg2$fitted.values, percent = TRUE,
    col = "#4daf4a", lwd = 4, print.auc = TRUE, add = TRUE, print.auc.y = 40)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Question 3: Does student's learning conditions really impact students' final grade math score and Portuguese scores in average?

Frequency Table

```
xtabs(~FinalAvg + internet, data = FM.q3data)
```

```
##          internet
## FinalAvg  no yes
##        0  35 156
##        1  22 157
```

```
xtabs(~FinalAvg + romantic, data = FM.q3data)
```

```
##          romantic
## FinalAvg  no yes
##        0 124  67
##        1 127  52
```

Simple Logistic regression

The model multi-regression model for final math is initially formed by

$$FinalMath = \beta_0 + \beta_1\, internet + \beta_2\, romantic + \beta_3\, freetime + \beta_4\, normtraveltime + \epsilon_j \quad j = 1, ..., n$$

After fitting, the model is

$$\begin{aligned} FinalMath = -0.81655 + 0.50518\, internet - 0.30615\, romantic - 0.09102\, freetime \\ +1.23613\, normtraveltime \quad j = 1, ..., n \end{aligned}$$

- The result showed that all the independent variables are insignificant, means that internet access, romantic relationship, free time, and travel time does not explained the final average score well.

```r
q3.logreg1 = glm(FinalAvg ~ ., data = FM.q3data, family = "binomial")
summary(q3.logreg1)
```

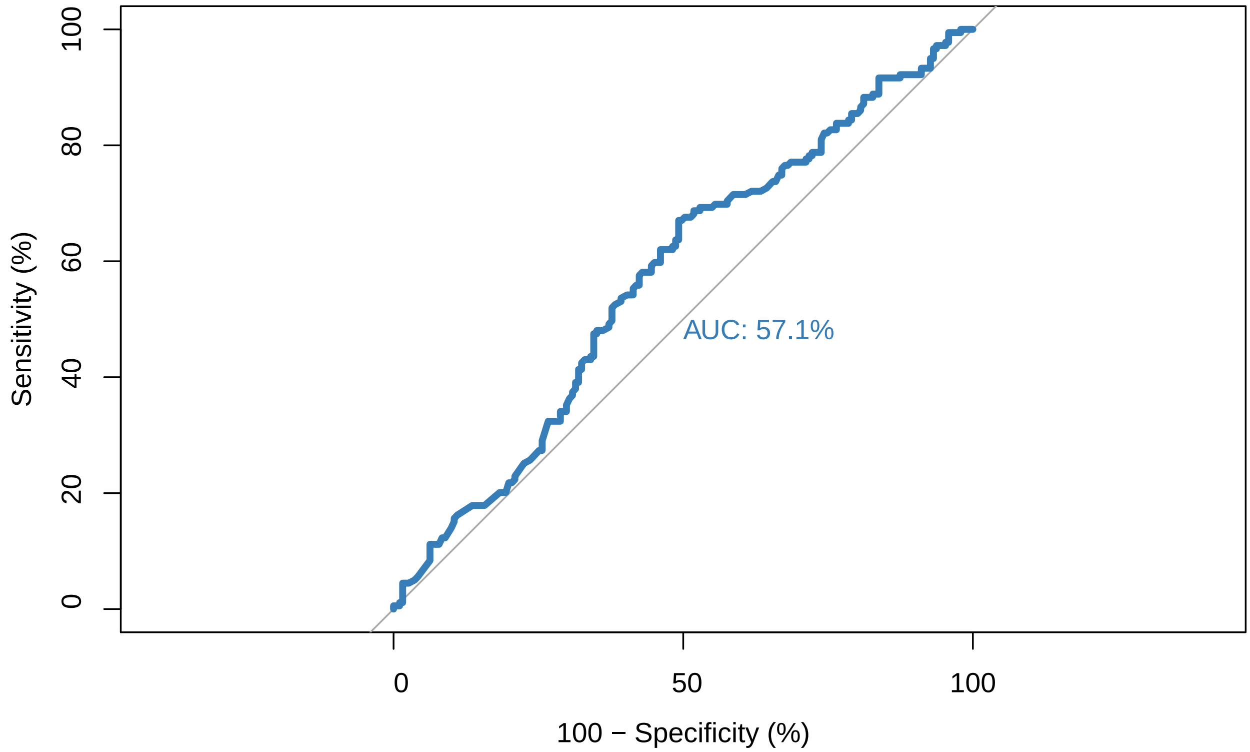
```
##
## Call:
## glm(formula = FinalAvg ~ ., family = "binomial", data = FM.q3data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3924  -1.1485  -0.8739   1.1665   1.5481
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.81655    0.81401  -1.003   0.3158
## internetyes     0.50518    0.29972   1.686   0.0919 .
## romanticyes    -0.30615    0.22620  -1.353   0.1759
## freetime       -0.09102    0.10751  -0.847   0.3972
## normtraveltime  1.23613    1.27994   0.966   0.3342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 512.54  on 369  degrees of freedom
## Residual deviance: 506.46  on 365  degrees of freedom
## AIC: 516.46
##
## Number of Fisher Scoring iterations: 4
```

AUC PLot

```r
roc(FM.q3data$FinalAvg, q3.logreg1$fitted.values, plot = TRUE,
    legacy.axes = TRUE, percent = TRUE, col = "#377eb8", lwd = 4,
    print.auc = TRUE)
```

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

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




```
##
## Call:
## roc.default(response = FM.q3data$FinalAvg, predictor = q3.logreg1$fitted.values,     percent = TRUE,
##
## Data: q3.logreg1$fitted.values in 191 controls (FM.q3data$FinalAvg 0) < 179 cases (FM.q3data$FinalAv
## Area under the curve: 57.12%
```