Statistical Inference

Project PhaseII

Hamidreza Aliakbary khoyi

810196514

Spring 2021

Question 1

Part 1

In this part I chose 2 categorical variables:"Sex" and "Mjob" which signifies job associated to mother of student. Since Mjob has more than 2 level, for construction confidence intervals we are two choose any possible pair selection of Mjob levels and find confidence interval for difference of proportion for selected pair based on Sex of students.

Table of this 2 categorical variable is:

Mjob	sex F	M
at_home	42	17
health	19	15
other	74	67
services	54	49
teacher	19	39

Checking conditions for inference for comparing two independent propotions:

- Independence:
  i. Random sample/assignment
  ii. If sampling without replacement, n<10% of population.
- Sample size/skew:
  Samples should meet success failure situation which indicated each of cases should have at least 10 cases. Based on table above, all of the cells are above 10 so this condition met.
- So all of conditions met.

$$Confidence\ interval: \widehat{p_1} - \widehat{p_2} \pm z^* \times \sqrt{\widehat{p_1} \times \frac{1-\widehat{p_1}}{n_1} + \widehat{p_2} \times \frac{1-\widehat{p_2}}{n_2}}$$

Writing a R code to compute all this Cis is necessary. Q1_CI_calculator does god job to calculate all corresponding CIs. First number is max of CI second is min of CI.

$$CI_{at_{home}\ VS\ health}: (0.17010789, 0.13597386)$$
$$CI_{at_{home}\ VS\ other}: (0.19897568, 0.17510774)$$
$$CI_{at_{home}\ VS\ Services}: (0.20024738, 0.17493774)$$
$$CI_{at_{home}\ VS\ teacher}: (0.39833223, 0.37022417)$$
$$CI_{health\ VS\ other}: (0.04965048\ ,0.01835119)$$
$$CI_{health\ VS\ Services}: (0.05075774\ ,0.01834563)$$
$$CI_{health\ VS\ teacher}: (0.24855805, 0.23191660)$$
$$CI_{other\ VS\ services}: (0.01121729, -0.01011559)$$
$$CI_{other\ VS\ teacher}: (0.20953056, 0.18494241)$$
$$CI_{services\ VS\ teacher}: (0.20968060, 0.18369068)$$

Part b

$$H_0: Mothers\ job\ is\ idepentant\ from\ student'sex.$$
$$H_A: Mothers\ job\ is\ not\ idepentant\ from\ student'sex.$$

For independence test I used χ^2 test.
Conditions for this test:

- Independence:
  - Random sample/assignment
  - If sampling without replacement, n<10% of population.
  - Each case only contributes to one cell.
- Sample size: each cell must have at least 5 expected cases.

$$Expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

$$text\ statistic\ :\ \frac{point\ estimate\ -\ null\ value}{SE\ of\ point\ estimate}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad and \qquad df = (R-1)(C-1)$$

Table for Mjob/sex:

Mjob	sex F	M
at_home	42	17
health	19	15
other	74	67
services	54	49
teacher	19	39

All conditions met.

Result of hypothesis:

```
"X^2: "
17.48356
"DF: "
4
"P-value:"
0.001556439
```

Due to fact that p-value is less than 0.05, we reject Null hypothesis and there is enough evidence that Sex of students and their mother job is dependent.

Question 2

Chosen categorical variable: Internet

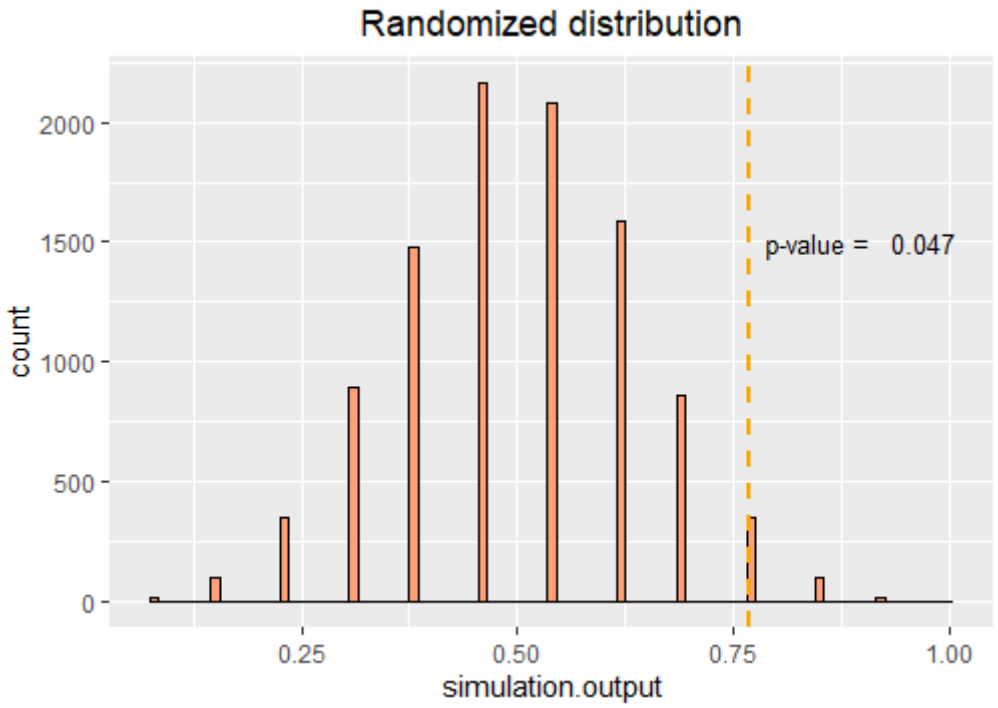
$$H_0: p = 0.5$$

$$H_A: P > 0.5$$

Conditions:

- Independence:
  - Random sample/assignment
  - If sampling without replacement, n>10% population
- Sample size/skew:
  - $n \times \hat{p} = 13 \times \frac{10}{13} = 10 \geq 10 \rightarrow correct$
  - $n \times (1 - \widehat{p}) = 13 \times \frac{3}{13} = 3 < 10 \rightarrow not\ correct$

Since conditions didn't meet, we use simulation.



Since p-value is less than 0.05, we reject null hypothesis, and there is enough evidence that probability of having internet is more than 0.5 .

Question 3

Part a

Chosen categorical variable is Fjob.

Probability distribution of chosen categorical variable is:

```
   at_home     health      other   services    teacher
0.05063291 0.04556962 0.54936709 0.28101266 0.07341772
```

Original samples:

```
at_home   health    other services  teacher
     20       18      217      111       29
```

Conditions for chi-square test:

- Independence:
  - Random sample/assignment
  - If sampling without replacement, n<10% of population.
  - Each case only contributes to one cell.
- Sample size: each cell must have at least 5 expected cases.

$$H_0: randomly\ selected\ samples\ have\ same\ distribution\ of\ original\ sample$$
$$H_A: randomly\ selected\ samples\ do\ not\ have\ same\ distribution\ of\ original\ sample$$

Randomly selected samples:

```
Q3.unbiasedsample
 at_home   health    other services  teacher
       4        5       56       24       11
```

Chi square test result:

```
        Chi-squared test for given probabilities

data:  ub.tab
X-squared = 2.7083, df = 4, p-value = 0.6078
```

Since p-value is more than 0.05, we fail to reject null hypothesis and randomly selected samp ls are closly to have same distribution as orginal sample.

$$H_0: biased\ \ samples\ have\ same\ distribution\ of\ original\ sample$$
$$H_A: biased\ samples\ do\ not\ have\ same\ distribution\ of\ original\ sample$$

Randomly selected samples with 80% tendency to Fjob = services:

```
at_home   health    other services  teacher
     5        6       30       55        4
```

Chi square test result:

```
        Chi-squared test for given probabilities

data:  b.tab
X-squared = 39.046, df = 4, p-value = 6.817e-08
```

Since p-value is less than 0.05, we reject null hypothesis and biased sampls do not have same distribution as orginal sample.

Part b

Other chosen categorical variable: Mjob

By using chi-square test we can test dependency chance of these two variables.

$$Expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

$$test\ statistic : \frac{point\ estimate\ -\ null\ value}{SE\ of\ point\ estimate}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad and \qquad df = (R-1)(C-1)$$

$$H_0: Mjob\ and\ Fjob\ are\ independent\ variables.$$

$$H_A: Mjob\ and\ Fjob\ are\ dependent\ variables.$$

Conditions for chi-square test:

- Independence:
  - Random sample/assignment
  - If sampling without replacement, n<10% of population.
  - Each case only contributes to one cell.
- Sample size: each cell must have at least 5 expected cases.

Mjob	Fjob: at_home	health	other	services	teacher
at_home	7	2	33	15	2
health	0	6	17	10	1
other	5	2	104	24	6
services	6	4	42	43	8
teacher	2	4	21	19	12

Chi-square result:

```
        Pearson's Chi-squared test

data:  table(StudentPerformance[, c("Mjob", "Fjob")])
X-squared = 73.381, df = 16, p-value = 2.534e-09

Warning message:
In chisq.test(table(StudentPerformance[, c("Mjob", "Fjob")])) :
  Chi-squared approximation may be incorrect
```

Since there are sells with value less than 5, I combined athome, health and teacher columns and save it in at_home column, and rerun the test, the results was:

Mjob	Fjob: at_home	services	teacher
at_home	11	15	2
health	7	10	1
other	13	24	6
services	18	43	8
teacher	18	19	12

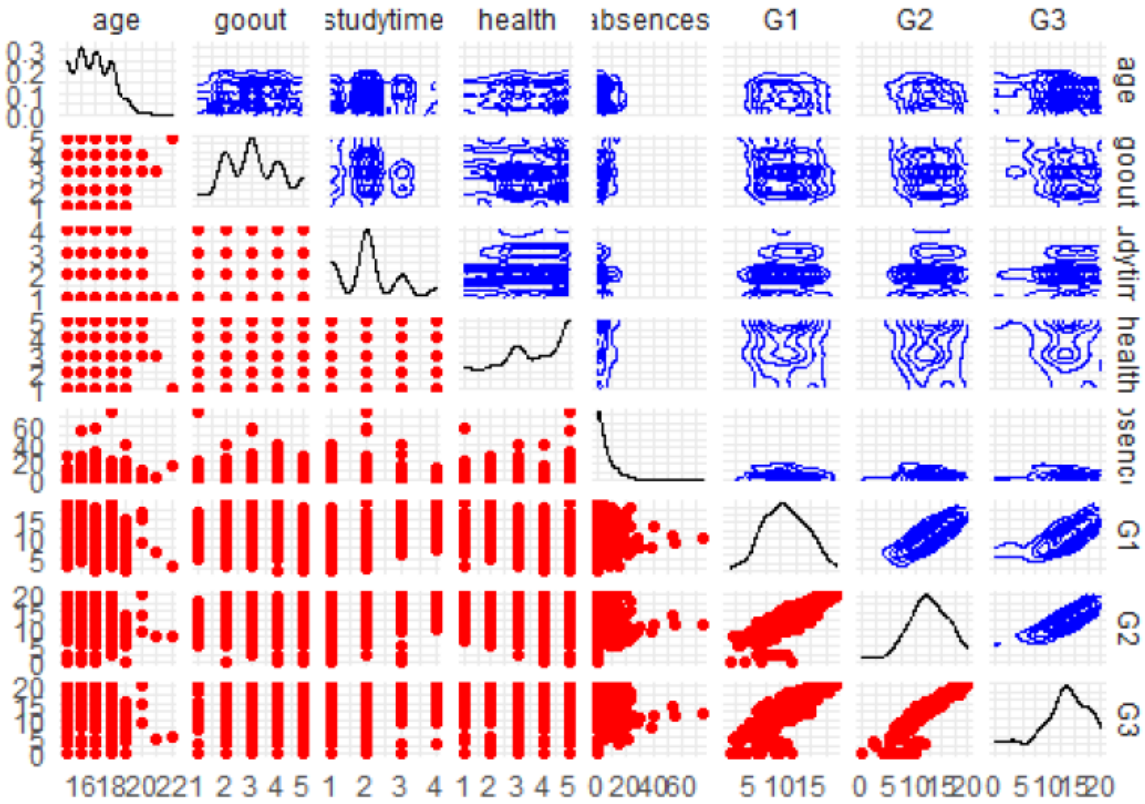
```
        Pearson's Chi-squared test

data:  q3.tab[, sel]
X-squared = 42.454, df = 8, p-value = 1.113e-06
```

Since p-value is less than 0.05, we reject null hypothesis and there is enough evidence that selected variables are dependent.

Question 4

Part a

Based on phase I gout and failure has negative correlation with G1, G2 and G3. And also study time ha a positive correlation with these 3 too. Observing correlogram of these variables from previous phase, I decided to choose G3 as response variable and G1 and study time as our explanatory variable. Based on correlogram :



Based on this plot, we can see high correlations between grades that is reasonable based on the fact that students' performance during course almost stay constant. And also pick study time as a traditional cause of getting good grades and also having a good positive correlation between it and grades.

Part b

Conditions for linear regression:

- Residuals and Fitted : Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, which is good.
- Normal Q-Q : Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line.
- Scale-Location :(or Spread-Location). Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.
- Residuals vs Leverage : Used to identify in uential cases, that is extreme values that might in uence the regression results when included or excluded from the analysis.

Study time:

Fitted:

```
Call:
lm(formula = G3 ~ studytime, data = StudentPerformance)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4076  -2.6972   0.7806   3.6529   8.1824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0225     0.6750  16.331  < 2e-16 ***
studytime     0.7950     0.3066   2.593  0.00987 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.108 on 393 degrees of freedom
Multiple R-squared:  0.01682,   Adjusted R-squared:  0.01432
F-statistic: 6.723 on 1 and 393 DF,  p-value: 0.009875
```

Coefficients:

```
(Intercept)   studytime
 11.0225345   0.7950239
```

$$G_3 = 0.7950239 \times studytime + 11.0225345$$

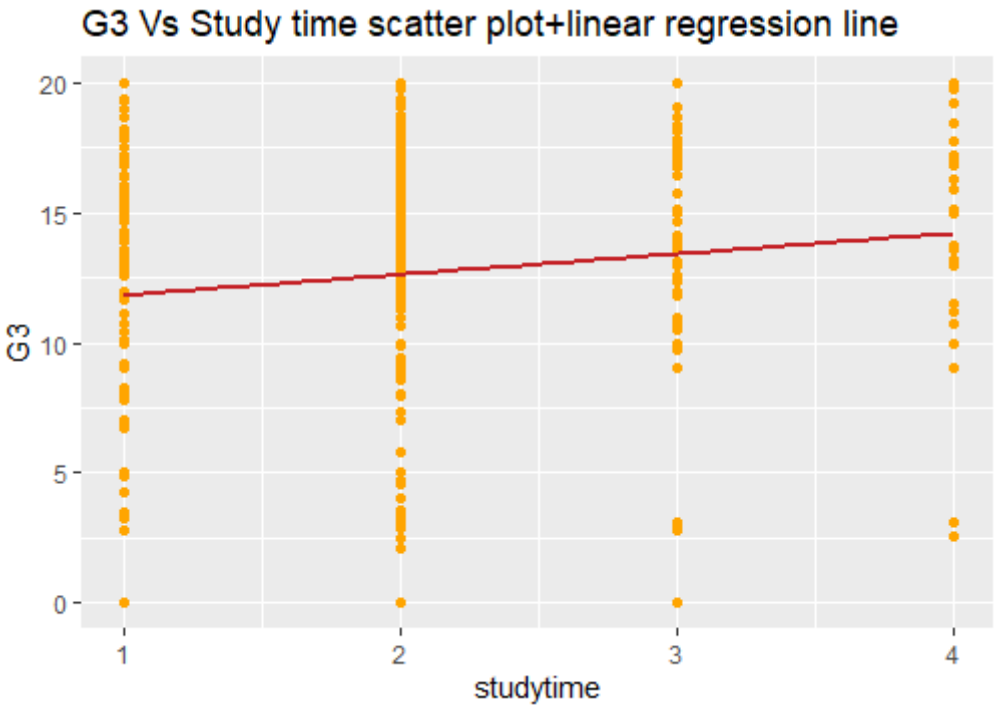
Intercept: when study time is zero, studenst required to get 11.0225345on average.

Slope: for each unit increase in study time, G3 requires to be 0.7950239 higher on average.

Least square:

10253.66

Plot:



G2:

Fitted:

```
Call:
lm(formula = G3 ~ G2, data = StudentPerformance)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0217  -0.4872   0.2822   1.1949   3.8342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.49813    0.33123  -4.523 8.09e-06 ***
G2           1.15199    0.02561  44.985  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.077 on 393 degrees of freedom
Multiple R-squared:  0.8374,    Adjusted R-squared:  0.837
F-statistic:  2024 on 1 and 393 DF,  p-value: < 2.2e-16
```

Coefficients:

```
(Intercept)          G2
  -1.498133    1.151988
```

$$G_3 = 1.151988 \times G_2 - 1.498133$$

Intercept: if G2 is zero, G3 is expected to be intercept which here is negative and impossible, it indicated that G2 scores was an average higher than absolute value of intercept.

Slope: by one unit increase in G2, G3 is expected to increase 1.151988 unites on average.

Least square:

```
> leastquare.first
[1] 1696.009
```

Plot:

Part c

Judginf based on adjuster R squared and p-value of model, we have:

Based on p-value, p-value of study time is 0.009875 and p-value of G2 is 2.2e-16 which indicates that G2 is more significant.

Based on adjusted R-squared, for studytime it is0.01432 and for G2 its value is 0.837 which again signifies G2 is more significant variable.

Part d

Adjusted R-square:

Based on adjusted R-squared, for studytime it is0.011832 and for G2 its value is 0.837 which again signifies G2 is more significant variable.

ars.first	0.836546925385211
ars.second	0.0118021479501507

ANOVA:

both variables:

```
Response: G3
           Df Sum Sq Mean Sq   F value Pr(>F)
G2          1 8733.0  8733.0 2020.2814 <2e-16 ***
studytime   1    1.5     1.5    0.3503 0.5543
Residuals 392 1694.5     4.3
```

Each independently:

```
 Response: G3
            Df  Sum Sq Mean Sq F value   Pr(>F)
 studytime   1   175.4 175.400  6.7227 0.009875 **
 Residuals 393 10253.7  26.091
```

```
Response: G3
           Df Sum Sq Mean Sq F value    Pr(>F)
G2          1   8733  8733.0  2023.6 < 2.2e-16 ***
Residuals 393   1696     4.3
```

Based on p-value, p-value of study time is 0.009875 and p-value of G2 is 2.2e-16 which indicates that G2 is more significant.

Part e

1. Having less p-value in modl and ANOVA
2. Having more adjusted saure
3. Maybe selecting variables witch is more confident to predict response variable.

Part F

H_0: *explanatory variables are not significant predictor of response variable*

H_A: *explanatory variables are significant predictor of response variable*

For G2:

```
Call:
lm(formula = G3 ~ G2, data = samples.train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3402 -0.5968  0.1276  1.1248  3.5846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.04101    0.66119  -4.599 1.41e-05 ***
G2           1.26504    0.05126  24.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.976 on 88 degrees of freedom
Multiple R-squared:  0.8738,    Adjusted R-squared:  0.8723
F-statistic: 609.1 on 1 and 88 DF,  p-value: < 2.2e-16
```

For study time:

```
Call:
lm(formula = G3 ~ studytime, data = samples.train)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8472  -2.3214   0.9437   3.8770   9.0174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.5503     1.5794   6.047  3.5e-08 ***
studytime     1.4323     0.7277   1.968   0.0522 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.443 on 88 degrees of freedom
Multiple R-squared:  0.04217,   Adjusted R-squared:  0.03128
F-statistic: 3.874 on 1 and 88 DF,  p-value: 0.05218
```

Based on model summaries, G2's p-value is less than 0.05 that means we reject null hypothesis regarding G2 and it is a significant predictor. However, studytime's pvalue is more than 0.05 and we fail to reject null hypothesis and it is not a significant predictor.

G2 CI: (1.256593, 1.273487)

Studytime CI: (1.312389, 1.552211)

We are 95% confident that for each additional point on G2, G1 is expected on average to be lower by 1.256593 to 1.273487points.

We are 95% confident that for each additional hour on study time, G1 is expected on average to be lower by 1.312389 to 1.552211points.

Predicted table is:

	pred.G2	pred.st	actual
1	11.454296	12.41490	12.910329
2	11.544141	13.84720	11.830718
3	12.917128	10.98259	13.386484
4	8.992853	10.98259	7.825943
5	18.313362	12.41490	16.855962
6	10.874480	12.41490	11.000000
7	21.374076	12.41490	19.398768
8	12.132461	12.41490	12.916115
9	11.102932	10.98259	10.733352
10	15.286641	10.98259	14.702594

Predicted residuals' table is:

	abs.pred.G2...actual.	abs.pred.st...actual.
1	1.4560321	0.4954296
2	0.2865770	2.0164850
3	0.4693555	2.4038890
4	1.1669103	3.1566518
5	1.4573994	4.4410635
6	0.1255204	1.4148990
7	1.9753082	6.9838686
8	0.7836540	0.5012160
9	0.3695800	0.2492428
10	0.5840467	3.7199995

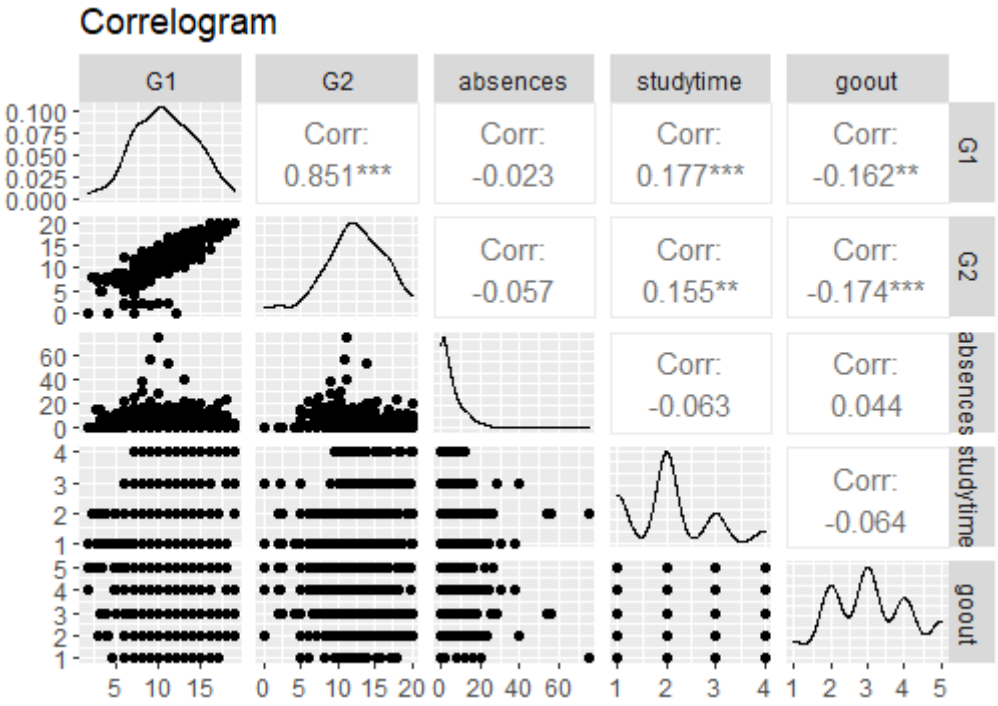
Determining 10% margin as success we get, error margin of 2. So:

Success rate of G2 is 100% but success rate of study time is 40%.

Question 5

Part a

Correlogram plot corresponding chosen explanatory variables:



I choose G3 as my response variable. My chosen explanatory variables are:

- G1
- G2
- Absences
- Studytime
- Gout

With a wise look at this plot, we can see G1 and G2 are more correlated, intuitively we can say that G3 is more correlated with G1 and G2, thus these 2 variables play more significant role in G3 prediction. Furthermore, absences is less correlated with G1 and G2 so it will next best predictor, based on the fact that predictors used to be not correlated with each other.

Part b

Summary of model:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.95339    0.54212  -3.603 0.000355 ***
G1           0.11113    0.05643   1.969 0.049628 *
G2           1.07860    0.04858  22.202  < 2e-16 ***
absences     0.03089    0.01304   2.370 0.018296 *
studytime   -0.07782    0.12594  -0.618 0.537002
goout        0.04498    0.09481   0.474 0.635513
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.06 on 389 degrees of freedom
Multiple R-squared:  0.8417,    Adjusted R-squared:  0.8397
F-statistic: 413.7 on 5 and 389 DF,  p-value: < 2.2e-16
```

With a wise look at above figure, among all chosen explanatory variables, gout and studytime seems to be non-significant. Since p-value < 0.05, the model as a whole is significant.

Part c

Based on value of adjusted R^2, 83.97% of variation in response variable is described by the model.

Part d

Higher R2 doesn't necessarily guarantee that the model ts the data well, we might face overfitting if we are not careful. Adjusted R2 can be a good indicator of when the model fits the data well, it compares the explanatory power of regression models that contain different numbers of predictors. Adjusted R2 is around 84% in our fitted model. The fact that R2 and Adjusted R2 are this close is very good which means we don't have overfitting in our model.

Part e

Forward selection – p value:

```
Forward Selection Method
---------------------------

Candidate Terms:

1. G1
2. G2
3. absences
4. studytime
5. goout

We are selecting variables based on p value...


Forward Selection: Step 1

+ G2

                        Model Summary
-----------------------------------------------------------------
R                       0.915       RMSE                 2.077
R-Squared               0.837       Coef. Var           16.434
Adj. R-Squared          0.837       MSE                  4.316
```

```
Pred R-Squared          0.836       MAE                   1.342
------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
--------------------------------------------------------------------------
Regression     8733.050          1       8733.050    2023.627    0.0000
Residual       1696.009        393          4.316
Total         10429.059        394
--------------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------
-------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lo
wer     upper
----------------------------------------------------------------------------
-------------
(Intercept)    -1.498         0.331                -4.523    0.000    -2.
149    -0.847
         G2     1.152         0.026        0.915    44.985    0.000     1.
102     1.202
----------------------------------------------------------------------------
-------------



Forward Selection: Step 2

+ absences

                        Model Summary
------------------------------------------------------------------
R                       0.917       RMSE                  2.063
R-Squared               0.840       Coef. Var            16.323
Adj. R-Squared          0.839       MSE                   4.257
Pred R-Squared          0.838       MAE                   1.356
------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
--------------------------------------------------------------------------
Regression     8760.143          2       4380.071    1028.805    0.0000
Residual       1668.916        392          4.257
Total         10429.059        394
--------------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------
-------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lo
wer     upper
----------------------------------------------------------------------------
-------------
(Intercept)    -1.731         0.342                -5.066    0.000    -2.
403    -1.059
         G2     1.156         0.025        0.918    45.361    0.000     1.
106     1.206
```

```
   absences     0.033         0.013        0.051     2.523    0.012     0.
007     0.058
-------------------------------------------------------------------------
--------------


Forward Selection: Step 3

+ G1

                         Model Summary
-----------------------------------------------------------------
R                       0.917       RMSE                 2.056
R-Squared               0.841       Coef. Var           16.268
Adj. R-Squared          0.840       MSE                  4.229
Pred R-Squared          0.838       MAE                  1.358
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F          Sig.
-------------------------------------------------------------------------
Regression     8775.599          3       2925.200    691.733    0.0000
Residual       1653.460        391          4.229
Total         10429.059        394
-------------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------
--------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lo
wer     upper
-------------------------------------------------------------------------
--------------
(Intercept)    -1.913         0.354                -5.410    0.000    -2.
609    -1.218
         G2     1.077         0.048        0.855    22.257    0.000     0.
982     1.172
   absences     0.032         0.013        0.049     2.433    0.015     0.
006     0.057
         G1     0.107         0.056        0.073     1.912    0.057    -0.
003     0.217
-------------------------------------------------------------------------
--------------


No more variables to be added.

Variables Entered:

+ G2
+ absences
+ G1


Final Model Output
------------------

                         Model Summary
-----------------------------------------------------------------
R                       0.917       RMSE                 2.056
R-Squared               0.841       Coef. Var           16.268
Adj. R-Squared          0.840       MSE                  4.229
```

```
Pred R-Squared          0.838       MAE                  1.358
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF     Mean Square       F          Sig.
-------------------------------------------------------------------------
Regression     8775.599          3       2925.200     691.733     0.0000
Residual       1653.460        391          4.229
Total         10429.059        394
-------------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------
--------------
      model       Beta     Std. Error     Std. Beta       t         Sig       lo
wer      upper
-------------------------------------------------------------------------------
--------------
(Intercept)     -1.913          0.354                   -5.410     0.000     -2.
609     -1.218
         G2      1.077          0.048         0.855     22.257     0.000      0.
982      1.172
   absences      0.032          0.013         0.049      2.433     0.015      0.
006      0.057
         G1      0.107          0.056         0.073      1.912     0.057     -0.
003      0.217
-------------------------------------------------------------------------------
--------------
```

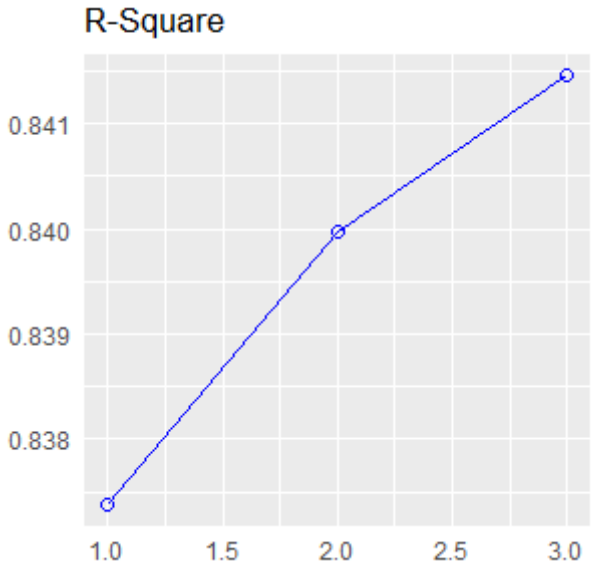
Final model G3~G1+G2+absences

Forward Rsquared:

+G2 => adj Rsqrd = 0.8376

+G2+absences => 0.840

+G2+absences+G1 => 0.843



Backward elimination:

P value:

```
Backward Elimination Method
---------------------------

Candidate Terms:

1 . G1
2 . G2
3 . absences
4 . studytime
5 . goout

We are eliminating variables based on p value...

x goout

Backward Elimination: Step 1

 Variable goout Removed

                        Model Summary
-----------------------------------------------------------------
R                       0.917       RMSE                 2.058
R-Squared               0.842       Coef. Var           16.280
Adj. R-Squared          0.840       MSE                  4.235
Pred R-Squared          0.838       MAE                  1.364
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F         Sig.
-------------------------------------------------------------------------
Regression     8777.306          4       2194.326    518.109    0.0000
Residual       1651.753        390          4.235
Total         10429.059        394
-------------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------
--------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lo
wer     upper
-------------------------------------------------------------------------------
--------------
(Intercept)    -1.784         0.408                -4.371    0.000    -2.
587    -0.982
         G1     0.110         0.056        0.076     1.960    0.051     0.
000     0.221
         G2     1.077         0.048        0.856    22.242    0.000     0.
982     1.172
   absences     0.031         0.013        0.048     2.390    0.017     0.
006     0.057
  studytime    -0.080         0.126       -0.013    -0.635    0.526    -0.
327     0.167
-------------------------------------------------------------------------------
--------------


x studytime

Backward Elimination: Step 2
```

Variable studytime Removed

Model Summary

R	0.917	RMSE	2.056
R-Squared	0.841	Coef. Var	16.268
Adj. R-Squared	0.840	MSE	4.229
Pred R-Squared	0.838	MAE	1.358

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	8775.599	3	2925.200	691.733	0.0000
Residual	1653.460	391	4.229		
Total	10429.059	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.913	0.354		-5.410	0.000	-2.609	-1.218
G1	0.107	0.056	0.073	1.912	0.057	-0.003	0.217
G2	1.077	0.048	0.855	22.257	0.000	0.982	1.172
absences	0.032	0.013	0.049	2.433	0.015	0.006	0.057

No more variables satisfy the condition of p value = 0.3

Variables Removed:

x goout
x studytime

Final Model Output

Model Summary

R	0.917	RMSE	2.056
R-Squared	0.841	Coef. Var	16.268
Adj. R-Squared	0.840	MSE	4.229
Pred R-Squared	0.838	MAE	1.358

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of

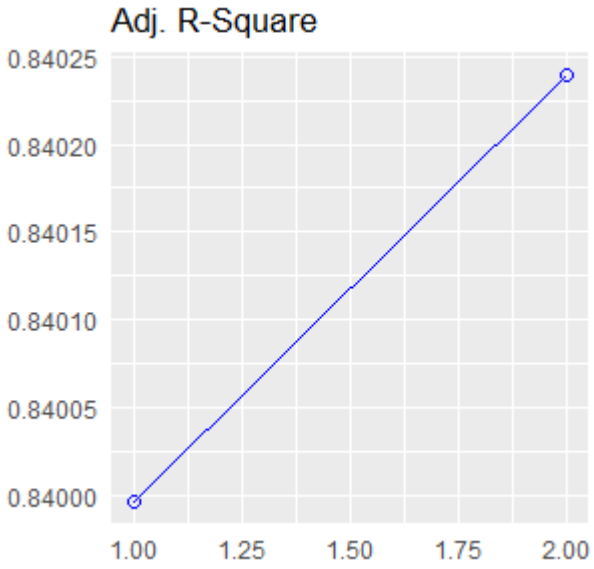
	Squares	DF	Mean Square	F	Sig.
Regression	8775.599	3	2925.200	691.733	0.0000
Residual	1653.460	391	4.229		
Total	10429.059	394			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-1.913	0.354		-5.410	0.000	-2.609	-1.218
G1	0.107	0.056	0.073	1.912	0.057	-0.003	0.217
G2	1.077	0.048	0.855	22.257	0.000	0.982	1.172
absences	0.032	0.013	0.049	2.433	0.015	0.006	0.057

Final model: G3~G1+G2+absences

Adj- R squared:



-goout => 0.84

-goout-studytime => 0.8403

Final model: G3~G1+G2+absences

Part f

Conditions for linear regression :

- **Linear relationships between x and y**: Each (numerical) explanatory variable linearly related to the response variable Check using residuals plots (e vs. x) Looking for a random scatter around 0 Instead of scatterplot of y vs. x: allows for considering the other variables that are also in the model, and not just the bivariate relationship between a given x and y
- **Nearly normal residuals:** we look for random scatter of residuals around 0 This translates to a nearly normal distribution of residuals centered at 0 Check using histogram or normal probability plot
- **Constant variability of residuals :** Residuals should be equally variable for low and high values of the predicted response variable Check using residuals plots of residuals vs. predicted (e vs. y) Residuals vs. predicted instead of residuals vs. x because it allows for considering the entire model (with all explanatory variables) at once Residuals randomly scattered in a band with a constant width around 0 (no fan shape) Also worthwhile to view absolute value of residuals vs. predicted to identify unusual observations easily

Part e

Cross validation for part b:

```
Linear Regression

395 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 315, 317, 315, 316, 317
Resampling results:

  RMSE      Rsquared   MAE
  2.049076  0.8400572  1.378047

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross validation for part e:

```
Linear Regression

395 samples
  3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 316, 316, 317, 315, 316
Resampling results:

  RMSE      Rsquared   MAE
  2.039993  0.8487241  1.382552

Tuning parameter 'intercept' was held constant at a value of TRUE
```

As it is obvious that in part e model, RMSE is less than part b which is good and a depiction of what a better model is.
Also, Rsquared metric increased in part e indicating better model was trained and result is sati sfying.

Question 6

Part A

Selected Response variable is Absences. I converted this numerical variable to categorical, it's value is 1 if No. of absences is more than 8, and it is 0 otherwise. Chosen explanatory variables that I thought could be a good fit for model are:

- G1, G2 and G3
- Failures
- Study time
- Gout
- romanticyes

Summary of Trained model is shown below:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.213011	0.138532	1.538	0.12496	
G1	-0.013034	0.013296	-0.980	0.32754	
G2	-0.052708	0.017006	-3.099	0.00208	**
G3	0.065857	0.012517	5.262	2.37e-07	***
failures	0.093195	0.040438	2.305	0.02172	*
studytime	0.002653	0.029536	0.090	0.92848	
goout	0.070616	0.022061	3.201	0.00148	**
romanticyes	0.067608	0.052132	1.297	0.19545	

$$\log\left(\frac{p}{1-p}\right) = 0.213011 - 0.013034 \times G1 - 0.052708 \times G2 + 0.065857 \times G3 + 0.093195 \times failures + 0.002653 \times study\ time + 0.070616 \times goout + 0.067608 \times romanticyes$$

intercept: keeping all other predictors zero, the log odds ratio / odds radio of absences is -0.213011 / exp(0.213011)

G1: keeping all other predictors zero by single unit increase in G1, the log odds ratio / odds radio of absences is -0.013034 / exp(-0.013034)

G2: keeping all other predictors zero by single unit increase in G2, the log odds ratio / odds radio of absences is -0.052708 / exp(-0.052708)

G3: keeping all other predictors zero by single unit increase in G3, the log odds ratio / odds radio of absences is $+0.065857$ / exp($+0.065857$)

failures: keeping all other predictors zero by single unit increase in failures, the log odds ratio / odds radio of absences is 0.093195 / exp(0.093195)

studytime: keeping all other predictors zero by single unit increase in studytime, the log odds ratio / odds radio of absences is 0.002653 / exp(0.002653)

goout: keeping all other predictors zero by single unit increase in gout, the log odds ratio / odds radio of absences is 0.070616 / exp(0.070616)

romantycies: keeping all other predictors zero by single unit increase in romantycies, the log odds ratio / odds radio of absences is 0.067608 / exp(0.067608)

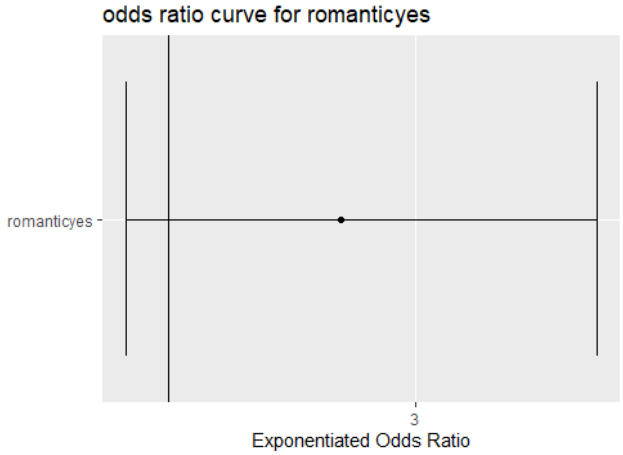
Part b

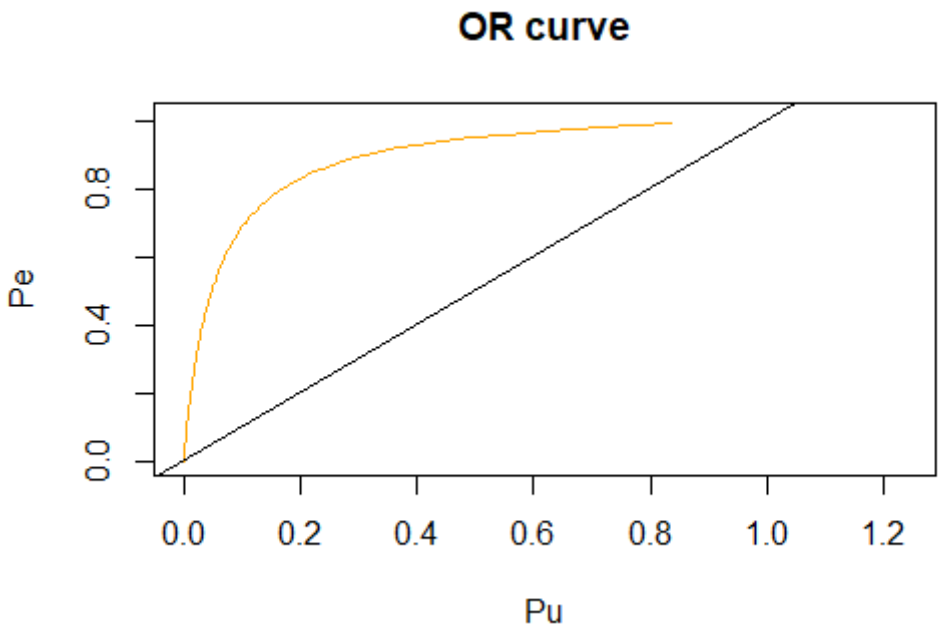
Choosing romanticyes as categorical variable for this part, we are to calculate odds ratio and then plot the curve:

LOR $CI \rightarrow (Estimate - Z^*_{95\%} \times std.Error, Estimate + Z^*_{95\%} \times std.Error)$

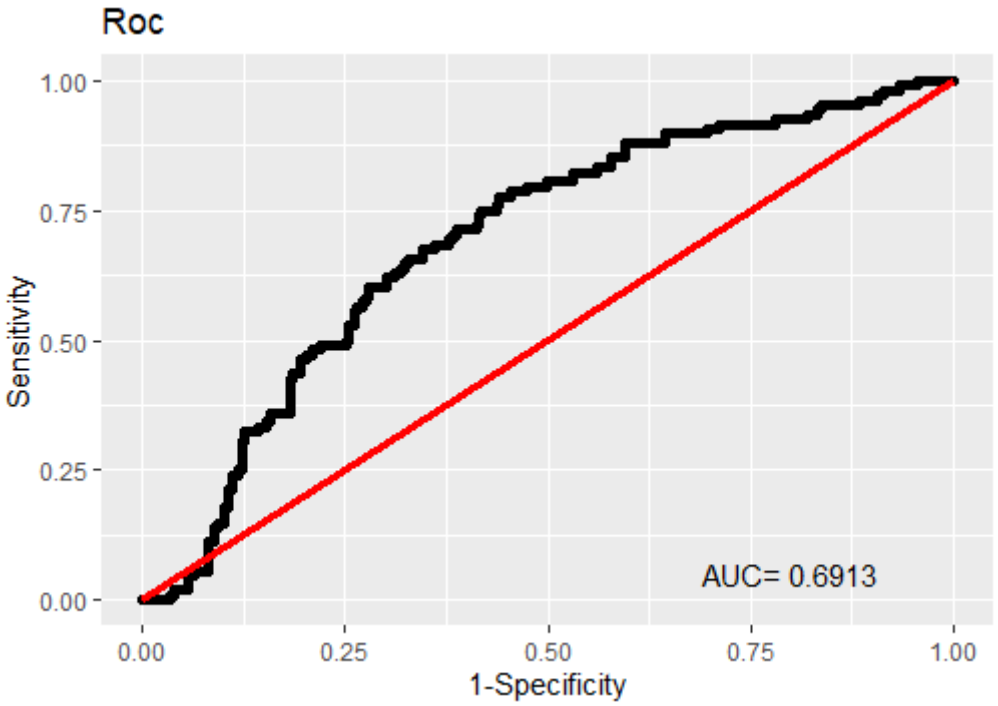
OR CI $(e^{0.067608-1.64\times0.052132},\ e^{0.067608+1.64\times0.052132}) = (0.9822706, 1.165447)$

OR curve for this categorical variable is:





Part c



Since AUC is 0.6913, model is not failed. It is a good model based on AUC. This plot illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Part d

Based on summary of model:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.24766	0.12630	1.961	0.05062	.
G1	0.01976	0.01212	1.630	0.10391	
G2	-0.05107	0.01550	-3.294	0.00108	**
G3	0.03282	0.01141	2.876	0.00425	**
failures	0.07831	0.03687	2.124	0.03430	*
studytime	-0.03539	0.02693	-1.314	0.18951	
goout	0.01039	0.02011	0.517	0.60564	
romanticyes	0.11438	0.04753	2.406	0.01658	*

Variables that have p-values less than 0.05, are significant. These variables are:

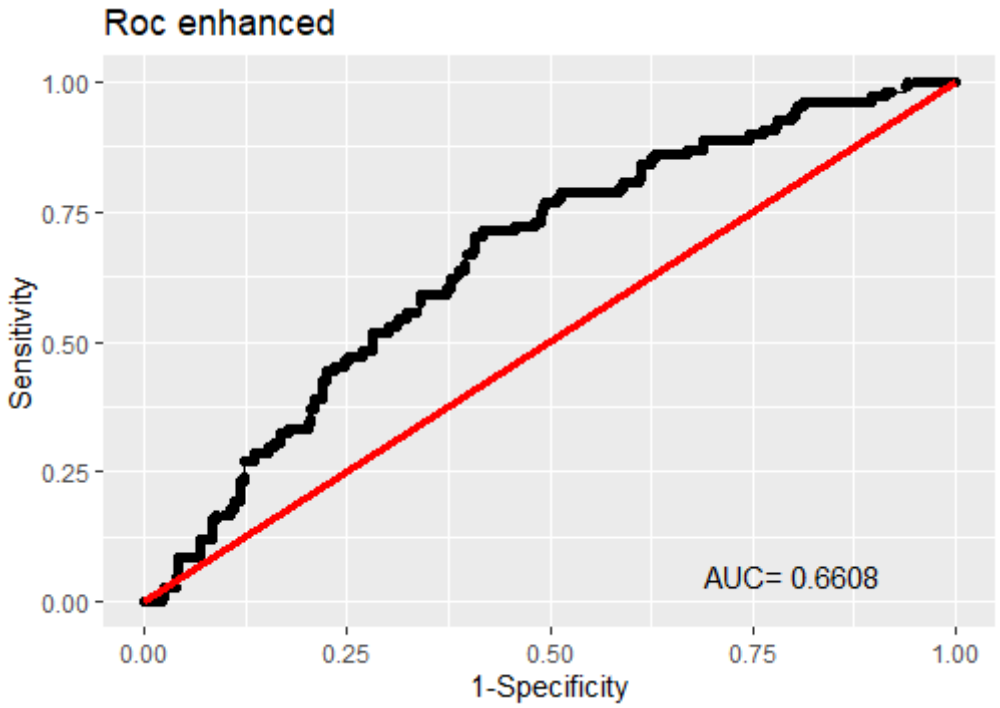
- G2
- G3
- Failures
- Romanticyes

Reconstructing the model, results will be:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25827    0.08830   2.925  0.00365 **
G2          -0.04106    0.01343  -3.058  0.00238 **
G3           0.03571    0.01135   3.147  0.00178 **
failures     0.08258    0.03652   2.261  0.02428 *
romanticyes  0.12013    0.04705   2.553  0.01106 *
```

As it is clear, all selected variables are significant. Despite the fact that, choosing most significant variables will state the model is significant with this variables, it is not guaranteed to have better model based AUC-ROC plot, which is shown below:

Part e



By omitting some variables, model's new AUC is less than before but as you can see ROC plot is completely above red line which is good.

New GLm summary:

```
Call:
glm(formula = Response ~ G2 + G3 + failures + romantic, data = StudentPerf
ormance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6261  -0.2830  -0.2141   0.5747   0.8529

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25827    0.08830   2.925  0.00365 **
G2          -0.04106    0.01343  -3.058  0.00238 **
G3           0.03571    0.01135   3.147  0.00178 **
failures     0.08258    0.03652   2.261  0.02428 *
romanticyes  0.12013    0.04705   2.553  0.01106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1911938)
```

```
    Null deviance: 78.471  on 394  degrees of freedom
Residual deviance: 74.566  on 390  degrees of freedom
AIC: 474.41

Number of Fisher Scoring iterations: 2
```

Question 7

First GLM model called:

```
Call:
glm(formula = Response ~ ., data = StudentPerformance)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.53347  -0.16125  -0.02439   0.12822   0.71991

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5727511  0.2528640   2.265  0.02408 *
X            -0.0003776  0.0001900  -1.988  0.04760 *
schoolMS      0.0712605  0.0458770   1.553  0.12120
sexM         -0.0108716  0.0261016  -0.417  0.67728
age           0.0193575  0.0154559   1.252  0.21120
Fjobhealth   -0.0075410  0.0784609  -0.096  0.92348
Fjobother     0.0556574  0.0556653   1.000  0.31803
Fjobservices  0.0170865  0.0574909   0.297  0.76648
Fjobteacher   0.1782529  0.0694316   2.567  0.01064 *
Mjobhealth   -0.0168924  0.0533220  -0.317  0.75157
Mjobother    -0.0237897  0.0373993  -0.636  0.52510
Mjobservices  0.0254472  0.0400474   0.635  0.52554
Mjobteacher  -0.0660271  0.0467440  -1.413  0.15863
goout         0.0180671  0.0108942   1.658  0.09808 .
internetyes   0.0651700  0.0340249   1.915  0.05621 .
romanticyes  -0.0228445  0.0261565  -0.873  0.38302
studytime    -0.0231084  0.0154218  -1.498  0.13487
failures      0.0504275  0.0208483   2.419  0.01605 *
health       -0.0048166  0.0087280  -0.552  0.58138
absences     -0.0043278  0.0015673  -2.761  0.00604 **
G1            0.0032185  0.0068741   0.468  0.63991
G2            0.0136471  0.0084542   1.614  0.10732
G3           -0.0702312  0.0062501 -11.237  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.05342459)

    Null deviance: 64.390  on 394  degrees of freedom
Residual deviance: 19.874  on 372  degrees of freedom
AIC: -11.882

Number of Fisher Scoring iterations: 2
```

Observing the output of GLM and writing a code to find variables that have p-values which are less than 0.05, final significant variables will be:

```
"Fjobteacher" "failures"    "absences"    "G3"
```

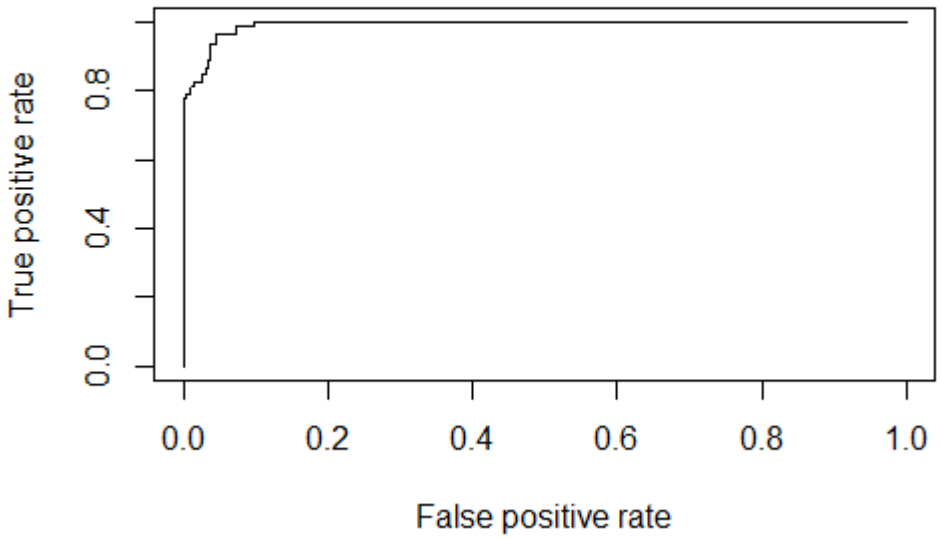
That Fjonteacher is a single level of whole Fjob column so, we should treat all of variables of a column completl ey same and we don't consider this variable. Then, final best model is :

```
"failures"    "absences"    "G3"
```

Second call for best GLM:

```
Call:
glm(formula = Response ~ failures + absences + G3, data = StudentPerforman
ce)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.52323  -0.16888  -0.02625   0.13642   0.71110

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.926730   0.042197  21.962  < 2e-16 ***
failures     0.072860   0.019664   3.705 0.000242 ***
absences    -0.004592   0.001489  -3.084 0.002189 **
G3          -0.056943   0.002837 -20.075  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.05562953)

    Null deviance: 64.390  on 394  degrees of freedom
Residual deviance: 21.751  on 391  degrees of freedom
AIC: -14.23

Number of Fisher Scoring iterations: 2
```

Best model ROC Curve:



Final precision:

Percision: 91.89873%

Rcode:

```
library(ROCR)
library(ggplot2)
library(Deducer)
library(GGally)
library(olsrr)
library(caret)
library(psych)


#Q1
##Part a
Q1.chosenvar <- table(StudentPerformance[,c("internet","Mjob")])
Q1_CI_calculator <- function(sel1, sel2, data){
  sel1.names <- names(table(data[,sel1]))
  print(sel1.names)
  sel2.names <- names(table(data[,sel2]))
  print(sel2.names)
  print("----------------")
  n.names1 <- length(sel1.names)
  print(n.names1)
  n.names2 <- length(sel2.names)
  print(n.names2)
  print("------------------")
  tab <- table(StudentPerformance[, c(sel1, sel2)])
  print(tab)
  CI <- c()
  NCI <- c()
  i=1
  j=1
  for(i in 1:(n.names1-1)){
```

```
    k <- (i+1)
    for(j in k:n.names1){
      n_1 <- sum(tab[sel1.names[i],])
      #print(tab[sel1.names[i],])
      p_1 <- tab[sel1.names[i],sel2.names[1]]/n_1
      n_2 <- sum(tab[sel1.names[j],])
      p_2 <- tab[sel1.names[j],sel2.names[1]]/n_2
      SE <- sqrt(p_1*(1-p_1)/n_1+p_2*(1-p_2)/n_2)
      delta.p <- p_1 - p_2
      CI <- c(CI, delta.p + c(1,-1)*pnorm(0.975, lower.tail = F)*SE)
      NCI <- c(NCI, sel1.names[i], sel1.names[j])
    }
  }
  return(data.frame(CI, NCI))
}
Q1_CI_calculator("Mjob", "sex", StudentPerformance)

#Part b
run_pchi <- function(sel1, sel2, data){
  tab <- table(StudentPerformance[, c(sel1, sel2)])
  sel1.names <- names(table(data[,sel1]))
  print(sel1.names)
  sel2.names <- names(table(data[,sel2]))
  print(sel2.names)
  print("----------------")
  n.names1 <- length(sel1.names)
  print(n.names1)
  n.names2 <- length(sel2.names)
  print(n.names2)
  print("------------------")
```

```
  x_2 <- 0
  for(i in 1:n.names1){
    for(j in 1:n.names2){
      expected <- sum(tab[sel1.names[i],])*sum(tab[,sel2.names[j]])/sum(tab)
      x_2 <- x_2 + (tab[i,j]-expected)**2/expected
    }
  }
  DF <- (n.names2-1)*(n.names1-1)
  print("X^2: ")
  print(x_2)
  print("DF: ")
  print(DF)
  print("P-value:")
  return(pchisq(x_2, DF, lower.tail = FALSE))
}

run_pchi("Mjob", "sex", StudentPerformance)

#Q2
set.seed(1)
small.sample <- StudentPerformance[sample(nrow(StudentPerformance), 13), ]$internet
table(small.sample)
simulation.output <- c()
no.simulation <- 10000
for(i in 1:no.simulation){
  mn <- mean(sample(0:1, 13, replace = TRUE))
  simulation.output <- c(simulation.output, mn)
}
pval <- length(simulation.output[simulation.output>=10/13])/no.simulation
p<-ggplot(data.frame(simulation.output), aes(x=simulation.output)) +
```

```
  geom_histogram(color="black", fill="lightsalmon", binwidth = .010)+
  geom_vline(aes(xintercept=10/13),
             color="orange", linetype="dashed", size=1)+
  ggtitle("Randomized distribution")+
  annotate("text", x = 11/13 , label = 'p-value =', y = 1500 , size = 3.4) +
  annotate("text", x = 0.97 , label = pval , y = 1500 , size = 3.4) +
  theme(plot.title = element_text(hjust = 0.5))
p


#Q3
#Part a

Q3.chosen.categvar <- StudentsPerformance$Fjob
N <- 100
table(Q3.chosen.categvar)
Q3.unbiasedsample <- sample(StudentsPerformance$Fjob,
                            N,
                            replace = FALSE)


Q3.biasedsample <- sample(StudentsPerformance$Fjob,
                          N,
                          prob = ifelse(Q3.chosen.categvar == "services", 0.8, 0.2))
ub.tab <- table(Q3.unbiasedsample)
b.tab <- table(Q3.biasedsample)


ub.tab
chisq.test(ub.tab,
           p = c(prop.table(table(Q3.chosen.categvar))))
b.tab
chisq.test(b.tab,
```

```
        p = c(prop.table(table(Q3.chosen.categvar))))

#Part b
#Q3.b.sampled <- StudentPerformance[sample(nrow(StudentPerformance), N), ]
q3.tab<-table(StudentPerformance[,c("Mjob","Fjob")])
chisq.test(table(StudentPerformance[,c("Mjob","Fjob")]))
sel <- c('at_home','services','other')
chisq.test(q3.tab[,sel])
q3.tab[,sel]

#Q4
##Part b
Q4.model1 <- lm(G3~G2, data = StudentPerformance)
Q4.model2 <- lm(G3~studytime, data = StudentPerformance)
leastsquare.first <- sum((Q4.model1$residuals)^2)
leastsquare.second <- sum((Q4.model2$residuals)^2)

Q4.model1$coefficients
Q4.model2$coefficients

my_graph <- ggplot(StudentPerformance, aes(x = G2, y = G3)) +
  geom_point(col='orange') +
  stat_smooth(method = "lm",
              col = "#C42126",
              se = FALSE,
              size = 1)+
  ggtitle("G3 Vs G2 scatter plot+linear regression line")

my_graph
```

```
my_graph <- ggplot(StudentPerformance, aes(x = studytime, y = G3)) +
  geom_point(col='orange') +
  stat_smooth(method = "lm",
              col = "#C42126",
              se = FALSE,
              size = 1)+
  ggtitle("G3 Vs Study time scatter plot+linear regression line")

my_graph

#Part d
anova(lm(G3~G2+studytime, data = StudentPerformance))
anova(lm(G3~studytime, data = StudentPerformance))
anova(lm(G3~G2, data = StudentPerformance))
#DF=0!!!!!!:|:|:|:|:|:|?

ars.second <- 1-(1-cor(StudentPerformance$studytime,
StudentPerformance$G3)^2)*(394/392)
ars.first <- 1-(1-cor(StudentPerformance$G2, StudentPerformance$G3)^2)*(394/392)

#Part f
samples <- StudentPerformance[sample(nrow(StudentPerformance), 100), ]
samples.train <- samples[1:90,]
samples.test <- samples[91:100,]
Q4.sampled.firstmodel <- lm(G3~G2, data = samples.train)
Q4.sampled.secondmodel <- lm(G3~studytime, data = samples.train)

predictsample<-function(intercept, slope, x, y){
  x.hat <- intercept + slope*x
  correct <- sum(x.hat == y)
```

```
  return(x.hat)
}
pred.G2 <- predictsample(Q4.sampled.firstmodel$coefficients[1],
            Q4.sampled.firstmodel$coefficients[2],
            samples.test$G2,
            samples.test$G3)
pred.st <- predictsample(Q4.sampled.secondmodel$coefficients[1],
            Q4.sampled.secondmodel$coefficients[2],
            samples.test$studytime,
            samples.test$G3)
actual <- samples.test$G3
pred.tb<- data.frame(pred.G2,pred.st, actual)
res.tb<- data.frame(abs(pred.G2-actual),abs(pred.st-actual))
#how to return percision????
SE <- 0.7277
m<-1.4323
c( m + c(1,-1)*pnorm(0.975, lower.tail = F)*SE)
SE <- 0.05126
m<-1.26504
c( m + c(1,-1)*pnorm(0.975, lower.tail = F)*SE)


#Q5
##Part a
featurePlot(x=temp[,1:8], y=temp[,8:16], plot="pairs")
G2 <- StudentPerformance$G2
G1 <- StudentPerformance$G1
G3 <- StudentPerformance$G3
absences <- StudentPerformance$absences
studytime <- StudentPerformance$studytime
goout <- StudentPerformance$goout
```

```
selected <- data.frame(G1,G2, absences,studytime,goout)
ggpairs(selected, title = "Correlogram")

##Part b
Q5.MLR <- lm(G3~G1+G2+absences+studytime+goout, data = StudentPerformance)
summary(Q5.MLR)

#Part e
forward.p <- ols_step_forward_p(Q5.MLR, details = TRUE)
plot(forward.p)

backward.p <- ols_step_backward_p(Q5.MLR, details = TRUE)
plot(backward.p)

Q5.MLR.enhanced <- lm(G3~G1+G2+absences+failures, data = StudentPerformance)
summary(Q5.MLR.enhanced)

#Part e
train_test <- trainControl(method = "cv", number = 5)
Q5.Partb.crossvalidation <- train(G3 ~ G1 + G2 + absences + studytime + goout,
                data = StudentPerformance,
                trControl = train_test,
                method = "lm")

Q5.Parte.crossvalidation <- train(G3 ~ G1 + G2 + absences,
                        data = StudentPerformance,
                        trControl = train_test,
                        method = "lm")
```

Q5.Partb.crossvalidation

Q5.Parte.crossvalidation

```
#Q6
##a
Response <- StudentPerformance$absences
N <- length(Response)
for (i in 1:N){
  if (Response[i]<8){
    Response[i] <- 0
  }else{
    Response[i] <- 1
  }
}
Q6.glm <- glm(Response~ G1 + G2 + G3 + failures + studytime + goout + romantic  , data = StudentPerformance )
summary(Q6.glm)

##b
df <- data.frame(boxLabels = c("romanticyes"),
                 boxOdds = c(1.069946),
                 boxCILow = c(0.9822706),
                 boxCIHigh = c(1.165447))

ggplot(data = df,
       mapping = aes(y = forcats::fct_inorder(f = rev(x = boxLabels)))) +
  geom_vline(xintercept = 1) +
  geom_point(mapping = aes(x = boxOdds)) +
  geom_errorbarh(mapping = aes(xmin = boxCILow,
```

```
                        xmax = boxCIHigh)) +
  coord_trans(x = scales::exp_trans()) +
  scale_x_continuous(breaks = log(x = 0.5 * (1:10)),
                     minor_breaks = NULL,
                     labels = (0.5 * (1:10))) +
  labs(x = "Exponentiated Odds Ratio",
       y = "") +
  ggtitle("odds ratio curve for romanticyes")


py <- function(x, mdl) {
  return ((abs(summary(mdl)$coefficients[3])*x/(1-x)) / (1 +
(abs(summary(mdl)$coefficients[3])*x/(1-x))))
}


Pu <- py(seq(0, 1.01, 0.01), Q6.glm)
Pe<-seq(0, 1.01, 0.01)
plot(Pu, Pe, type = "l", col = "orange") +
  abline(a=0, b=1)+
  title("OR curve")



#Part c
a<- rocplot(Q6.glm)
a$labels$title<- "Roc"
a


#Part d
summary(Q6.glm)


#Part e
```

```
Q6.glm.enhanced <- glm(Response~G2 + G3 + failures + romantic  , data =
StudentPerformance )

summary(Q6.glm.enhanced)

a<- rocplot(Q6.glm.enhanced)

a$labels$title<- "Roc enhanced"

a


#Part f




#Q7

StudentPerformance <- read.csv("F:\\Semester 8\\Statistical
inference\\Project\\P2\\StudentsPerformance.csv", header = TRUE)

Response <- StudentPerformance$G1 + StudentPerformance$G2 + StudentPerformance$G3

N <- length(Response)

for (i in 1:N){
  if (Response[i]<25){
    Response[i] <- 1
  }else{
    Response[i] <- 0
  }
}

Q8.glm <- glm(Response~.  , data = StudentPerformance )

summary(Q8.glm)

p.values <- coef(summary(Q8.glm))[,4]

significant.variables <- c()

column.names <- colnames(StudentPerformance)

for(i in 1:23){
  if(p.values[i]<0.05){
    significant.variables <- c(significant.variables, names(p.values)[i])
```

```
  }
}
significant.variables
Q8.glm.enhanced <- glm(Response~ failures + absences + G3, data = StudentPerformance )
summary(Q8.glm.enhanced)
fitted <- Q8.glm.enhanced$fitted.values
p <- (exp(fitted))/(1+exp(fitted))

pred <- prediction( p, Response)
perf <- performance(pred,"tpr","fpr")
plot(perf)
cutoffs <- data.frame(cut=perf@alpha.values[[1]], fpr=perf@x.values[[1]],
                      tpr=perf@y.values[[1]])
cutoffs <- cutoffs[order(cutoffs$tpr, decreasing=TRUE),]
subset.cutoff <- subset(cutoffs, fpr < 0.2)
head(subset.cutoff)
threshold <- subset.cutoff$cut[1]

predicted.labels <- c()
for(prediction in p){
  if(prediction>threshold){
    predicted.labels <- c(predicted.labels, 1)
  }else{
    predicted.labels <- c(predicted.labels, 0)
  }
}
counter<-0
for(i in 1:395){
  if(predicted.labels[i]==Response[i]){
    counter<-counter+1
```

```
  }
}
print(100*counter/395)
```