Statistical Inference
Course Project
Phase II

Amin Asadi Sarijalou

810196410

Spring 2021

University of Tehran

ECE Department

Contents

Question 1:

1.A

Chosen Categorical Variables:

1. Mjob: at_home
2. Fjob: at_home

First of all, we find the proportion of the mothers and fathers who work at home.

After that we find the standard error as the square root of the sum of the variances of the two group. Then the confidence interval is obtained.

```
# 1.A

n <- nrow(data)
n1 <- length(data$Fjob)
n2 <- length(data$Mjob)


p1 <- table(data$Mjob)[["at_home"]] / n
p2 <- table(data$Fjob)[["at_home"]] / n

p_diff <- p2 - p1

se = sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
me = se * qnorm(0.975)

CI <- p_diff + c(-me, me)

cat("confidence interval = ", "(", CI[1], ",", CI[2], ")")
```

```
confidence interval =  ( -0.1400032 , -0.05746512 )
```

Result:

The interpretation of this confidence interval is:

$$We\ are\ 95\%\ confident\ that\ this\ interval\ captures\ the\ real\ difference\ of\ two\ proportions.$$

If we want to perform hypothesis test with this using this confidence interval, the hypotheses would be:

$$H_0: The\ two\ proportions\ are\ equal.$$

$$H_A: The\ two\ proportions\ are\ different.$$

As we can see above, the confidence interval doesn't include 0 thus the null hypothesis is rejected in favor of the alternative, hence the difference of the two proportions is statistically significant.

1.B

In order to check the dependency between Fjob and Mjob, we can use chi-square independency test.

Checking Conditions for Chi-Square Test:

1. Independence:
    - The students are randomly sampled
    - size < 10 % of total number of students
    - Each student contributes to one cell in the table
2. Sample size:
    - Each particular scenario has at least 5 expected cases

```
35  # 1.B
36  chisq.test(table(data$Fjob, data$Mjob))
37
38  ```
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  table(data$Fjob, data$Mjob)
X-squared = 73.381, df = 16, p-value = 2.534e-09
```

Result:

As we can see, the p-value obtained by this test is nearly zero which means these two variables are statistically independent.

Question 2:

Hypothesis Test is as below:

$$H_0: half\ of\ the\ students\ are\ have\ romantic\ relationship \rightarrow p = 0.5$$

$$H_A: less\ than\ half\ of\ the\ students\ have\ romantic\ relationship \rightarrow p < 0.5$$

Checking the conditions:

1. Independence:

$$We\ can\ assume\ that\ the\ samples\ are\ independent.$$

2. Sample size / skew:

$$15 \times 0.5 = 7.5 < 10$$

So, we can't assume that the distribution of the sample proportions are, nearly normal (but we continue testing.)

At first, we use a sample of size 15:

```
# 2
set.seed(42)
SAMPLE_SIZE <- 15
p_hat <- sum(data[sample(1:n, SAMPLE_SIZE), ]$romantic == "yes") / SAMPLE_SIZE
cat("proportion in small sample =", p_hat)
```

```
proportion in small sample = 0.2666667
```

In this sample the proportion is about 27%

Now we use simulation. In order to do this, we toss a fair coin 15 times and calculate the proportion heads (we label head as success) and repeat this process 10000 times:

```r
n_sim <- 10e4
sim_dist <- replicate(n_sim,
                      mean(sample(c("yes", "no"),
                                  size=SAMPLE_SIZE,
                                  rep=T)=="yes"))

p_value <- mean(sim_dist > p_hat)

cat("p-vlaue =", p_value, '\n')
if(p_value > 0.05) {
  cat("We don't have enough evidence to reject H0.")
}
if(p_value <= 0.05) {
  cat("H0 is rejected in favor of HA.")
}
```

```
p-vlaue = 0.94191
We don't have enough evidence to reject H0.
```

Result:

As we can see in the p-value obtained by means of simulation is quite large and thus we can't claim that less than half of the population are in romantic relationship.

Question 3

3.A

I chose father's job (Fjob) as the categorical variable.

First of all, we find the probability distribution in population:

```
# 3.A
# probability distribution
expected_table <- table(data$Fjob) / n
print(round(expected_table, 2))
```

```
at_home    health     other services   teacher
   0.05      0.05      0.55     0.28      0.07
```

Then we take two samples where one of them is random and the other is biased. The biased sample is biased on teachers (it has more than usual teachers):

```
SAMPLE_SIZE = 100

random_sample <- data[sample(1:n, SAMPLE_SIZE), "Fjob"]
random_sample_table <- table(random_sample)
print(random_sample_table)

prb <- ifelse(data$Fjob=="teacher",0.9, 0.1)
biased_sample <- data[sample(n, SAMPLE_SIZE, prob = prb), "Fjob"]

biased_sample_table <- table(biased_sample)
print(biased_sample_table)
```

```
random_sample
 at_home   health    other services  teacher
       4        3       52       32        9
biased_sample
 at_home   health    other services  teacher
       3        2       52       18       25
```

Goodness of Fit:

Checking Conditions for Chi-Square Test:

1. Independence:
    - The students are randomly sampled
    - 100 > 10 % of total number of students, but we continue
    - Each student contributes to one cell in the table
2. Sample size:
    - Each particular scenario has at least 5 expected cases

We use chi-square test to test the goodness of fit:

For random sample:

```
chisq.test(random_sample_probs, p=exptected_probs)
```

```
Chi-squared approximation may be incorrect
        Chi-squared test for given probabilities

data:  random_sample_probs
X-squared = 1.8277, df = 4, p-value = 0.7674
```

For biased sample:

```
chisq.test(biased_sample_probs, p=exptected_probs)
```

```
Chi-squared approximation may be incorrect
        Chi-squared test for given probabilities

data:  biased_sample_probs
X-squared = 48.535, df = 4, p-value = 7.301e-10
```

Result:

As we can see, the p-value of the goodness of fit test for random sample is very high and greater than 0.05 so we can't reject null hypothesis.

But the p-value of the goodness of fit test for biased sample is nearly zero so we conclude that the distribution of the biased sample is different with the distribution of the population.

Therefore, both results match with our expectations.

3.B

I chose mother's job (Mjob) as the categorical variable. We want to see whether parents' job are dependent to eachother.

We use chi-square for this purpose.

```
# 3.B
chisq.test(table(data$Fjob, data$Mjob))
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  table(data$Fjob, data$Mjob)
X-squared = 73.381, df = 16, p-value = 2.534e-09
```

Result:

Because p-value is nearly zero, we can conclude that fathers and mothers are dependent. This is actually as I expected to be. Because more often people marry someone who has the same job as them (maybe their colleague).

Question 4

I chose the variables G1 as response variable and I think predicting its future value is meaningful.

I chose G2 and study time as explanatory variables because I believe they have high correlation with G1. I didn't choose G3 because it has high linear correlation with G2 and considering the fact that I have already chosen G2, G3 wouldn't help for prediction purpose.

4.A

I believe the variables G2 is the most significant variable since if a student has higher grade in G2, so that student is someone who studies for the exams therefore in general has high grades, hence that student have higher grade in G1 too.

4.B

a)

For G2:

```
model_by_G2<- lm(formula = G1 ~ G2, data)
print(summary(model_by_G2))
```

Summary:

```
Call:
lm(formula = G1 ~ G2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5525 -1.1545 -0.0471  1.0380 10.2153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78475    0.29527   6.045 3.49e-09 ***
G2           0.73313    0.02283  32.115  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.852 on 393 degrees of freedom
Multiple R-squared:  0.7241,    Adjusted R-squared:  0.7234
F-statistic:  1031 on 1 and 393 DF,  p-value: < 2.2e-16
```

For studytime:

```
model_by_study <- lm(formula = G1 ~ studytime, data)
print(summary(model_by_study))
```

Summary:

```
Call:
lm(formula = G1 ~ studytime, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6592 -2.7566 -0.0149  2.3726  8.2434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2732     0.4585  20.224  < 2e-16 ***
studytime     0.7417     0.2083   3.561 0.000415 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.47 on 393 degrees of freedom
Multiple R-squared:  0.03125,   Adjusted R-squared:  0.02879
F-statistic: 12.68 on 1 and 393 DF,  p-value: 0.0004154
```

b)

For G2:

predictive equation:

$$\widehat{G_1} = 1.78 + 0.73 \times G_2$$

Interpretation of parameters:

- Intercept: This means that when G_2 is zero, the expected value of G_1 is 1.78 on average.
- Slope of G_2: This means that a unit increase in G_2 causes the G_1 to be higher on average by about 0.73 points.

For studytime:

predictive equation:

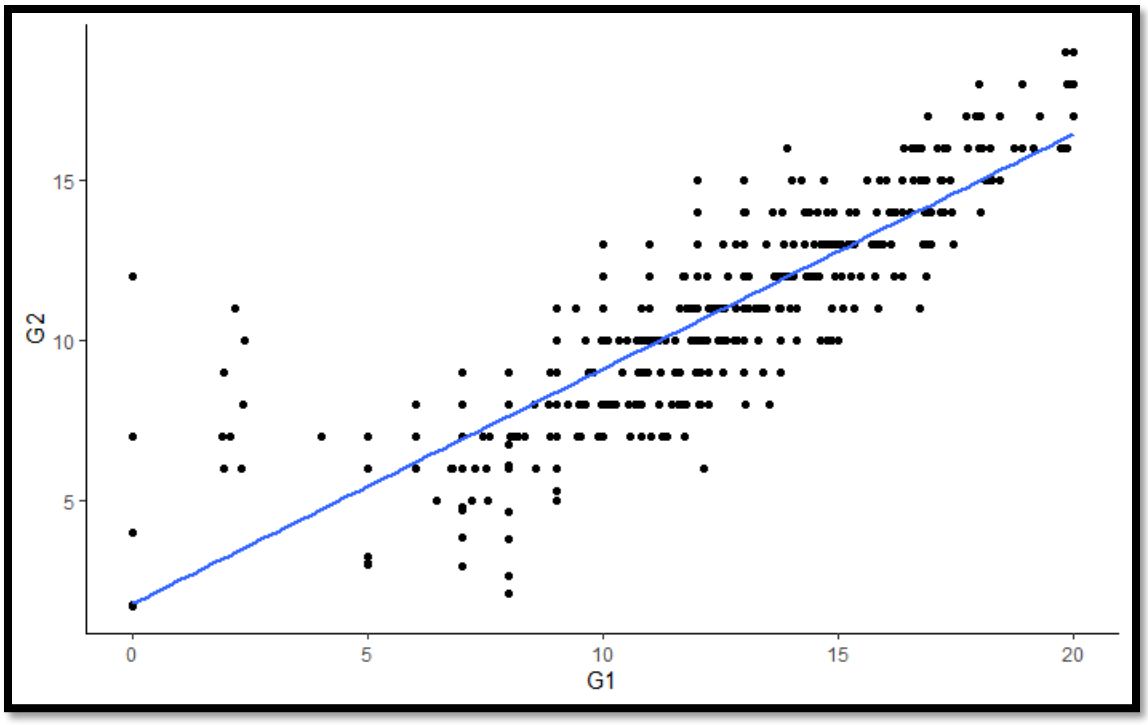
$$\widehat{G_1} = 9.27 + 0.74 \times studytime$$

Interpretation of parameters:

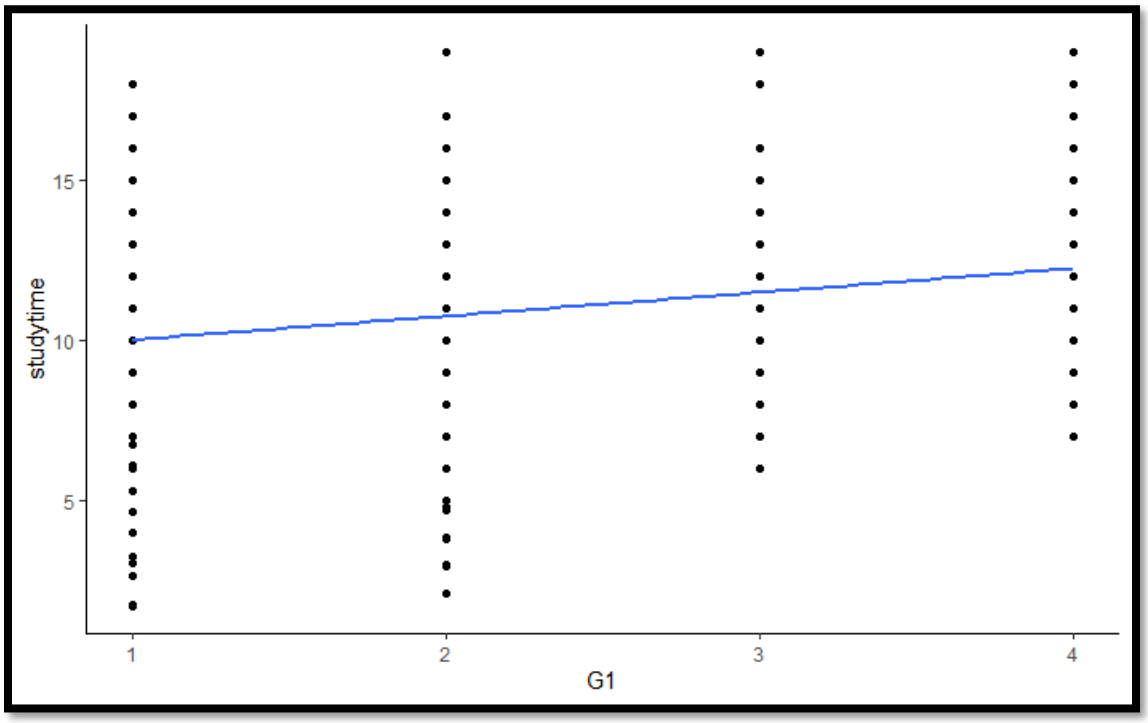
- Intercept: This means that if a student has not studies at all, their score is on average expected to be 9.27.
- Slope of studytime: This means that an hour increase in study time, causes the G_1 to be higher on average by about 0.74 points.

c)

For G2:



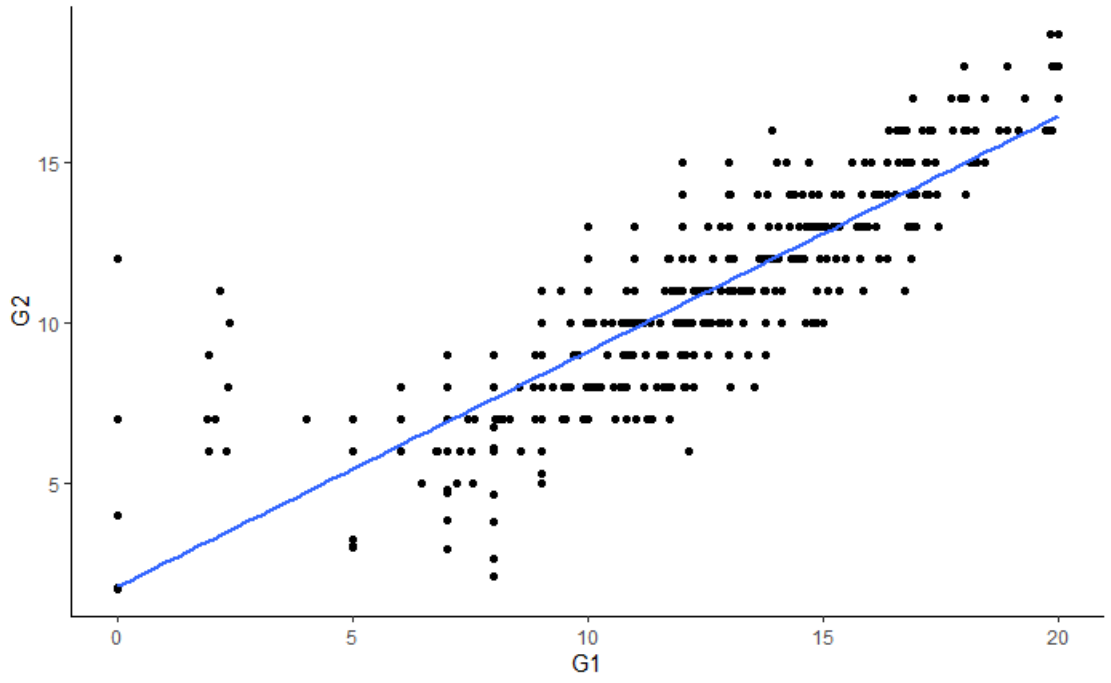
For studytime:



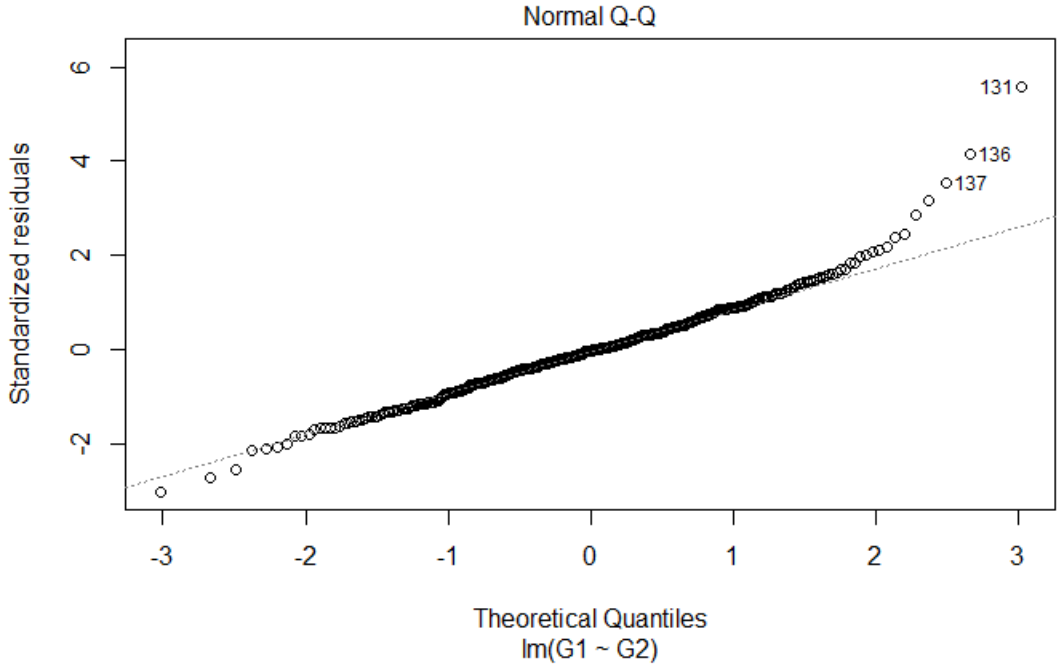
Checking Conditions:

We check 1.linearity 2.normal residuals 3.constant variability for both variables:

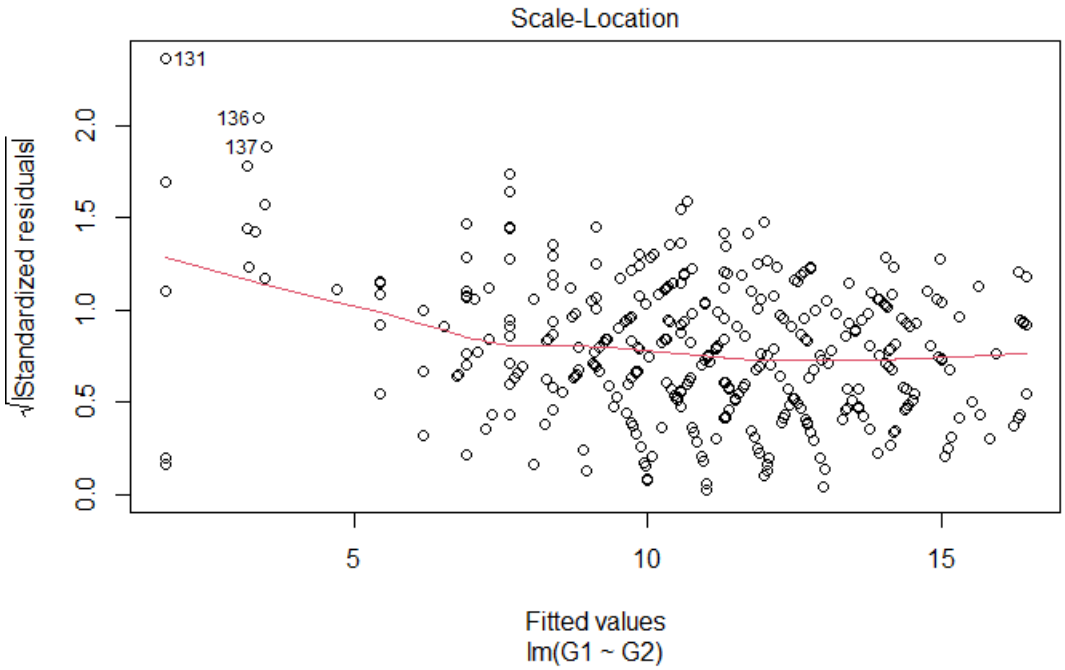
For G2:



---> linear

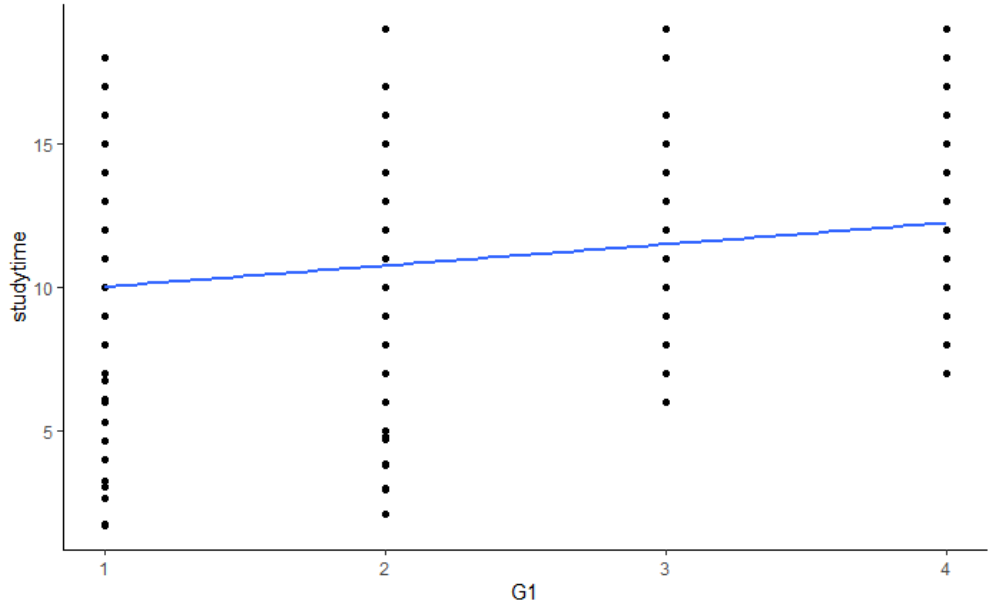


---> nearly normal

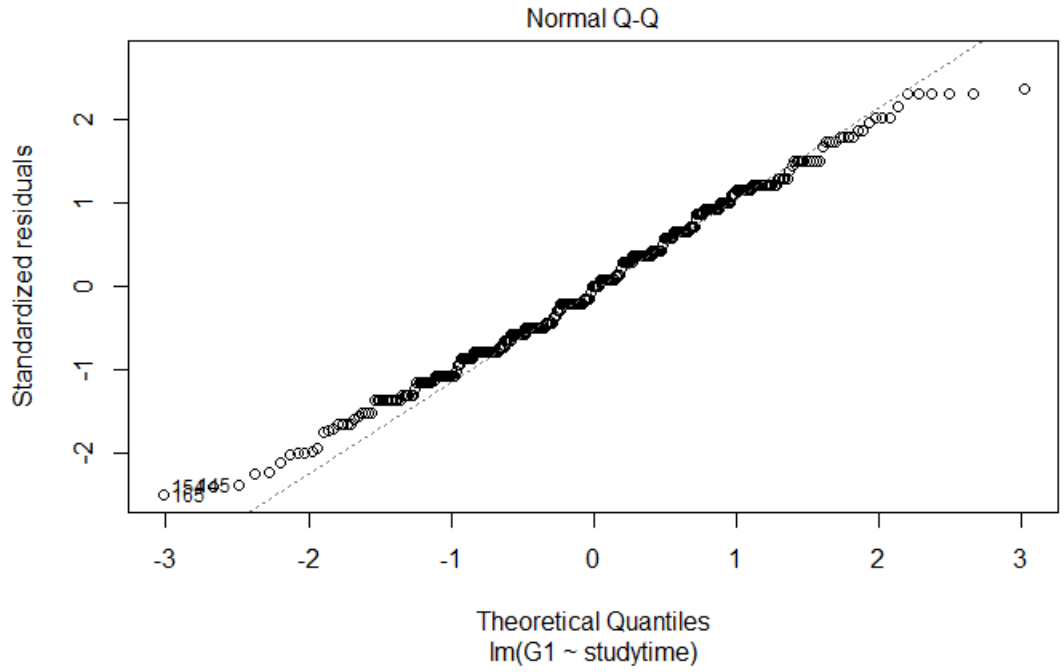


---> constant variability

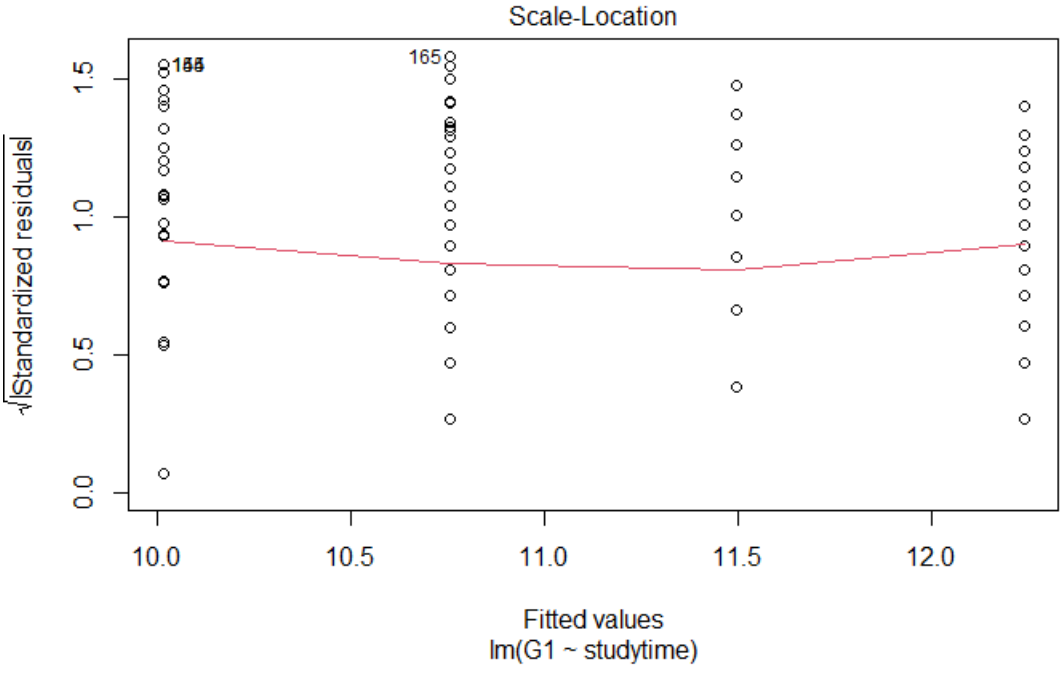
For studytime:



---> linear



---> nearly normal



---> constant variability

So, the conditions are met for both variables.

4.C

Because variable G2 has much more adjusted R^2, it explains the variability of G_1 a lot better than study time. Also, it's p-value it's is much smaller (nearly zero) and hence it is more significant when we consider the models separately. Therefore, it seems that G2 is a more significant predictor.

4.D

```
# 4.D

print(summary(model_by_study))

cat('###############################################\n\n')

print(summary(model_by_G2))
```

```
Call:
lm(formula = G1 ~ studytime, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6592 -2.7566 -0.0149  2.3726  8.2434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2732     0.4585  20.224  < 2e-16 ***
studytime     0.7417     0.2083   3.561 0.000415 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.47 on 393 degrees of freedom
Multiple R-squared:  0.03125,   Adjusted R-squared:  0.02879
F-statistic: 12.68 on 1 and 393 DF,  p-value: 0.0004154

###############################################


Call:
lm(formula = G1 ~ G2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5525 -1.1545 -0.0471  1.0380 10.2153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78475    0.29527   6.045 3.49e-09 ***
G2           0.73313    0.02283  32.115  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.852 on 393 degrees of freedom
Multiple R-squared:  0.7241,    Adjusted R-squared:  0.7234
F-statistic:  1031 on 1 and 393 DF,  p-value: < 2.2e-16
```

As we can see in the summary of the two models, the model with G2 has a much higher adjusted R-squared and hence it better explains the variability of G1.

```
print(anova(model_by_study))

cat('################################################\n\n')

print(anova(model_by_G2))
```

```
Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq F value    Pr(>F)
studytime   1  152.7 152.654  12.678 0.0004154 ***
Residuals 393 4732.1  12.041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
################################################

Analysis of Variance Table

Response: G1
           Df Sum Sq Mean Sq F value    Pr(>F)
G2          1 3537.0  3537.0  1031.4 < 2.2e-16 ***
Residuals 393 1347.7     3.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see in the two anova tables, although the p-value of the both models are significant, the model with G2 has a more significant pvalue (it is nearly zero).

So, in conclusion G2 is a better predictor of G1 than studytime.

4.E

A good predictor should have these characteristics:

- It should meet the conditions for linear regression:
    - Have linear relationship with the response variable
    - Residuals should be nearly normally distributed
    - The variability of points around the regression line should be roughly constant
- It should have a significant p-value
- It should well explain the variability of the response variable

4.F

a)

Fist we take sample of size 100 and then build the two models.

```
# 4.F
sample_100 <- data[sample(1:n, 100), ]

train.data <- sample_100[1:90, ]
test.data <- sample_100[91:100, ]

# 4.a
model_by_study_90<- lm(formula = G1 ~ studytime, train.data)
model_by_G2_90<- lm(formula = G1 ~ G2, train.data)
```

```
Call:
lm(formula = G1 ~ studytime, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4448 -1.6300 -0.4982  2.4578  8.5457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.3665     1.0005   8.362  8.4e-13 ***
studytime     1.0878     0.4921   2.210   0.0297 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.511 on 88 degrees of freedom
Multiple R-squared:  0.0526,    Adjusted R-squared:  0.04184
F-statistic: 4.886 on 1 and 88 DF,  p-value: 0.02967

################################################


Call:
lm(formula = G1 ~ G2, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7577 -1.3113  0.1321  0.8556  9.4376

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56245    0.66594   3.848 0.000225 ***
G2           0.66158    0.05269  12.557  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.159 on 88 degrees of freedom
Multiple R-squared:  0.6418,    Adjusted R-squared:  0.6377
F-statistic: 157.7 on 1 and 88 DF,  p-value: < 2.2e-16
```

We can see in the summary above that both explanatory variables are a significant predictor of the response variable.

b)

```r
# 4.b
pe.studytime = summary(model_by_study_90)$coefficients[2]
se.studytime = summary(model_by_study_90)$coefficients[4]
me.studytime = qnorm(.975) * se.studytime
CI.studytime = pe.studytime + c(-me.studytime, me.studytime)
cat("confidence interval of slope of studytime is =", CI.studytime)
```

```
> cat("confidence interval of slope of studytime is =", CI.studytime)
confidence interval of slope of studytime is = 0.1232817 2.052339
```

This means we are 95% confident that for each 1 hour increase in study time, the response variable(G1) increases on average by 0.12 to 2.05 points.

```
pe.G2 = summary(model_by_G2_90)$coefficients[2]
se.G2 = summary(model_by_G2_90)$coefficients[4]
me.G2 = qnorm(.975) * se.G2
CI.G2 = pe.G2 + c(-me.G2, me.G2)
cat("confidence interval of slope of G2 is =", CI.G2)
```

```
> cat("confidence interval of slope of G2 is =", CI.G2)
confidence interval of slope of G2 is = 0.5583159 0.7648428
```

This means we are 95% confident that for each 1 unit increase in G2, the response variable(G1) increases on average by 0.55 to 0.76 points.

c)

Here we perform prediction using two models on test data:

```
# 4.f.c
predictions.studytime <- model_by_study_90 %>% predict(test.data)
predictions.G2 <- model_by_G2_90 %>% predict(test.data)
```

d)

```
# 4.f.d
cat("RMSE of the prediction by studytime model = ", RMSE(predictions.studytime, test.data$G1),'\n')
cat("RMSE of the prediction by G2 model = ", RMSE(predictions.G2, test.data$G1), '\n')
```

```
RMSE of the prediction by studytime model =  3.906663
RMSE of the prediction by G2 model =  2.191077
```

Result

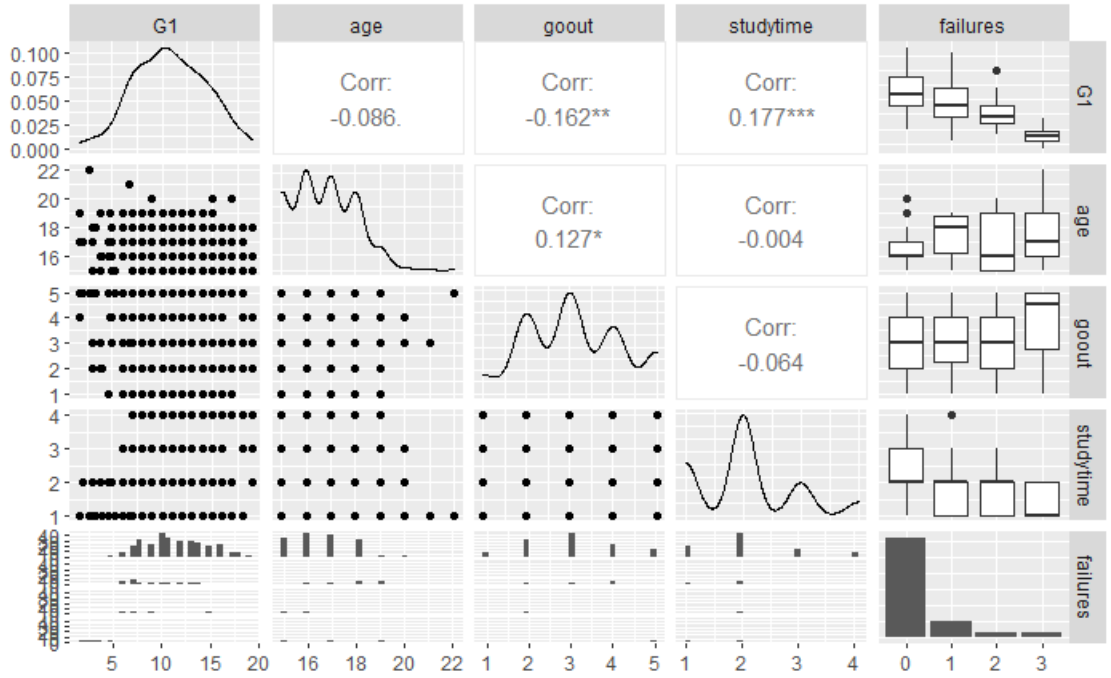
Root Mean Square Error is a popular and well enough metric for evaluating the performance of a linear regression prediction.

As we can see, the RMSE of the prediction by G2 is about 66% of the RMSE of the prediction by studytime. This is reasonable because as we stated before, G2 is in general a better predictor that studytime so it fits better and hence it has smaller RMS error.

Question 5

5.A

For better presentation, I plotted the correlogram in two plots (the explanatory variables are separated two groups, both of which include G1):



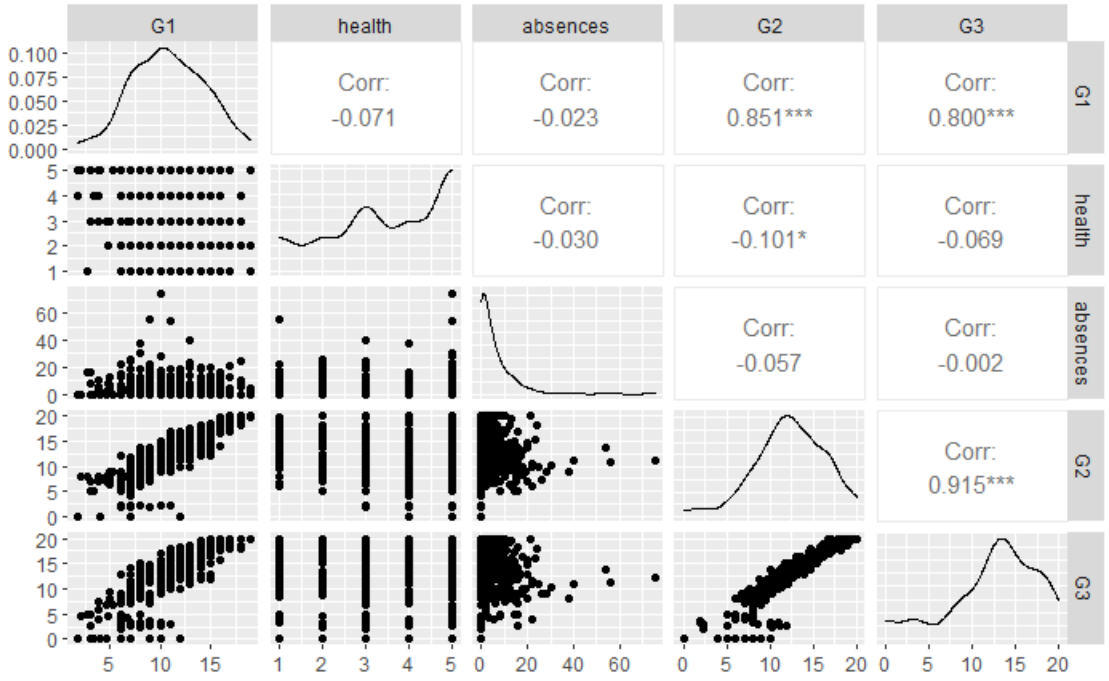
We see here that failures maybe highly correlated with G1. We further examine it:

```
cor(data$failures, data$G1)
```

```
[1] -0.4630003
```

The correlation between these variables is nearly 0.5 which is high. This is because the ones who have lower failures, have higher grades.

Here we see that G2 and G3 are much correlated with G1. This high correlation is because the students who have high grade in one course, are hard worker so in general they have high grades, hence the high correlation between G1 and G2 and G3.



So, we decide to pick G2, G3 and failures as our predictors.

5.B

```
mlr <- lm(formula = G1 ~ G2 + G3 + failures, data)
summary(mlr)
```

```
Call:
lm(formula = G1 ~ G2 + G3 + failures, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9610 -1.1728 -0.0933  1.0815 10.1912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.12189    0.36092   5.879 8.85e-09 ***
G2           0.63419    0.05663  11.198  < 2e-16 ***
G3           0.07354    0.04782   1.538    0.125
failures    -0.15656    0.15401  -1.017    0.310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.845 on 391 degrees of freedom
Multiple R-squared:  0.7277,    Adjusted R-squared:  0.7256
F-statistic: 348.2 on 3 and 391 DF,  p-value: < 2.2e-16
```

5.C

As we can see R^2 is about 72%. Therefore 72% of the variation in the response variable is explained by this model.

5.D

As we can see in the summary of the model, the p-value of the F-statistic is nearly zero which means that the model as a whole is good. Also, because G2 is a very significant predictor and the fact that 72% of the variability of the response variable is explained by this model, I conclude that the model fits the data well.

5.E

In order to find the best model, I used both “backward” and “forward” stepwise model selections methods by finding the best "$adjusted\ R^2$". For this purpose, I implemented a general function which:

- For forward selection, at each step it iterates over the remaining variables and for each it adds it to the model. Finally, it adds the variable which adding it increases the $adjusted\ R^2$ most. It repeats this process until $adjusted\ R^2$ is not increased or all the variables are included.
- For backward selection, at each step it iterates over the included variables and for each it removes it from the model. Finally, it removes the variable which removing it increases the $adjusted\ R^2$ most. It repeats this process until $adjusted\ R^2$ is not increased or all the variables are removed.

This is a helper function which takes response and explanatory variables and generates the formula for the model:

```
create_formula = function(response, vars) {
  as.formula(paste(response, paste(vars, collapse=" + "), sep=" ~ "))
}
```

And this is the main function which finds the model having greatest adjusted R^2 in a stepwise manner. It is given a parameter which specifies the direction of the steps (forward or backward):

```r
stepwise_selection = function(method) {
  full.vars <- c("school","sex","age","Fjob","Mjob","goout","internet","romantic","studytime","failures","health","absences","G2","G3"
  selected <- rep(0, length(full.vars))
  max_total_parameter <- 0
  for(i in seq(1, length(full.vars))) {
    remaining.vars <- full.vars[selected == 0]

    if(length(remaining.vars) == 0) {
      break
    }
    max_step_parameter <- 0
    max_step_var <- NULL
    for(j in seq(1, length(remaining.vars))) {
      if(method == "forward") {
        temp.vars <- c(full.vars[selected == 1], remaining.vars[j])
      }
      else if(method == "backward") {
        temp.vars <- setdiff(full.vars[selected == 0], remaining.vars[j])
      }
      formula <- create_formula("G1", temp.vars)
      model.temp.summary <- summary(lm(formula, data=data))
      if(model.temp.summary$adj.r.squared > max_step_parameter) {
        max_step_var <- remaining.vars[j]
        max_step_parameter <- model.temp.summary$adj.r.squared
      }
    }
    selected[which(full.vars == max_step_var)] = 1
    if(max_total_parameter < max_step_parameter) {
      max_total_parameter = max_step_parameter
    }
    else {
      break
    }
  }
  if(method == "forward")
    return(lm(create_formula("G1", full.vars[selected == 1]), data=data))
  else if(method == "backward")
    return(lm(create_formula("G1", full.vars[selected == 0]), data=data))
}
```

The best model obtained by forward and backward methods are stored in best_forward and best_backward variables, respectively.

```r
best_forward <- stepwise_selection(method="forward")
best_backward <- stepwise_selection(method="backward")
```

As we can see in the summary of the best_forward, the model has 11 variables and the adjusted R^2 of it is 0.7496.

```
summary(best_forward)
```

```
Call:
lm(formula = create_formula("G1", full.vars[selected == 1]),
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5966 -1.1936 -0.0871  1.0282  8.8232

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.05638    1.42265  -0.743  0.45822
sexM          0.34254    0.19448   1.761  0.07899 .
age           0.23640    0.07477   3.162  0.00169 **
Fjobhealth   -0.68310    0.59485  -1.148  0.25155
Fjobother    -0.88516    0.42073  -2.104  0.03605 *
Fjobservices -0.99938    0.43426  -2.301  0.02192 *
Fjobteacher   0.26422    0.52798   0.500  0.61706
Mjobhealth    0.11737    0.40385   0.291  0.77150
Mjobother    -0.52023    0.28265  -1.841  0.06648 .
Mjobservices  0.14808    0.30226   0.490  0.62448
Mjobteacher  -0.16313    0.35541  -0.459  0.64650
goout        -0.06643    0.08281  -0.802  0.42298
internetyes  -0.24216    0.25421  -0.953  0.34140
romanticyes   0.44815    0.19545   2.293  0.02240 *
studytime     0.23497    0.11541   2.036  0.04246 *
failures     -0.25161    0.15373  -1.637  0.10252
G2            0.60549    0.05529  10.951  < 2e-16 ***
G3            0.08723    0.04625   1.886  0.06005 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.762 on 377 degrees of freedom
Multiple R-squared:  0.7604,    Adjusted R-squared:  0.7496
F-statistic: 70.38 on 17 and 377 DF,  p-value: < 2.2e-16
```

As we can see in the summary of the best_backward, the model has 9 variables and the adjusted R^2 of it is 0.7498.

```
summary(best_backward)
```

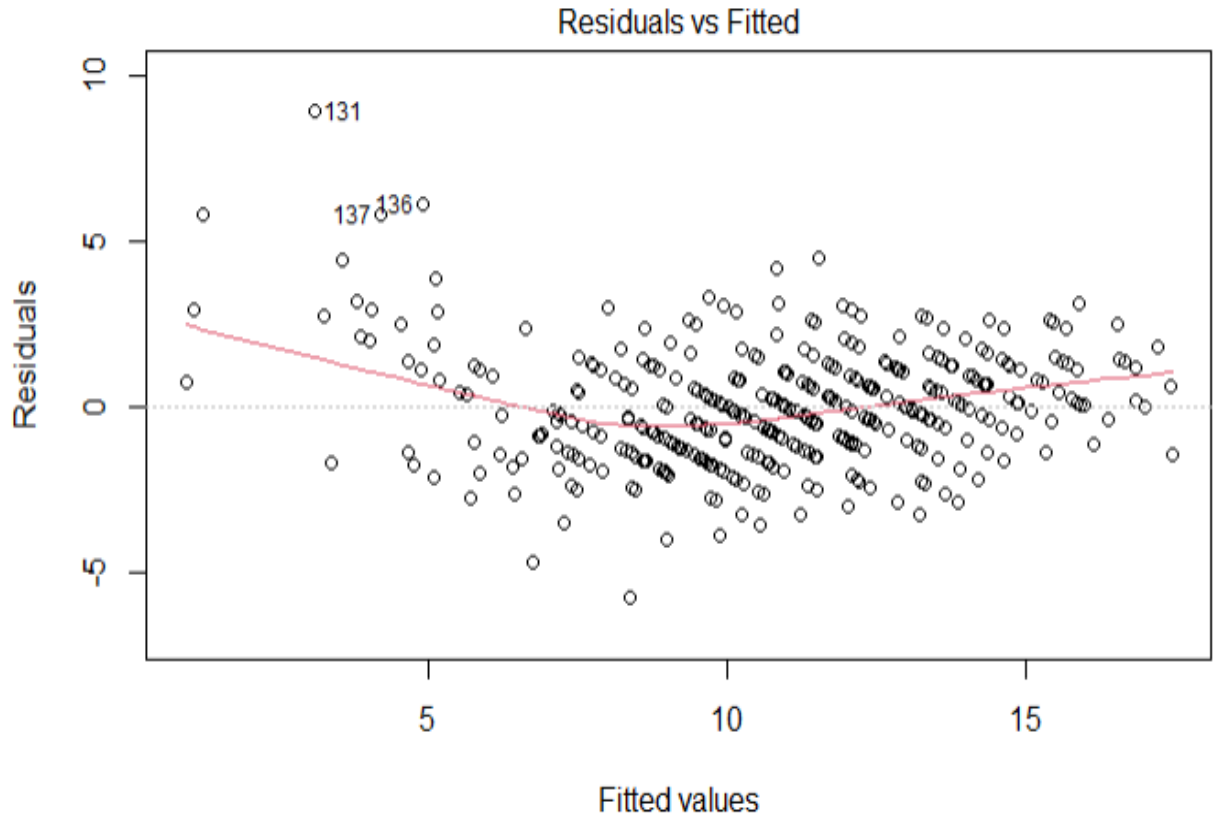
```
Call:
lm(formula = create_formula("G1", full.vars[selected == 0]),
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7186 -1.1896 -0.0835  1.0698  8.8874

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.41975    1.39278  -1.019   0.3087
sexM          0.32723    0.19384   1.688   0.0922 .
age           0.23749    0.07378   3.219   0.0014 **
Fjobhealth   -0.63760    0.59217  -1.077   0.2823
Fjobother    -0.88135    0.41966  -2.100   0.0364 *
Fjobservices -1.00216    0.43386  -2.310   0.0214 *
Fjobteacher   0.29469    0.52540   0.561   0.5752
Mjobhealth    0.01250    0.39515   0.032   0.9748
Mjobother    -0.56947    0.27890  -2.042   0.0419 *
Mjobservices  0.07675    0.29565   0.260   0.7953
Mjobteacher  -0.24944    0.34556  -0.722   0.4708
romanticyes   0.42823    0.19378   2.210   0.0277 *
studytime     0.22820    0.11516   1.982   0.0482 *
failures     -0.25709    0.15360  -1.674   0.0950 .
G2            0.60935    0.05492  11.095   <2e-16 ***
G3            0.08497    0.04618   1.840   0.0666 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.761 on 379 degrees of freedom
Multiple R-squared:  0.7593,    Adjusted R-squared:  0.7498
F-statistic:  79.7 on 15 and 379 DF,  p-value: < 2.2e-16
```

Result:

Adjusted R^2 of the backward method is slightly better. Furthermore, it used less variables (9 < 11) which is more efficient. Therefore, the model obtained by backward method is better.

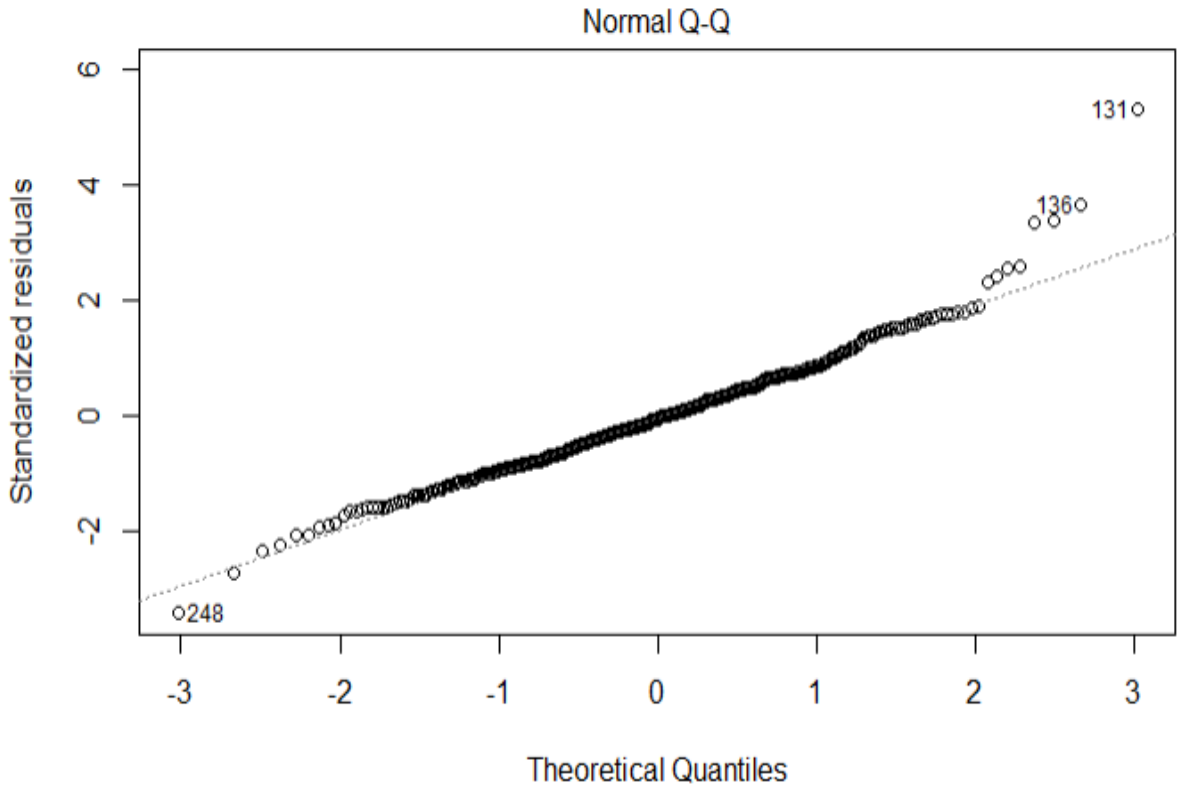
5.F

Linearity Check



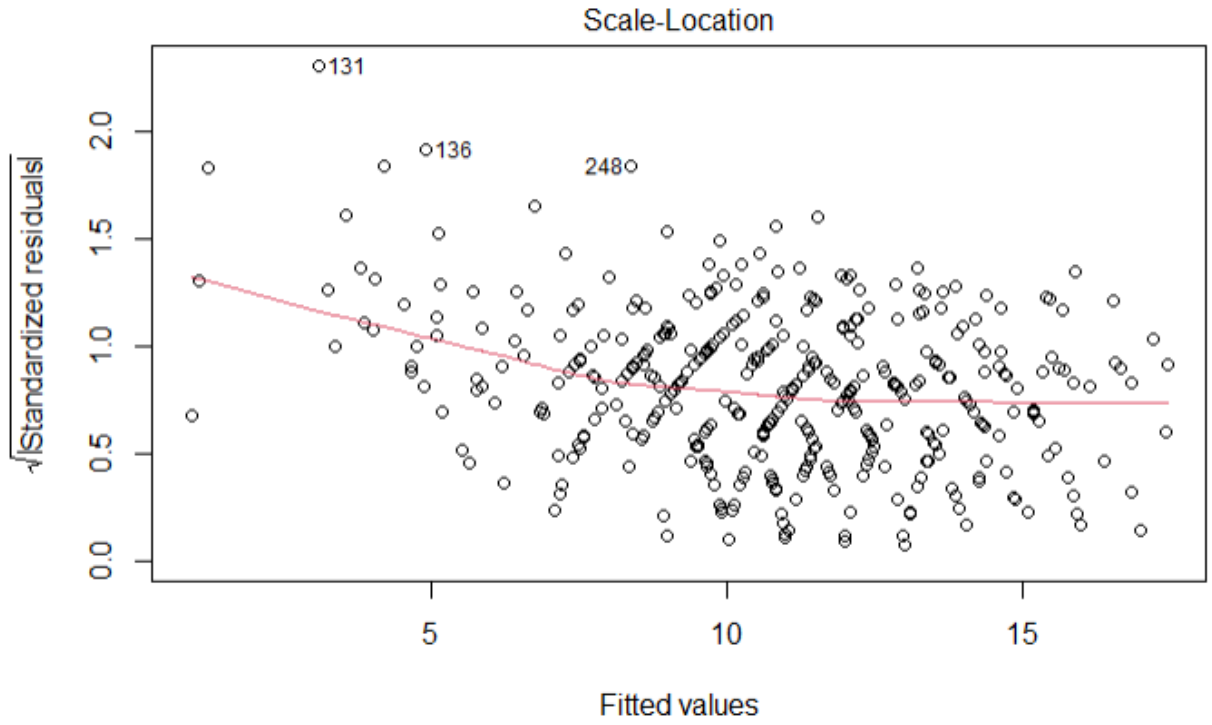
As we can see in the above plot, the residuals are nearly zero and also there cannot be seen any particular pattern in the residuals. So, this condition is held.

Nearly Normal Residuals Check QQ-Plot



As we can see in the QQ-Plot, except for the tails, the plot residuals distribution is very similar to normal distribution because the QQ-Plot of it fits the QQ-Plot of the normal distribution.

Constant Variability Check



Scale-Location plot is a plot which shows if residuals are spread equally along the ranges of predictors. It can be seen that except for the initial points, the variability (variances) of the residual points remains constant with the value of the fitted outcome variable.

5.G

Root Mean Squared Error (RMSE), measures the average prediction error of a model when in the process of prediction of an outcome for an observation. It is calculated as the average difference between predicted and actual label. When the RMSE is lower, model is better.

5-Fold cross validation for the model in part B:

```
cv.lm(mlr, k=5)
```

```
Mean absolute error          :  1.419079
Sample standard deviation    :  0.06614729

Mean squared error           :  3.4773
Sample standard deviation    :  0.6552709

Root mean squared error      :  1.85868
Sample standard deviation    :  0.1681104
```

5-Fold cross validation for the model in part E which we found using backward method:

```
cv.lm(best_backward, k=5)
```

```
Mean absolute error          :  1.360368
Sample standard deviation    :  0.04777727

Mean squared error           :  3.173767
Sample standard deviation    :  0.2661166

Root mean squared error      :  1.780268
Sample standard deviation    :  0.07428209
```

Result:

We can see that the model obtained by backward method has smaller RMSE which means it has less error and hence is better.

Question 6

6.A

I chose “romantic” as the response variable which is of type (binary) categorical.

Also, I guess the set of variables “school” + “sex” + “age” + “goout” + “internet” + “studytime” + “failures” + “absences” + “G1” might explain the response variable accurately.

In the code below, first I one-hot encode the response variable to be able to be used by glm. Then I split the dataset into train and test groups.

```
set.seed(42)
data["romanticYes"] = as.numeric(data["romantic"] == "yes")

train.size <- floor(2/3 * nrow(data))
train.ind <- sample(seq_len(nrow(data)), size = train.size)
train.data <- data[train.ind, ]
test.data <- data[-train.ind, ]
```

Then I use the glm function with binomial family to train a logistic regression classifier.

```
clf_part_A <- glm(formula = romanticYes ~ school + sex + age + goout + internet + studytime + failures + absences + G1,
                  data=train.data,
                  family=binomial)
summary(clf_part_A)
```

```
Call:
glm(formula = romanticYes ~ school + sex + age + goout + internet +
    studytime + failures + absences + G1, family = binomial,
    data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5683  -0.8954  -0.6952   1.2443   2.0723

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.00646    2.08826  -2.397   0.0165 *
schoolMS     0.21212    0.44433   0.477   0.6331
sexM        -0.21450    0.29387  -0.730   0.4654
age          0.20076    0.12330   1.628   0.1035
goout       -0.07903    0.12505  -0.632   0.5274
internetyes  0.63338    0.41003   1.545   0.1224
studytime    0.27294    0.18370   1.486   0.1373
failures     0.26353    0.22007   1.197   0.2311
absences     0.03491    0.01857   1.880   0.0600 .
G1          -0.01356    0.04461  -0.304   0.7611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.45  on 262  degrees of freedom
Residual deviance: 311.63  on 253  degrees of freedom
AIC: 331.63

Number of Fisher Scoring iterations: 4
```

As we can see in the summary of the model, the variables sex, age and internet are the most significant among all.

The intercept is the log odds of being in romantic relationship when all explanatory variables are zero.

For each categorical explanatory variable, the estimate is the log odds ratio between the given level and the reference level when the other variables remain constant.

For each numerical explanatory variable, the estimate is how much the log odds ratio change when this variable increases 1 unit and other variables remain constant.

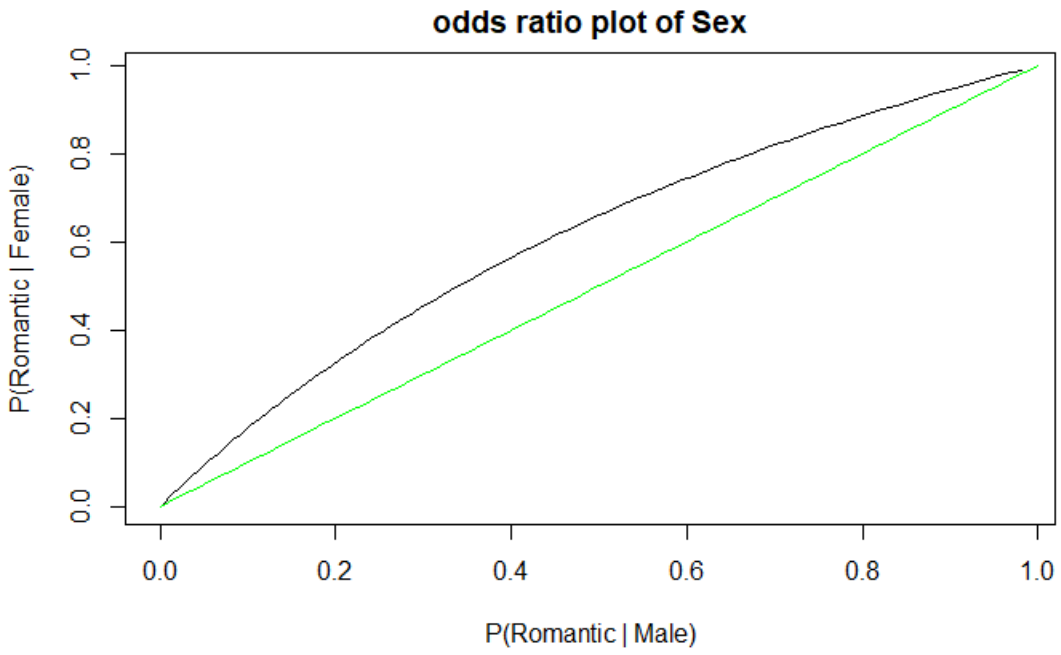
6.B

For each probability of P(Romantic | Female) we find the P(Romantic | Male) and then plot the curve.

```
# 6.B
# refrence level = male

probFemale <- seq(0, 0.99, 0.01)

OR_RATIO = abs(summary(classifier)$coefficients[3])
cat("OR RATIO =", OR_RATIO)
getY <- function(x) {
  return ((OR_RATIO*x/(1-x)) / (1 + (OR_RATIO*x/(1-x))))
}
probMale <- sapply(probFemale, getY)
```

```
OR RATIO = 0.4299028
```

```
plot(probMale, probFemale, type = "l", lty = 1,
     main="odds ratio plot of Sex",
     xlab = "P(Romantic | Male)",
     ylab = "P(Romantic | Female)")
lines(seq(0, 1, 0.01), seq(0, 1, 0.01), col="green")
```



The curve shows how the probability of “having romantic if sex is male” changes when we increase the “probability of having romantic if sex is female”.

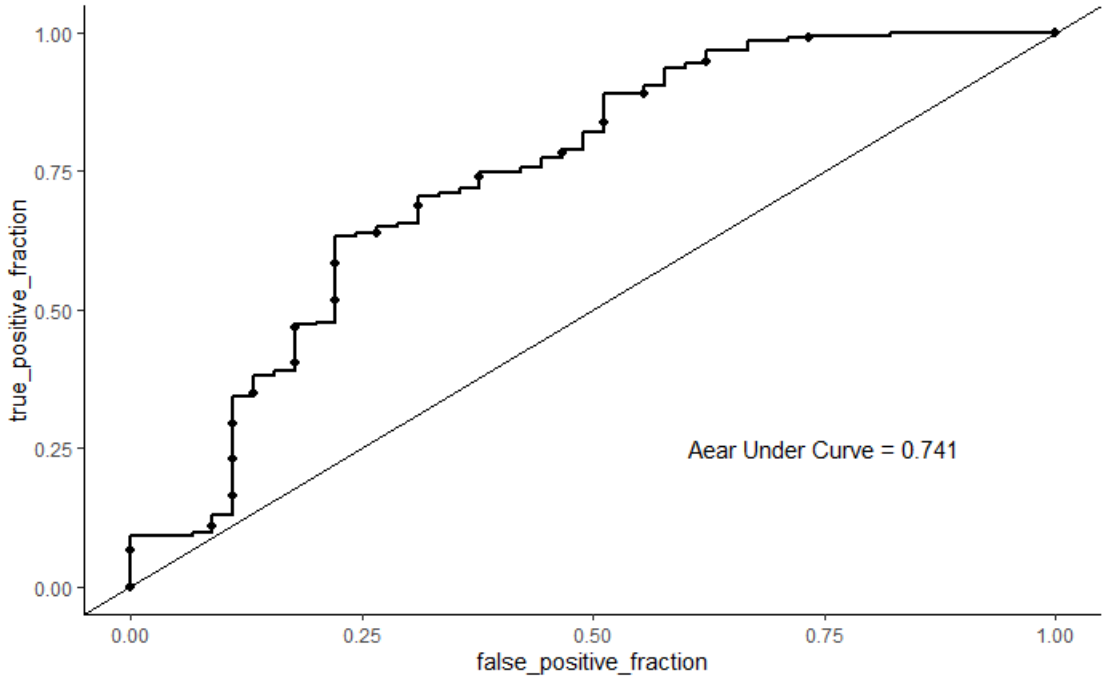
It can be seen that the curve is close to x=y line. This is because the OR Ratio = 0.42 and this is close to 1.

If OR Ratio were higher, the curve would be higher than this. If it were 1, it would be x=y line.

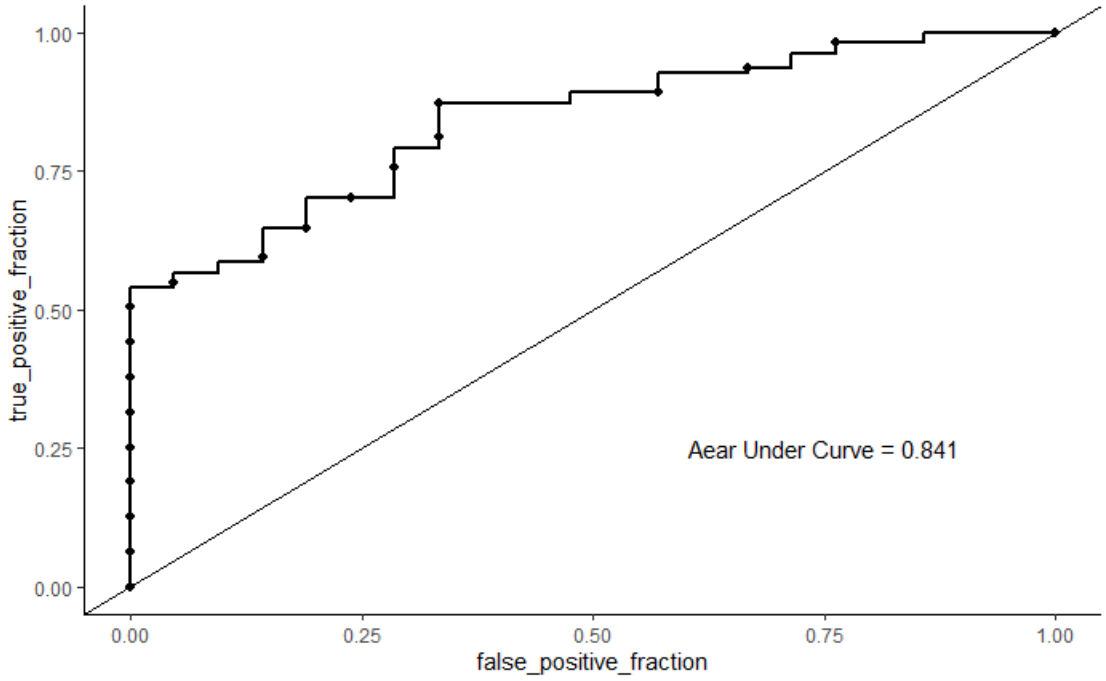
6.C

```
library(plotROC)
library(ggplot2)

train.data$prediction <- predict(classifier, newdata=train.data)

roc_curve <- ggplot(train.data,
                    aes(m = prediction,
                        d = romanticYes)) +
  geom_roc(n.cuts=20,
           labels=F) +
  theme_classic() +
  geom_abline(slope=1, intercept = 0)


show(roc_curve + annotate("text", x = .75, y = .25 , label =
                          paste("Aear Under Curve =",
                                round(calc_auc(roc_curve)["AUC"], 3))))

test.data$prediction <- predict(classifier, newdata=test.data)

roc_curve <- ggplot(test.data,
                    aes(m = prediction,
                        d = romanticYes)) +
  geom_roc(n.cuts=20,
           labels=F) +
  theme_classic() +
  geom_abline(slope=1, intercept = 0)

show(roc_curve + annotate("text", x = .75, y = .25 , label =
                          paste("Aear Under Curve =",
                                round(calc_auc(roc_curve)["AUC"], 3))))
```

ROC Curve For Train Data



ROC Curve For Test Data



We see that AUC for the train data is about 0.74 and for test data is about 0.84.

These values are acceptable specially for test data. The fact the AUC of the model is above 80% on test data indicates that the model has a good generalization.

6.D

As we saw in the summary of the model in part A, variable “absences” has the most significant p-value, hence having most importance role. This is maybe because the students who are in romantic relationship have higher tendency to be absent from classes.

```
clf_part_A <- glm(formula = romanticYes ~ school + sex + age + goout + internet + studytime + failures + absences + G1,
                  data=train.data,
                  family=binomial)
summary(clf_part_A)
```

```
Call:
glm(formula = romanticYes ~ school + sex + age + goout + internet +
    studytime + failures + absences + G1, family = binomial,
    data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5683  -0.8954  -0.6952   1.2443   2.0723

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.00646    2.08826  -2.397   0.0165 *
schoolMS     0.21212    0.44433   0.477   0.6331
sexM        -0.21450    0.29387  -0.730   0.4654
age          0.20076    0.12330   1.628   0.1035
goout       -0.07903    0.12505  -0.632   0.5274
internetyes  0.63338    0.41003   1.545   0.1224
studytime    0.27294    0.18370   1.486   0.1373
failures     0.26353    0.22007   1.197   0.2311
absences     0.03491    0.01857   1.880   0.0600 .
G1          -0.01356    0.04461  -0.304   0.7611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.45  on 262  degrees of freedom
Residual deviance: 311.63  on 253  degrees of freedom
AIC: 331.63

Number of Fisher Scoring iterations: 4
```

6.E

I chose the 4 most significant variable having p-value less than 0.2

```
clf_part_E <- glm(formula = romanticYes ~ age + absences + studytime + internet,
              data=train.data,
              family=binomial)
summary(clf_part_E)
```

```
Call:
glm(formula = romanticYes ~ age + absences + studytime + internet,
    family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5332  -0.8854  -0.7049   1.2616   1.9762

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.31767    1.88694  -3.348 0.000814 ***
age          0.26571    0.10627   2.500 0.012405 *
absences     0.03473    0.01789   1.941 0.052205 .
studytime    0.26613    0.16730   1.591 0.111669
internetyes  0.46466    0.39096   1.189 0.234630
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.45  on 262  degrees of freedom
Residual deviance: 314.97  on 258  degrees of freedom
AIC: 324.97

Number of Fisher Scoring iterations: 4
```

Result

It can be seen that the AIC has decreased from 331 to 324 which is because we reduced number of explanatory variables from 9 to 4. This decrease is acceptable because it is small and the new model is much smaller and hence more efficient.

Also, we see that the p-values have changed and now the most significant variable is age. This has happened because the p-values that are shown in the summary of a model are relative to that model and dependent to the whole set of explanatory variables.

The fact that age is a very significant predictor (p-value is about 0.01) is because students often get into romantic relationships in certain ages.

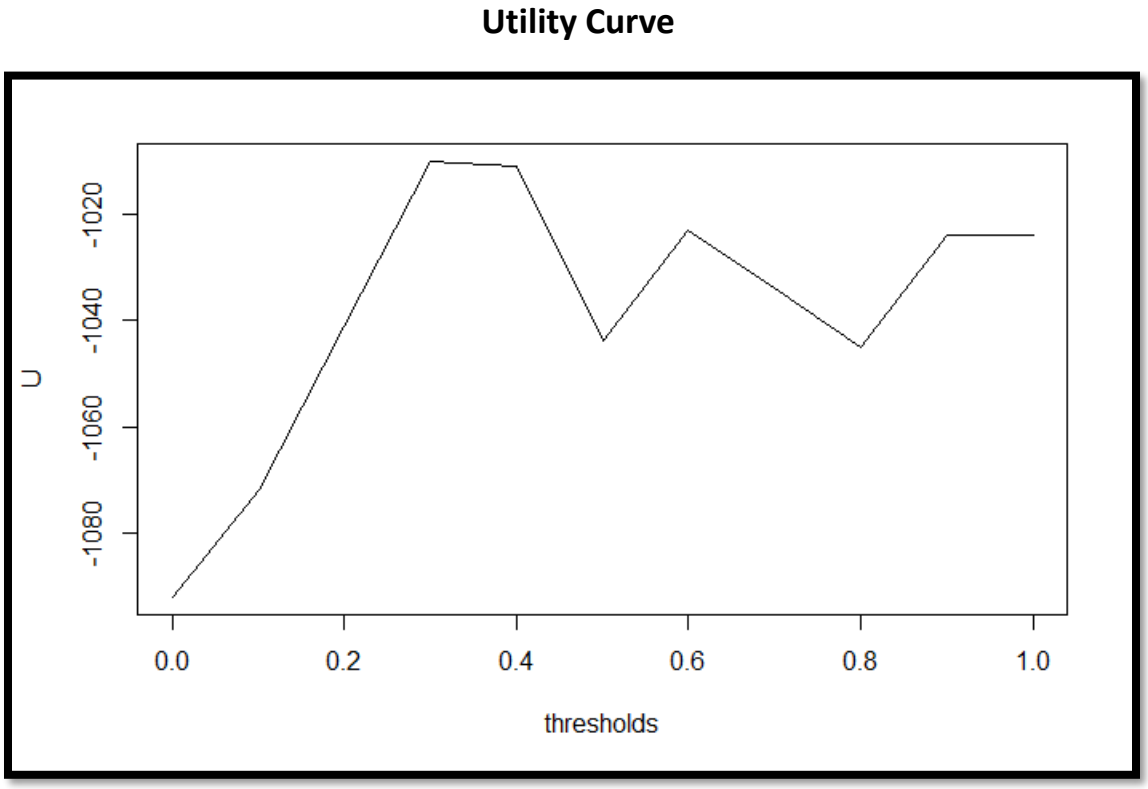
6.F

For all thresholds between 0 and 1 with steps of 0.1, I find the values: TP, TN, FP, FN and then calculate a utility based on them.

In this context, I think TP and TN has equally valuable but FP must be small because I don't want the students who are single to be classified as in relationship. Also, I don't want the reversal of this to happen either but it has a less important to me. Therefore, I assigned coefficients 1, 1, -20 and -10 respectively to these metrics.

```
plot_roc(clf_part_E)
```
```{r}
thresholds = seq(0, 1, 0.1)
U <- c()

for(i in seq(1, length(thresholds))) {
  predict_probs <- predict(clf_part_E, newdata=data)
  t <- confusionMatrix(data = as.factor(as.numeric(predict_probs > thresholds[i])),
                       reference=as.factor(data$romanticYes))$table

  TP = t[4]
  TN = t[1]
  FP = t[2]
  FN = t[3]
  U = c(U, TP + TN -20*FP - 10*FN)
}

plot(thresholds, U, type='l')
```

Result




```
max_th = thresholds[which(U == max(U))]

cat("Maximum Unitility happens when threshold =", max_th)
```

```
Maximum Unitility happens when threshold = 0.3
```

As we can see in the plot, the best threshold is 0.3.

Question 7

```
aca_prob_model <- glm(formula = academic_probation ~ school+sex+
                        age+Fjob+Mjob+goout+internet+romantic+
                        studytime+failures+health++absences,
              data=data,
              family=binomial)
summary(aca_prob_model)
```

```
Call:
glm(formula = academic_probation ~ school + sex + age + Fjob +
    Mjob + goout + internet + romantic + studytime + failures +
    health + +absences, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0923  -0.5442  -0.3376  -0.1423   3.2015

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.20806    2.67882  -2.317 0.020478 *
schoolMS     -0.27907    0.49832  -0.560 0.575463
sexM         -0.86234    0.37230  -2.316 0.020544 *
age           0.26653    0.15260   1.747 0.080710 .
Fjobhealth   -0.74944    1.39009  -0.539 0.589796
Fjobother     0.64252    0.73647   0.872 0.382973
Fjobservices  0.05696    0.76558   0.074 0.940689
Fjobteacher   1.19343    0.89879   1.328 0.184238
Mjobhealth   -1.94894    0.91234  -2.136 0.032664 *
Mjobother    -0.43731    0.46097  -0.949 0.342784
Mjobservices -0.59640    0.50373  -1.184 0.236430
Mjobteacher  -0.76851    0.66888  -1.149 0.250581
goout         0.44245    0.15278   2.896 0.003780 **
internetyes   0.31390    0.44338   0.708 0.478960
romanticyes   0.18694    0.34582   0.541 0.588800
studytime    -0.64080    0.24535  -2.612 0.009009 **
failures      1.96794    0.28984   6.790 1.12e-11 ***
health        0.00983    0.11999   0.082 0.934710
absences     -0.10901    0.03198  -3.409 0.000652 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 400.80  on 394  degrees of freedom
Residual deviance: 258.54  on 376  degrees of freedom
AIC: 296.54

Number of Fisher Scoring iterations: 6
```

Result:

It appears that variables: sex, Mjob, goout, studytime, failures and absences are significant predictors.

Among these, the variable “failures” is the most effective with p-value nearly equal to zero. This is reasonable and definitely matches my expectations because when

a student has many failures, they have very low grades and hence they are more probable to be an academic probation.