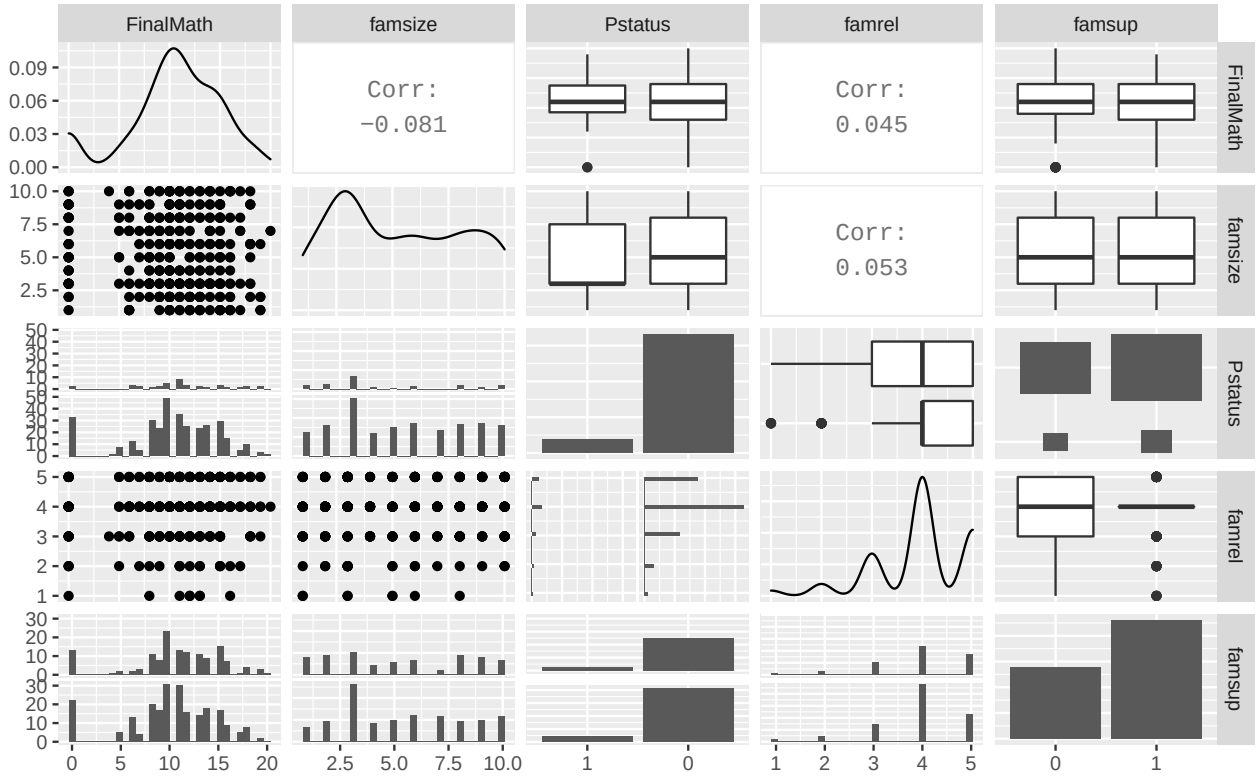
MVA Project - Multiregression Analysis

Chun-Jung Chen & Akshay Arora

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

Scotter Plot

```
GGally::ggpairs(data = d3.q1)
```



Fitting model 1

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

$$FinalMath = \beta_0 + \beta_1\ famsize + \beta_2\ Pstatus + \beta_3\ famrel + \beta_4\ famsup + \epsilon_j \quad j = 1, ..., n$$

After fitting, the model is

$$FinalMath = 11.43761 - 0.17823\ famsize - 0.67198\ Pstatus + 0.24481\ famrel - 0.44791\ famsup \quad j = 1, ..., n$$

- The result showed that only the intercet and the family size siginificantly impact final math score.
- The model performance is quite poor since the adjusted R-squared value is very low, indicating that the model does not explain the majority of data.
- The model may lack of other variables and may need further research or collect more data.

```
mod1 = lm(FinalMath ~ famsize + Pstatus + famrel + famsup, data = d3.q1)
summary(mod1)
```

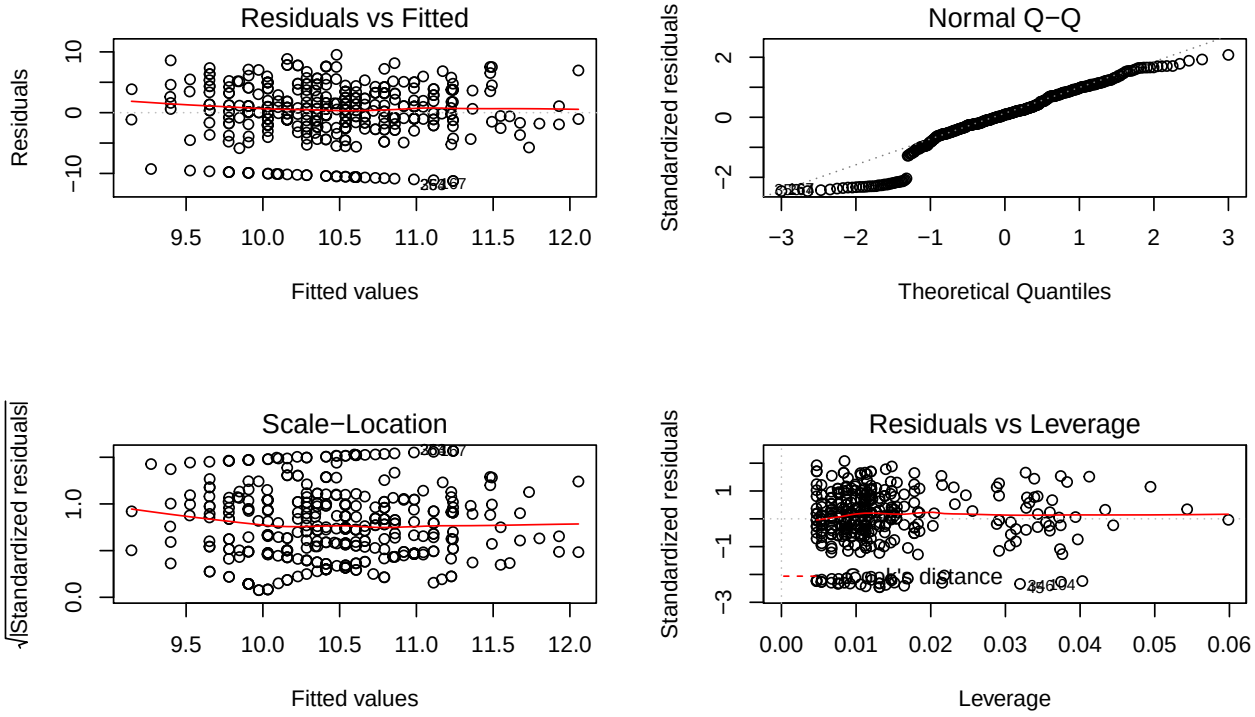
```
##
## Call:
## lm(formula = FinalMath ~ famsize + Pstatus + famrel + famsup,
##     data = d3.q1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2394  -2.0323   0.3929   3.2679   9.5199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.03811    1.33099   8.293 2.15e-15 ***
## famsize     -0.12632    0.08408  -1.502    0.134
## Pstatus0    -0.68979    0.79218  -0.871    0.384
## famrel       0.25400    0.26356   0.964    0.336
## famsup1     -0.44781    0.49434  -0.906    0.366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.603 on 365 degrees of freedom
## Multiple R-squared:  0.01329,    Adjusted R-squared:  0.002477
## F-statistic: 1.229 on 4 and 365 DF,  p-value: 0.298
```

Model validation

- According to the residual and fitted value plot, it does not show any patten, the linearity satisfied.
- In terms of Normal Q-Q plot, it is quite normal → Normality satisfied

In short, the model is qualified, but performance is poor.

```
par(mfrow = c(2, 2))
plot(mod1)
```



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

Fitting model 1

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

$$\begin{aligned} FinalMath = \beta_0 + \beta_1\ Medu1 + \beta_2\ Medu2 + \beta_3\ Medu3 + \beta_4\ Medu4 \\ +\beta_5\ Fedu1 + \beta_6\ Fedu2 + \beta_7\ Fedu3 + \beta_8\ Fedu4 \\ +\beta_9\ Mjob2 + \beta_10\ Mjob3 + \beta_11\ Mjob4 + \beta_12\ Mjob5 \\ +\beta_13\ Fjob2 + \beta_14\ Fjob3 + \beta_15\ Fjob4 + \beta_16\ Fjob5 \\ +\epsilon_j \quad j = 1, ..., n \end{aligned}$$

The coefficients are fitted below

- The result showed that none of the variables are significant to influence the first math score
- The model performance is quite poor since the adjusted R-squared value is very low, indicating that the model can only explain 7.5% of the data.
- However, p-value: 0.0001835 is quite small, means that the whole variables do explain some level of data.

```
mod2.1 = lm(FirstMath ~ Medu + Fedu + Mjob + Fjob, data = d3.q2)
summary(mod2.1)
```

```
##
## Call:
## lm(formula = FirstMath ~ Medu + Fedu + Mjob + Fjob, data = d3.q2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7583 -2.2251 -0.2339  2.1076  8.9645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.12742    3.04253   4.643 4.84e-06 ***
## Medu1       -2.01637    1.93981  -1.039    0.299
## Medu2       -1.54665    1.91583  -0.807    0.420
## Medu3       -1.55312    1.94562  -0.798    0.425
## Medu4       -0.77997    1.98846  -0.392    0.695
## Fedu1       -2.44738    2.32665  -1.052    0.294
## Fedu2       -1.16452    2.32819  -0.500    0.617
## Fedu3       -1.83956    2.33134  -0.789    0.431
## Fedu4       -1.34119    2.35326  -0.570    0.569
## Mjob2        1.20063    0.84879   1.415    0.158
## Mjob3       -0.09744    0.55432  -0.176    0.861
## Mjob4        0.84732    0.61626   1.375    0.170
## Mjob5        0.14733    0.79048   0.186    0.852
## Fjob2       -0.99821    1.18755  -0.841    0.401
## Fjob3       -0.75420    0.86831  -0.869    0.386
## Fjob4       -0.62816    0.89963  -0.698    0.485
```

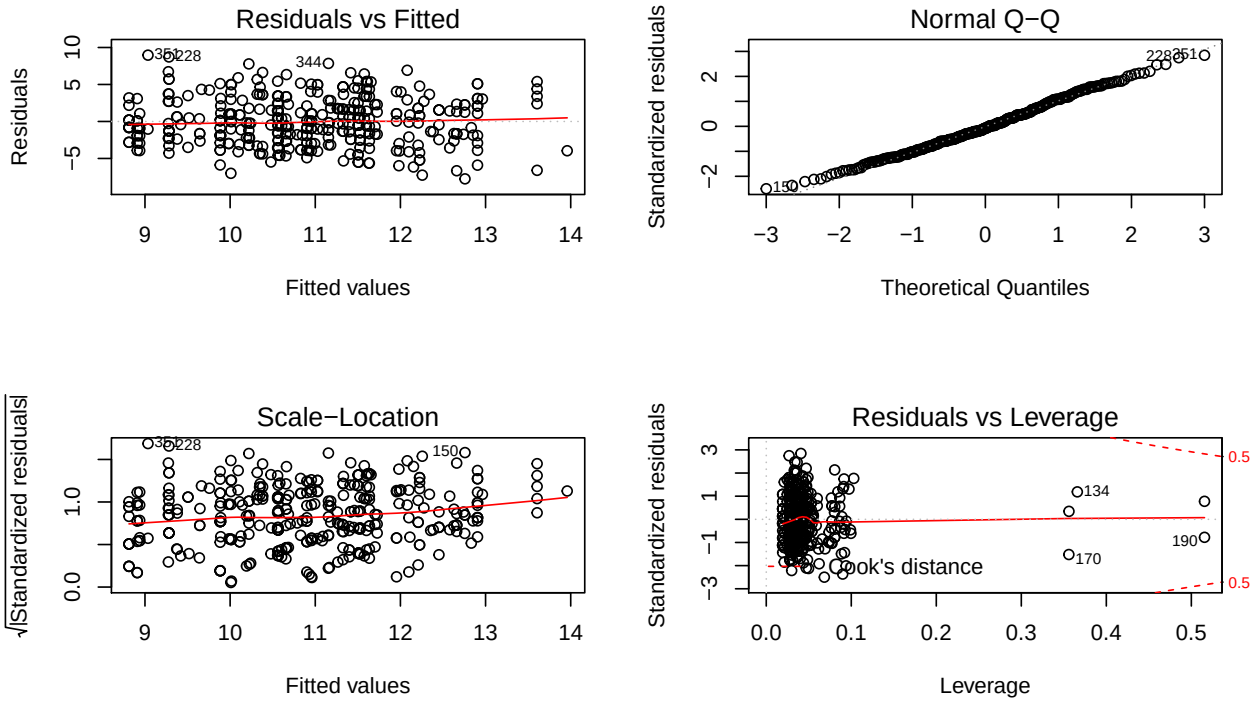
```
## Fjob5        0.75206    1.08091   0.696    0.487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.217 on 353 degrees of freedom
## Multiple R-squared:  0.1156, Adjusted R-squared:  0.0755
## F-statistic: 2.883 on 16 and 353 DF,  p-value: 0.0001835
```

Model validation

- According to the residual and fitted value plot, it does not show any patten but some outliers, the linearity satisfied.
- In terms of Normal Q-Q plot, it is quite normal → Normality satisfied

In short, the model is qualified, but performance is poor.

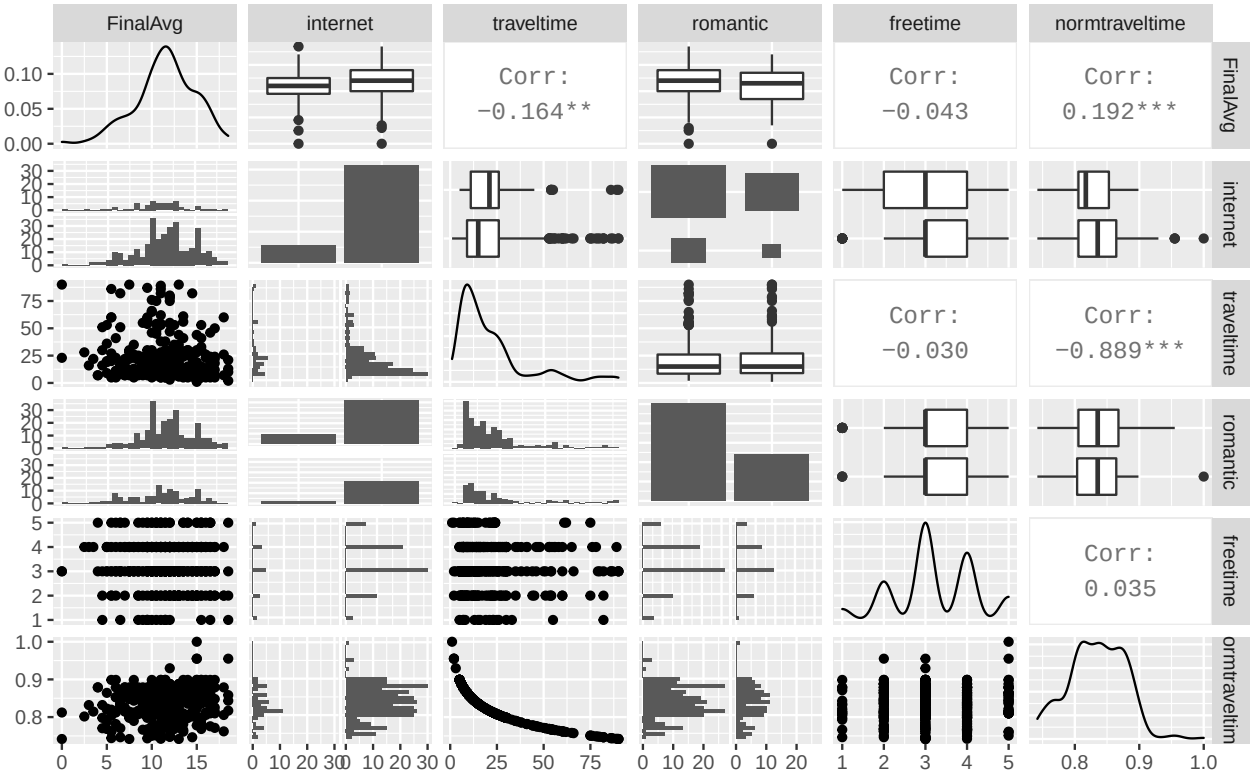
```
par(mfrow = c(2, 2))
plot(mod2.1)
```



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

Scatter Plot

```
GGally::ggpairs(data = d3.q3)
```



Fitting model 1

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

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

After fitting, the model is

$$FinalMath = 4.7694 + 0.9483\ internet - 0.8833\ romantic - 0.1832\ freetime + 8.0041\ normtraveltime \quad j = 1, ..., n$$

- The result showed that Internet, romantic relationship are significant to influence the average score of math and portuegese. Although the normal tranvel time is slightly insignificant, I would include it in the model as well.
- The model performance is still poor since the adjusted R-squared value is very low, indicating that the model can only explain 2.6% of the data.
- However, p-value: 0.007692 is quite small, means that the whole variables do explain some level of data.

```
mod3.1 = lm(FinalAvg ~ internet + romantic + freetime + normtraveltime,
    data = d3.q3)
summary(mod3.1)
```

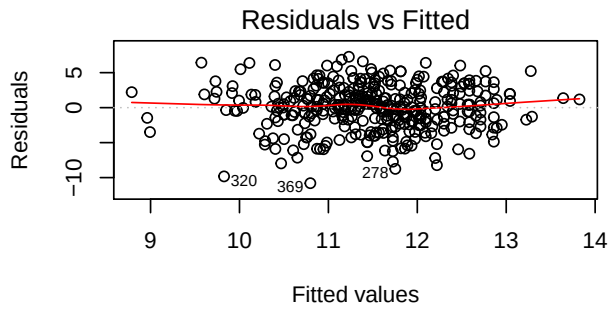
```
##
## Call:
## lm(formula = FinalAvg ~ internet + romantic + freetime + normtraveltime,
##     data = d3.q3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7971  -1.7256   0.1736   2.0938   7.2697
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.08932    3.27099   0.027  0.97823
## internetyes     0.86388    0.46671   1.851  0.06498 .
## romanticyes    -0.86022    0.35842  -2.400  0.01689 *
## freetime       -0.18252    0.17009  -1.073  0.28396
## normtraveltime 13.85861    3.89286   3.560  0.00042 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.205 on 365 degrees of freedom
## Multiple R-squared:  0.06098,    Adjusted R-squared:  0.05068
## F-statistic: 5.925 on 4 and 365 DF,  p-value: 0.0001251
```

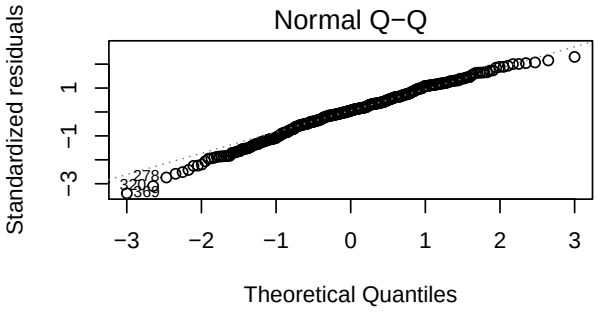
Model validation

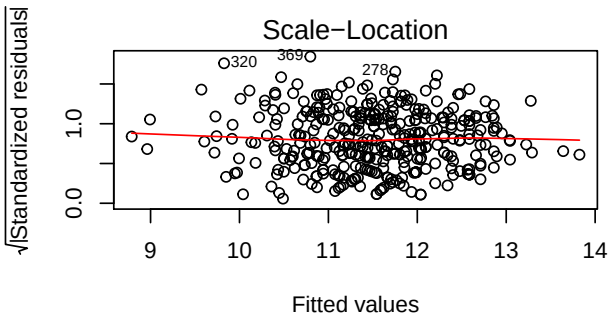
- According to the residual and fitted value plot, it does not show any patten but some outliers, the linearity satisfied.
- In terms of Normal Q-Q plot, it is quite normal → Normality satisfied

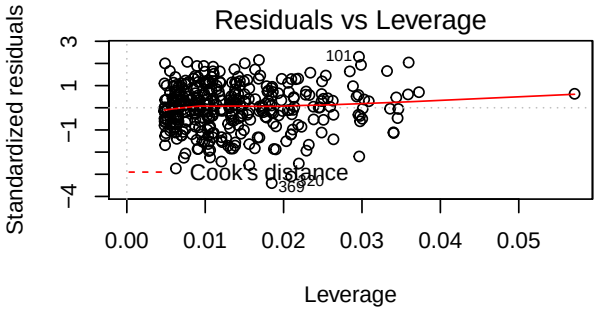
In short, the model is qualified, but performance is poor.

```
par(mfrow = c(2, 2))
plot(mod3.1)
```









Fitting model 2

The reverse multi-regression model for final math is conducted by

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

After fitting, the model is

$$FinalMath = 4.4438 + 0.9062\ internet - 0.8912\ romantic + 7.7374\ normtraveltime \quad j = 1, ..., n$$

- The result does not improve much since the correlation between dependent variable and independent variable is weak

```
mod3.2 = lm(FinalAvg ~ internet + romantic + normtraveltime,
    data = d3.q3)
summary(mod3.2)
```

```
##
## Call:
## lm(formula = FinalAvg ~ internet + romantic + normtraveltime,
##     data = d3.q3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7972  -1.8014   0.1987   2.1170   7.6353
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.3606     3.2447  -0.111 0.911559
## internetyes      0.8213     0.4651   1.766 0.078273 .
## romanticyes     -0.8675     0.3584  -2.420 0.015992 *
## normtraveltime  13.7385     3.8921   3.530 0.000469 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 366 degrees of freedom
## Multiple R-squared:  0.05801,    Adjusted R-squared:  0.05029
## F-statistic: 7.513 on 3 and 366 DF,  p-value: 6.832e-05
```

Model validation

- All the assumptions are satisfied including Linearity, Independency, normality, and Equal variance.

In short, the model is qualified, but performance is poor.

```
par(mfrow = c(2, 2))
plot(mod3.2)
```