PSTAT 126 Final Report

Hongjun Kim, Henry Hartwell, Kyra Maeda, Laurence Liao, Feiyu Fan

Dataset

We chose the Student Performance dataset from the UCI Machine Learning Repository. This dataset measures student achievement in two Portuguese secondary education schools across the subjects of Mathematics and Portuguese. For this project, we decided to only consider data from Mathematics courses because we didn't want to add another variable that could cause us to make false assumptions about the data.

The Mathematics dataset has 395 observations of students across 2 different Portuguese schools and has 30 variables that could be related to student performance including student demographics, sex, age, number of study hours, and level of education of the mother and father.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11

Student performance is ultimately measured by G3, which represents the final student grade in period 3 and is numeric from 1 to 20 with 1 being the lowest grade and 20 being the highest grade. G1 and G2, the final student grades in period 1 and period 2 respectively, were also captured to measure student performance over time.

It is important to note that 38 students had a G2 and G3 value of 0, indicating the grade could not be calculated. We removed these students from our dataset, giving us 357 observations in total to analyze.

We also analyzed the studytime variable, and wanted to see if study time could be a predictor for G3 (the final grade). The levels of study time are as follows and are measured in weekly number of study hours: 1 - <2 hours; 2 - 2 to 5 hours; 3 - 5 to 10 hours; 4 - >10 hours)

Regression Model

Question 1:

Does Study time affect G3 (Final Grade)?
For Hypothesis 1 we conducted a linear model between G3 as dependent variable and study time as independent variable. In Question 1, we are using a single predictor from the data set study time which results in no additional steps to be made for Model selection and Interaction Terms. AIC and BIC simply compared returned single values due to single predictors.

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.0659 -2.0659 -0.0659  1.9341  7.9341

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.4674     0.3318  34.558   <2e-16 ***
studytime2-5 hours  -0.4015     0.4072  -0.986   0.3248
studytime5-10 hours  1.0919     0.5309   2.057   0.0404 *
studytime>10 hours   1.1993     0.7295   1.644   0.1011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.183 on 353 degrees of freedom
Multiple R-squared:  0.03587,   Adjusted R-squared:  0.02768
F-statistic: 4.378 on 3 and 353 DF,  p-value: 0.004825
```

```{r}
AIC(model_simple_Q1)
BIC(model_simple_Q1)
```

```
[1] 1849.006
[1] 1860.639
```

Question 2:

Does the G1 and G2 (predictor) affect G3(response)?
For Hypothesis 2 we conducted a multiple linear model between G3 as dependent variable and G1 and G2 as the independent variables.

Summary of Linear and Multiple Models

```
Call:
lm(formula = G3 ~ G1, data = student)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6200 -1.0616 -0.1733  1.0501  3.7151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.51338    0.28042   5.397 1.24e-07 ***
G1           0.88832    0.02392  37.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.462 on 355 degrees of freedom
Multiple R-squared:  0.7953,    Adjusted R-squared:  0.7947
F-statistic:  1379 on 1 and 355 DF,  p-value: < 2.2e-16
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.2172 -0.1978 -0.1494  0.8119  2.8022

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.27529    0.16686    1.65   0.0999 .
G2           0.99031    0.01416   69.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8407 on 355 degrees of freedom
Multiple R-squared:  0.9323,    Adjusted R-squared:  0.9322
F-statistic:  4893 on 1 and 355 DF,  p-value: < 2.2e-16
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.1845 -0.2927 -0.1673  0.7056  2.8190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19482    0.16572   1.176 0.240541
G1           0.11167    0.03133   3.564 0.000415 ***
G2           0.88661    0.03226  27.486  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8272 on 354 degrees of freedom
Multiple R-squared:  0.9347,    Adjusted R-squared:  0.9343
F-statistic:  2533 on 2 and 354 DF,  p-value: < 2.2e-16
```

AIC and BIC for Multiple Model

After running a forward stepwise selection, the model that gave the lowest AIC value was when G1 and G2 were both included, suggesting that a multiple linear regression model was better than separate linear regression models.

```
Start:  AIC=837.66
G3 ~ 1

       Df Sum of Sq    RSS     AIC
+ G2    1    3458.1  250.9 -121.88
+ G1    1    2949.9  759.2  273.36
<none>              3709.0  837.66

Step:  AIC=-121.88
G3 ~ G2

       Df Sum of Sq    RSS     AIC
+ G1    1    8.6934 242.22 -132.47
<none>              250.92 -121.88

Step:  AIC=-132.47
G3 ~ G2 + G1

[1] 882.6531
[1] 898.164
```

Since the interaction between G1 and G2 was not statistically significant, we decided to remove it and keep the original multiple linear model.

```
Call:
lm(formula = G3 ~ G1 * G2, data = student)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3091 -0.3429 -0.1314  0.6538  2.7969

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.892018   0.511514   1.744   0.0821 .
G1          0.047245   0.054576   0.866   0.3873
G2          0.822473   0.054955  14.966   <2e-16 ***
G1:G2       0.005521   0.003833   1.440   0.1506
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8259 on 353 degrees of freedom
Multiple R-squared:  0.9351,    Adjusted R-squared:  0.9345
F-statistic:  1695 on 3 and 353 DF,  p-value: < 2.2e-16
```

Interpretation

Under this section, we further discuss the role of coefficients of regression model in the context of research questions, and the impact of each predictor on the response variable.

For research Question 1, recall the linear regression model:

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.0659 -2.0659 -0.0659  1.9341  7.9341

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          11.4674     0.3318  34.558   <2e-16 ***
studytime2-5 hours   -0.4015     0.4072  -0.986   0.3248
studytime5-10 hours   1.0919     0.5309   2.057   0.0404 *
studytime>10 hours    1.1993     0.7295   1.644   0.1011
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.183 on 353 degrees of freedom
Multiple R-squared:  0.03587,   Adjusted R-squared:  0.02768
F-statistic: 4.378 on 3 and 353 DF,  p-value: 0.004825
```

Given the linear model of final grade `G3` and study time, we utilize 4 levels of study time to predict `G3`: weekly study time is less than 2 hours, 2 to 5 hours, 5 to 10 hours, and greater than 10 hours. Moreover, We choose `study time < 2` as reference level and others as dummy variables.

The coefficient of intercept equals 11.467, meaning when `studytime` is less than 2 hours, the average final grade is expected to be 11.4674. This figure suggests a baseline for the minimum point of final grade ‘G3’ we can receive given the information on the ‘studytime.’ And the p-value < 0.05 suggests that the ‘studytime<2’ is statistical significance to the model.

For ‘studytime’ between 2 to 5 hours, the average final grade ‘G3’ is predicted to decrease by 0.4015 points, holding all other factors remain constant. This fact suggests a negative relationship between final grade and studying for 2 to 5 hours.

For `studytime`between 5 to 10 hours, the average final grade `G3` is predicted to increase by 1.0919 points, assuming all other factors remain constant.The study time category 5-10 hours is the only statistically significant predictor of the final grade (p=0.0404).This suggests that studying for 5-10 hours per week is associated with a significantly higher final grade compared to studying <2 hours.

For `studytime` being greater than 10 hours, the average final grade `G3` is predicted to increase by 1.1993 points, assuming all other factors remain constant. The p-value is 0.1011, which is above than the conventional threshold of 0.05. This suggests the effect of study time on G3 is not statistically significant.

From the summary, we also observe that the R Squared is 0.0357, which suggests that Only 3% of the variation in the final grades (`G3`) is explained by the `studytime` variable. This indicates a weak relationship between study time and the final grade. Moreover, the Adjusted R Squared which accounts for the number of predictors and the sample size, is slightly lower than R Squared . This indicates that adding additional predictors might improve the model's explanatory power.

```{r}
rse_Q1 <- summary(model_simple_Q1)$sigma
df_resid_Q1 <- df.residual(model_simple_Q1)
# calculate RSS
rss_Q1 <- (rse_Q1^2) * df_resid_Q1
cat("Residual Sum of Squares (RSS):", rss_Q1, "\n")
```

```
Residual Sum of Squares (RSS): 3575.987
```

Since `G3` values range from 0 to 20, an `RSS` of 3575.987 might represent significant error. Hence, a single predictor `studytime` does not contribute to the significance of final grade `G3`.

For research Question 2, recall the linear regression model:

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.1845 -0.2927 -0.1673  0.7056  2.8190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19482    0.16572   1.176 0.240541
G1           0.11167    0.03133   3.564 0.000415 ***
G2           0.88661    0.03226  27.486  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8272 on 354 degrees of freedom
Multiple R-squared:  0.9347,    Adjusted R-squared:  0.9343
F-statistic:  2533 on 2 and 354 DF,  p-value: < 2.2e-16
```

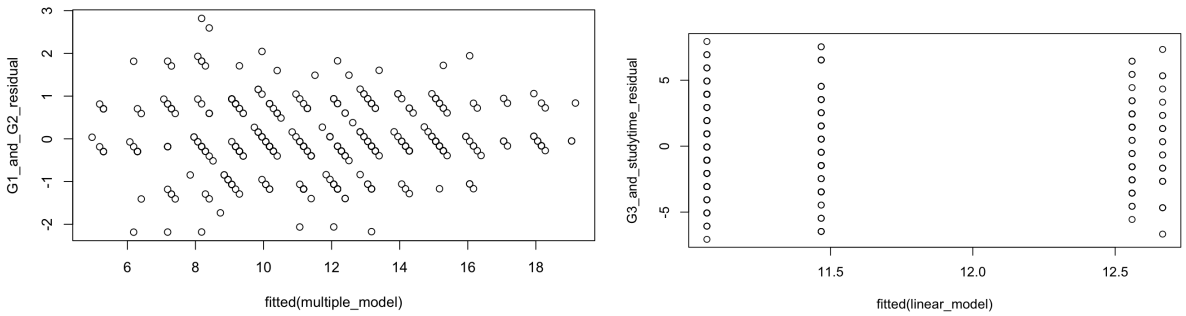
The intercept 0.19482 represents the expected value of `G3` when both `G1` and `G2` are zero. In this context, the intercept is not meaningful because grades of zero for `G1` and `G2` are unlikely in a real-world scenario.

For coefficient of G1 0.11167, we have for every one-unit increase in the first-period grade `G1`, the final grade `G3` is expected to increase by approximately 0.11167points, holding `G2` constant. This suggests that `G1` has a small positive impact on `G3.`

For coefficient of G2 0.88661, we have for every one-unit increase in the second-period grade `G2`, the final grade `G3` is expected to increase by approximately 0.88661points, holding `G1` constant. This indicates that `G2` has a much stronger impact on `G3` compared to `G1.`

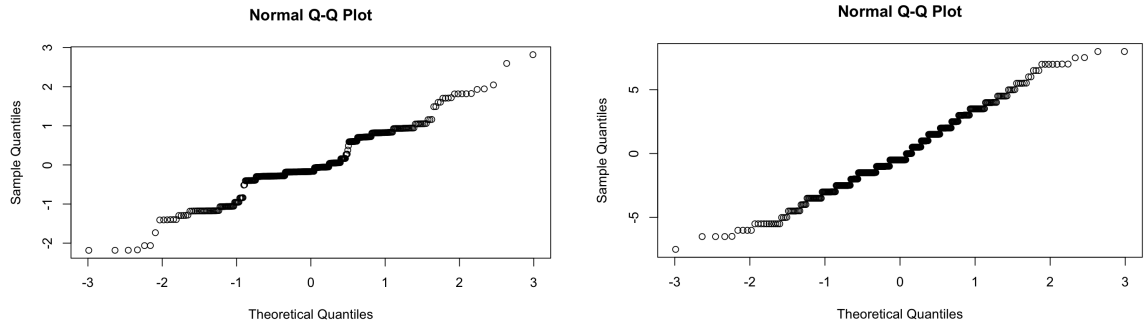
Assumptions: Diagnostic Checks

Linearity, homoscedasticity, and independence



The graph on the left plots the residuals of G1 and G2 when compared to G3 against the fitted values of the model. Similarly, the graph on the right plots the residuals of the study time variable when compared to G3 against the fitted values. The figures show that the plotted residuals are randomly scattered around 0 with no clear pattern. Both graphs indicate linearity, independence, and homoscedasticity as a result.

Normality of Errors

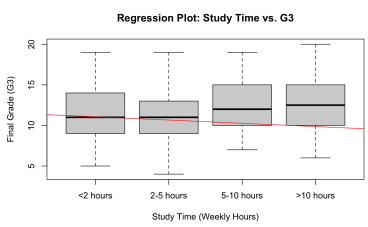


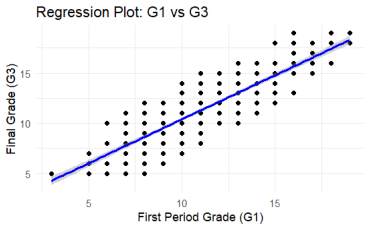
To check for independence, we used QQ plots to plot the residuals. The graph on the left represents the residuals for the G3 against G1 and G2 data and the graph on the right represents the residuals for the G3 against study time data. Both plots fit a relatively diagonal and linear pattern. While both plots show flattening data points at the tail ends, which could indicate the residuals are not perfectly normal.

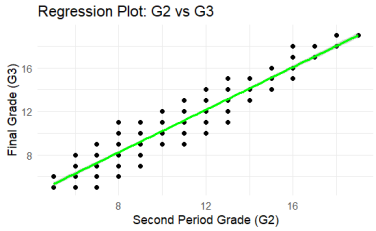
To make the residuals more normal, we recommend using a Box Cox transformation on the study time residuals but not the G1 and G2 residuals. We attempted a Box Cox transformation for the plot of G3 against G1 and G2, but since lambda equaled 1.03 (close to 1), the QQ plot did not improve since the residuals are already fairly normally distributed.

Conclusion

In conclusion, we see G1 and G2 were both significant predictors of G3 meaning the inclusion of the multiple regression model improved the fit. Here are the three regression plots.







1st regression plot shows horizontal pattern showing no strong relation between study time vs. G3. Study Time may not be a strong predictor of G3 in the dataset.
The 2nd and 3rd regression plot shows a positive relationship between G1 vs G3 and G2 vs G3. This means there is a positive linear relationship between these variables. Likely meaning that G1 and G2 are significant predictors for G3.

Through diagnostic checking, we found that the regression for G3 against G1 and G2 as well as the regression for G3 against studytime satisfies the 4 key regression assumptions. We checked for linearity, homoscedasticity, normality, and independence, which allows us to draw conclusions from our data. It is important to note that we recommend performing a Box Cox transformation on the G3 vs studytime data to improve the normality of the residuals.