PSTAT 126
Final Project

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

Dataset

- Student Performance dataset from UCI Machine Learning Repository
- Analyzes student performance data in 2 subjects (Math and Portuguese) for 2 Portuguese schools
    - We only analyzed data collected from students in the Math class
- G1 corresponds to grade in the first period, G2 is grade in the second period, and G3 is grade in the third period
- 357 observations and 30 variables (we removed 38 observations that had G2 and G3 values of 0)

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

Research Questions

1. Does Study time level affect G3 (Final Grade)?
    a. Levels for study time (weekly):
        i. 1 - < 2 hours
        ii. 2 - 2 to 5 hours
        iii. 3 - 5 to 10 hours
        iv. 4 - >10 hours
    b. Range for G3 (Final Grade): 5 ~ 19
2. Do G1 (1st Period Grade) and G2 (2nd Period Grade) affect G3 ?
    a. All numeric values from 1 to 20
    b. Range for G1 (1st Period Grade): 3-19
    c. Range for G2 (2nd Period Grade): 5-19

Methodology

Question 1: Does Study time level affect G3 (the final grade)?

- We ran a factor linear regression model with multiple levels since study time is a categorical variable.

Question 2: Do G1 and G2 affect G3?

- We ran two simple linear regression models that saw how G1 and G2 predicted G3 individually and a multiple linear regression model to study the effects that G1 and G2 have in predicting G3
- After running a forward stepwise selection, we chose the multiple linear regression model because it had the lowest AIC
- We considered adding an interaction term to see combined effect G1 and G2 could have on G3, but our regression showed that it was not statistically significant

Coefficients/Statistical Significance

Simple Linear Model (G3 ~ studytime)

Study hours:

- (< 2) : The expected G3 = 11.4674. Statistically significant
- (2-5) : 0.4015 decrease in G3. Not statistically significant
- (5-10) : 1.0919 increase in G3. Statistically significant
- (> 10) : 1.1993 increase in G3. Not statistically significant
- R^2 = 0.03587: 3.587% of variation is explained by 'studytime'
- RSS = 3575.99. Significant error

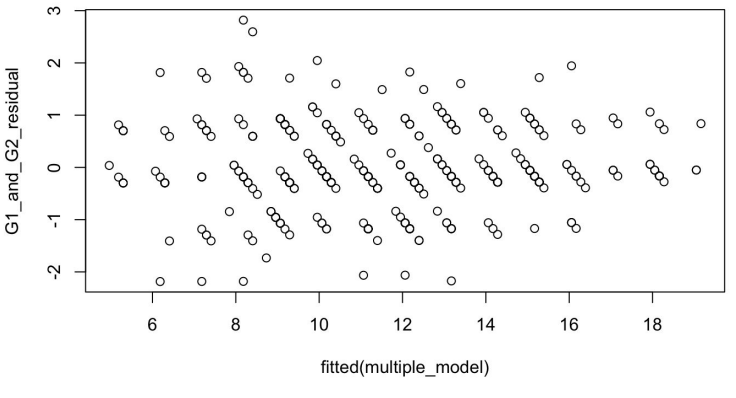
Multiple Linear Model (G3 ~ G1 + G2)

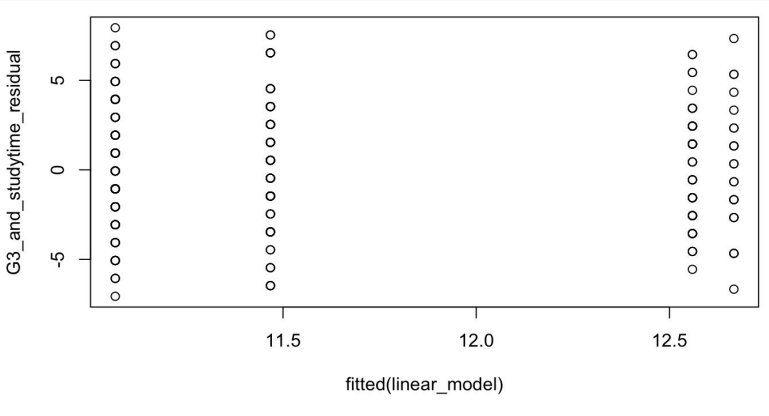
- Intercept = 0.19482: Expected value when G1 + G2 = 0
- G1: 0.11167 increase in G3. Statistically significant
- G2: 0.8861 increase in G3. Statistically significant
- R^2 = 0.9347: 93.5% of variation is explained by G1 and G2
- RSS = 242.22. Strong fit

```
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
```

```
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
```

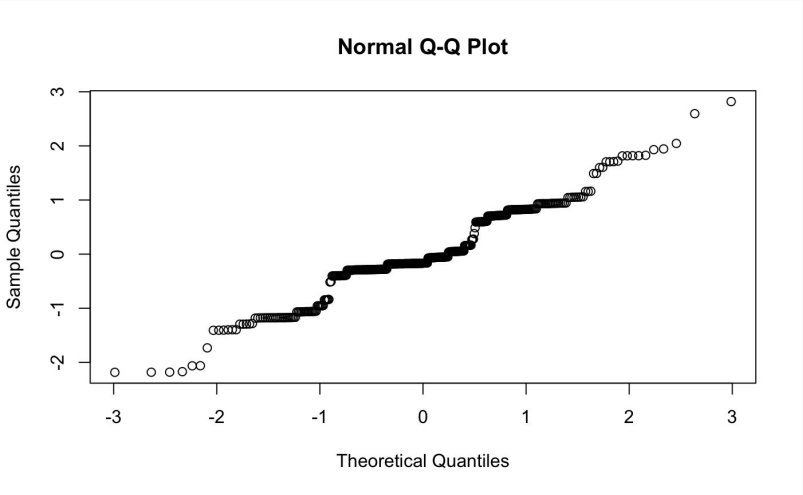
Diagnostic Checking



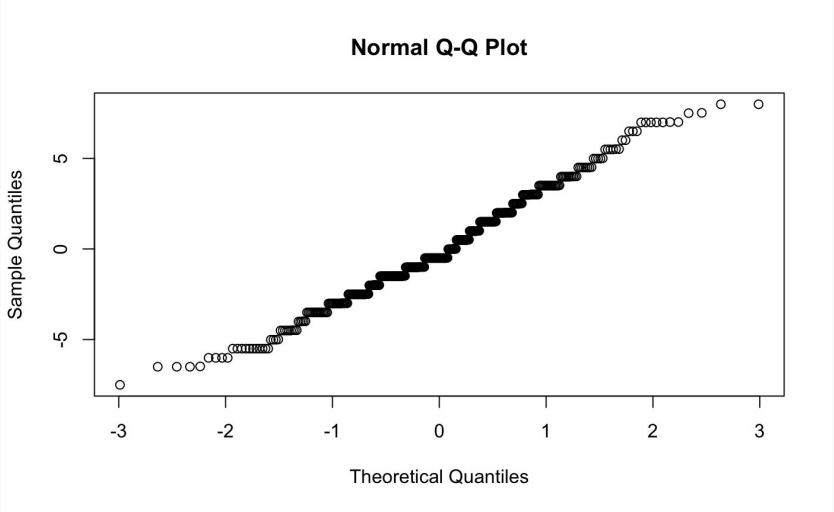


The residual plots show that the residuals for both graphs are scattered around 0 and there is no clear pattern. This indicates that the relationship between our predictors and response variables is linear and that the residuals are independent and exhibit homoscedasticity.

Diagnostic Checking



QQ plot for G3 against G1 and G2



QQ plot for G3 against study time

Both QQ plots are fairly linear and diagonal. We tested Box Cox transformations on both plots to try to improve the normality of the residuals, but only the plot on the right improved.

Conclusion

Overall, we found that the level of weekly study time hours does not significantly predict G3 for the number of weekly study hours <2, 2-5, and >10. However, study time hours for those who study between 5 and 10 hours per week can predict G3. The small R^2 suggests weekly study time hours is not enough to predict final grades.

We also found that G1 and G2 are statistically significant in predicting G3, with G2 being a stronger predictor.

Recommendations:

- Break down the data even further by separating the data by the school that it was collected from and repeat the same processes.
- Add more predictor variables along with the level of weekly study time hours to improve the model fit