Project 3 Word File

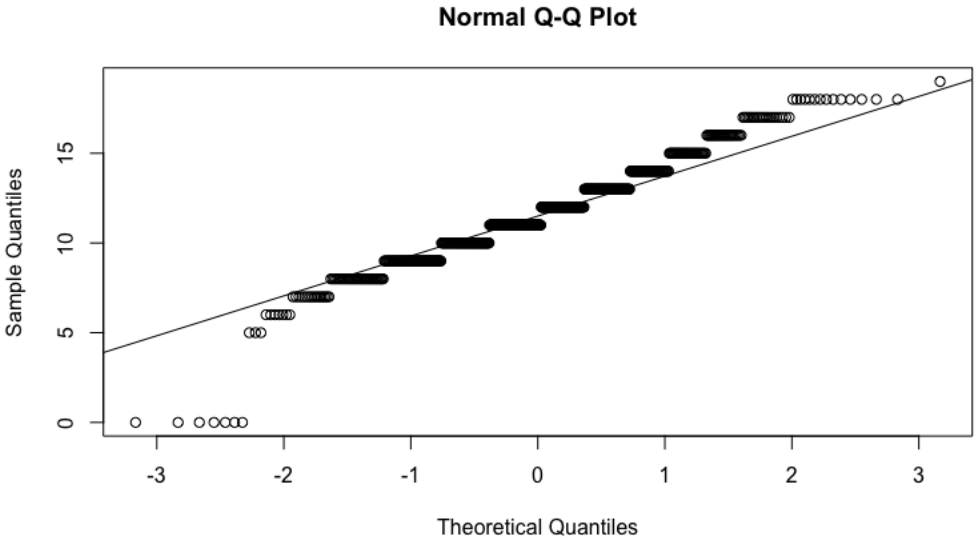
Section 1: Test of means

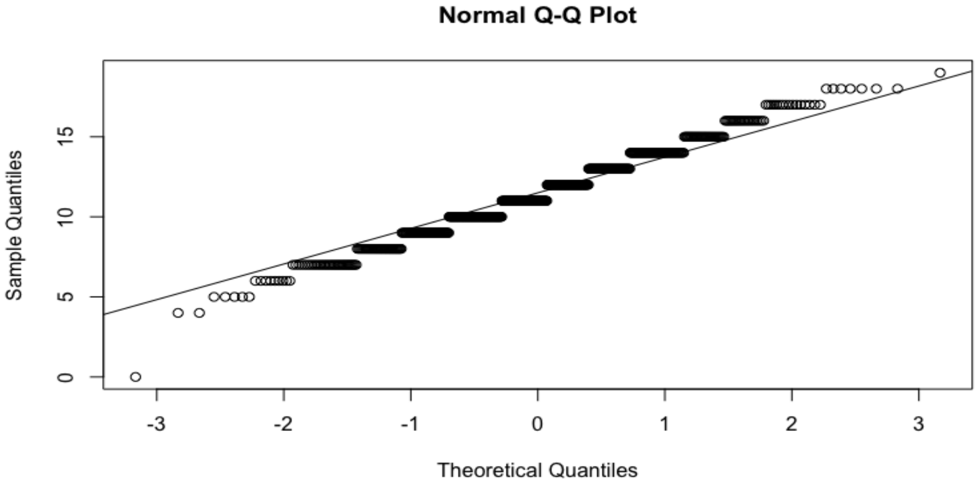
Question 1: Perform the appropriate t-test to determine if the means for G1 grade and G2 grades are the same for the Portuguese language dataset.

Test: Paired T-test

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the G1 and G2 columns of the Portuguese language dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: The means for G1 and G2 grades are the same for the Portuguese language dataset.

Research hypothesis: The G1 and G2 grades means in the Portuguese language dataset are significantly different.

Hypothesis mathematical notation:

p1 = population mean of G1 grades in the Portuguese language dataset
p2 = population mean of G2 grades in the Portuguese language dataset
α(alpha) = 0.05

Null hypothesis – H_0: p1 = p2, of p-value >/= 0.05
Research/Alternative hypothesis – H_R: p1 ≠ p2, if p-value < 0.05

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing the Paired T-test we got a p-value of 0.003341 which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

Conclusion as show in class in words: We come to the conclusion that we do have statistical evidence to show that the means of G1 and G2 grades are significantly different in the Portuguese language dataset and therefore we accept the research/alternative hypothesis.

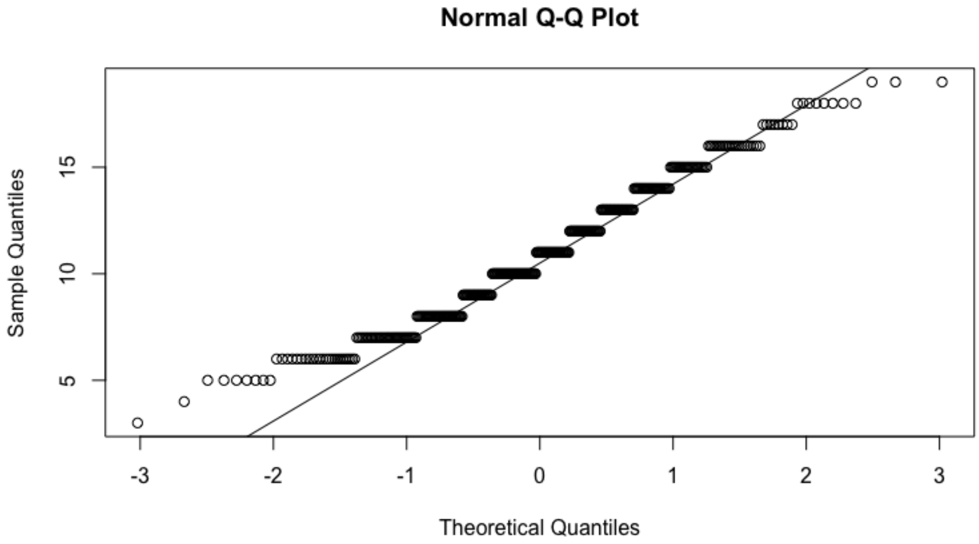
Conclusion as show in class mathematically: Accept H_R (p1 ≠ p2) since p-value (0.003341) < 0.05.

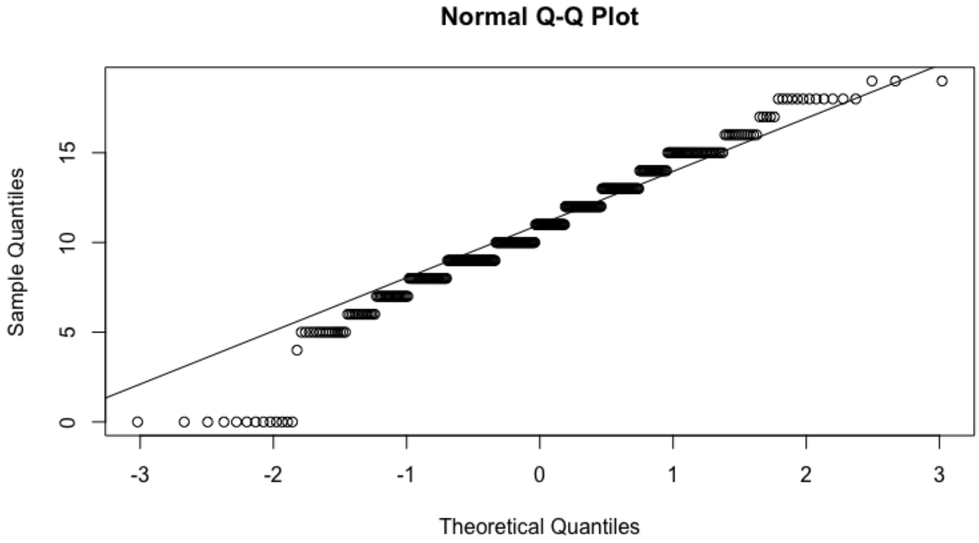
Question 2: Perform the appropriate t-test to determine if the means for G1 grade and G2 grades are the same for the Mathematics dataset.

Test: Paired T-test

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the G1 and G2 columns of the Mathematics dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: The means for G1 and G2 grades are the same for the Mathematics dataset.

Research hypothesis: The G1 and G2 grades means in the Mathematics dataset are significantly different.

Hypothesis mathematical notation:

p1 = population mean of G1 grades in the Mathematics dataset
p2 = population mean of G2 grades in the Mathematics dataset
α(alpha) = 0.05

Null hypothesis – H_0: p1 = p2, of p-value >/= 0.05
Research/Alternative hypothesis – H_R: p1 ≠ p2, if p-value < 0.05

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing the Paired T-test we got a p-value of 0.05014 which is greater than our alpha value of 0.05. This means that we fail to reject the null hypothesis.

Conclusion as show in class in words: We come to the conclusion that we do not have statistical evidence to show that the means of G1 and G2 grades are significantly different in the Mathematics dataset and therefore we accept the research/alternative hypothesis.

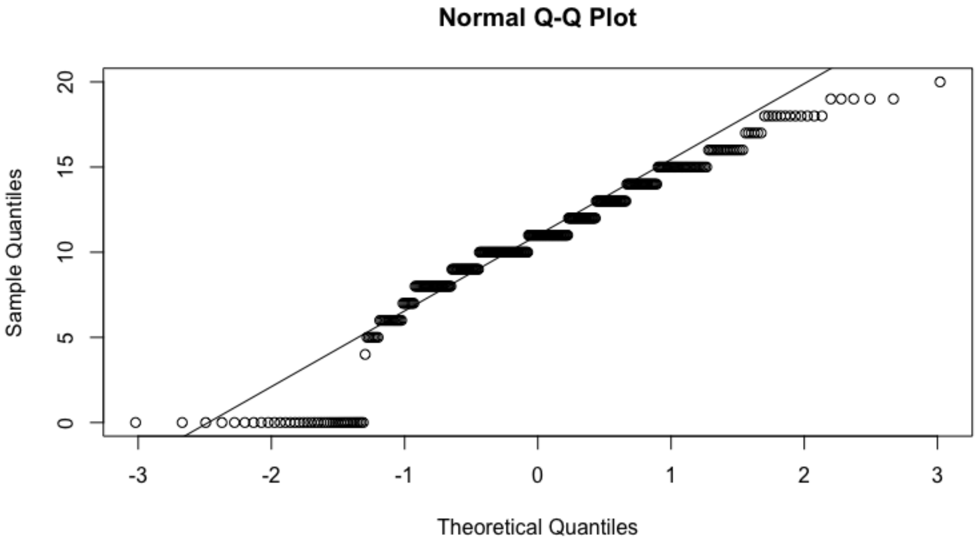
Conclusion as show in class mathematically: Do not accept H_R, p1 = p2, since p-value (0.05014) > 0.05.

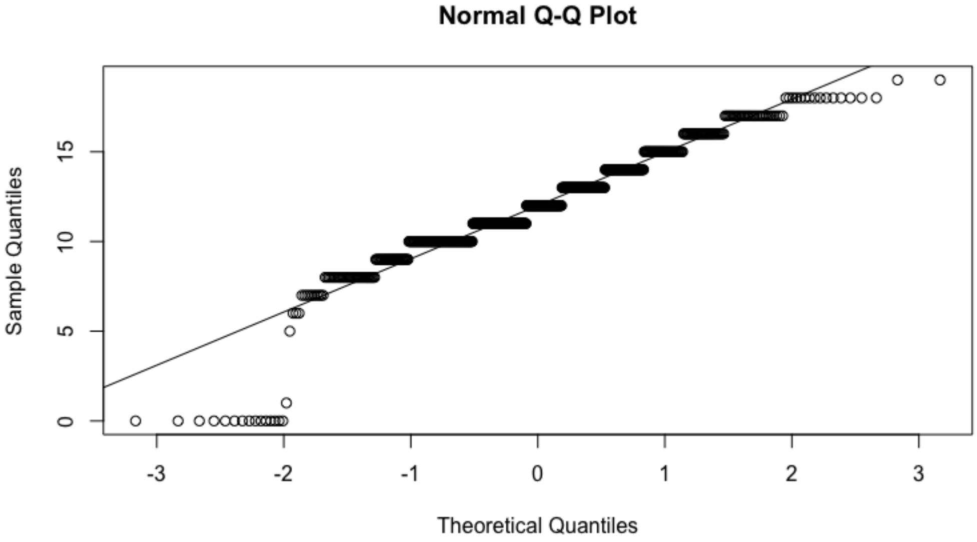
Question 3: Perform the appropriate t-test to determine if mean G3 grades for Mathematics are the same as the G3 grades for Portuguese.

Test: Two Sample T-test (Welch T-test)

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the G3 columns of the Mathematics and Portuguese language datasets. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: The means for G3 grades in the Mathematics and Portuguese language datasets are the same.

Research hypothesis: The G3 grades means in the Mathematics and Portuguese language datasets are significantly different.

Hypothesis mathematical notation:
p1 = population mean of G3 grades in the Mathematics dataset
p2 = population mean of G3 grades in the Portuguese language dataset
α(alpha) = 0.05

Null hypothesis – H_0: p1 = p2, of p-value >/= 0.05
Research/Alternative hypothesis – H_R: p1 ≠ p2, if p-value < 0.05

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing the Welch's Two Sample T-test we got a p-value of 2.215e-08 (0.00000002215)which is less than our alpha value of 0.05. This means that we reject the null hypothesis.

Conclusion as show in class in words: We come to the conclusion that we have statistical evidence to show that the G3 grades means in the Mathematics and Portuguese language datasets are significantly different and therefore we accept the research/alternative hypothesis.

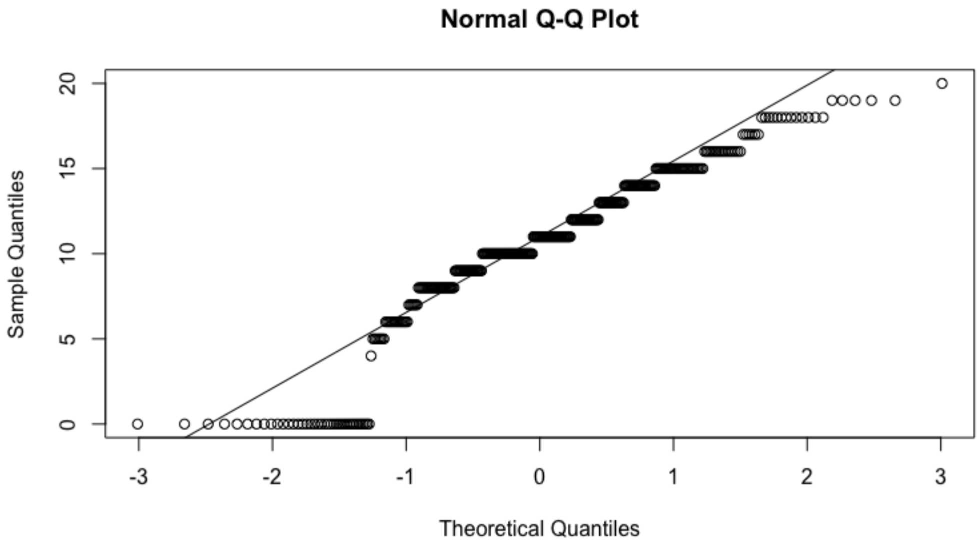
Conclusion as show in class mathematically: Accept H_R (p1 ≠ p2), since p-value (2.215e-08 (0.00000002215)) < 0.05.

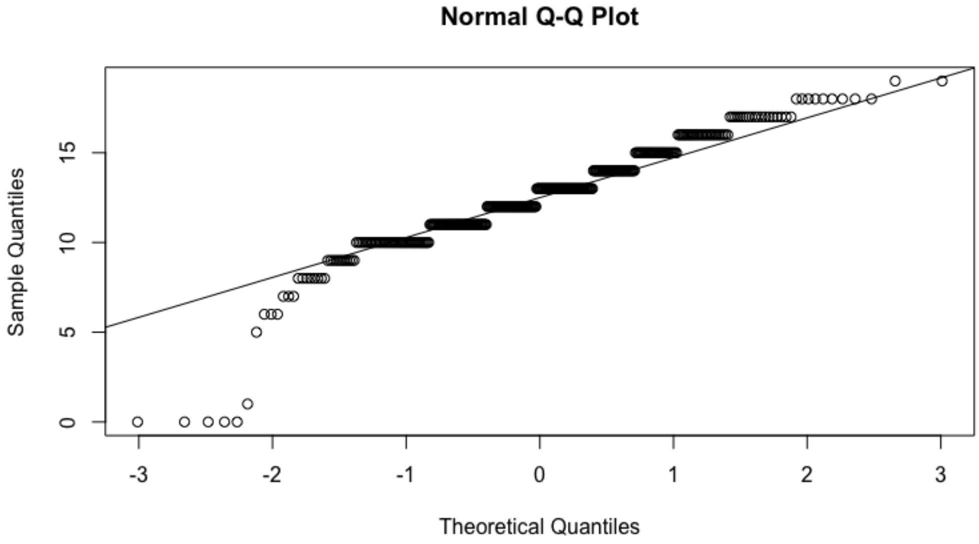
Question 4: Perform the appropriate t-test to determine if the means of G3 grades are the same for both Portuguese and Mathematics for students appearing in both datasets.

Test: Paired T-test

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the G3 columns of the merged Mathematics and Portuguese language dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: The means of G3 grades are the same for both Portuguese and Mathematics for students appearing in both datasets

Research hypothesis: The means of G3 grades for both Portuguese and Mathematics for students appearing in both datasets are significantly different.

Hypothesis mathematical notation:

p1 = population mean of G3 grades for Portuguese merged dataset
p2 = population mean of G3 grades for Mathematics merged dataset
α(alpha) = 0.05

Null hypothesis – H_0: p1 = p2, of p-value >/= 0.05
Research/Alternative hypothesis – H_R: p1 ≠ p2, if p-value < 0.05

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing the Paired T-test we got a p-value that is less than 2.2e-16 or p-value < 2.2e-16, which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

Conclusion as show in class in words: We come to the conclusion that we do have statistical evidence to show that the means of G3 grades are significantly different for both Portuguese and Mathematics for students appearing in both datasets and therefore we accept the research/alternative hypothesis.

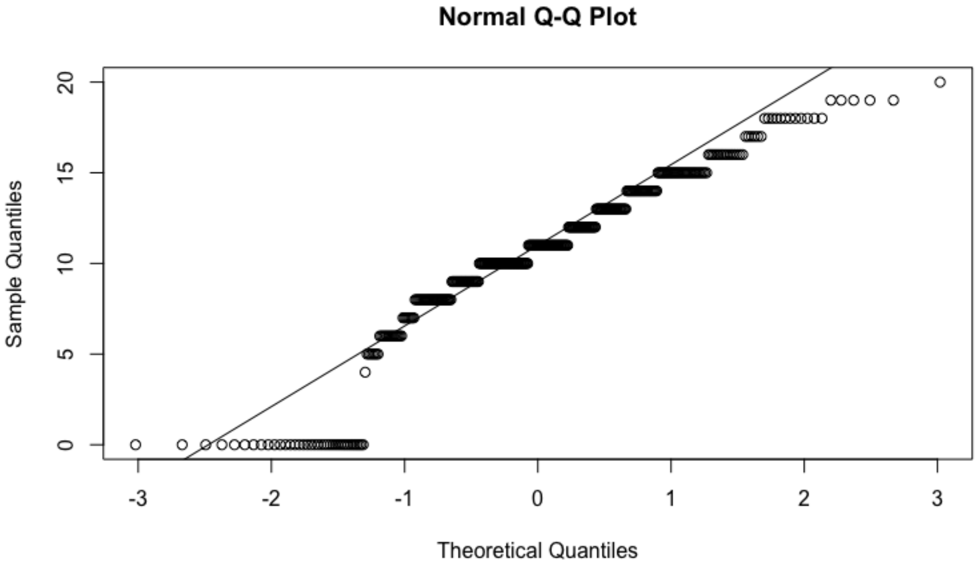
Conclusion as show in class mathematically: Accept H_R (p1 ≠ p2), since p-value
(p-value < 2.2e-16) < 0.05.

Section 2: ANOVA

Question 1: Perform a 2-way ANOVA examining the effect of weekday and weekend alcohol consumption on G3 grades for the Mathematics dataset.

Assumptions:

Normality



Comment: I used a QQ plot to determine the normality of the data for the G3 column of the Mathematics dataset. After examining the QQ plot we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Independence: We assume independence because the variables seem to be independent of each other and therefore this assumption passes, and we do not need to do any transformation on the data.

Hypotheses in Words:

Omnibus Null Hypothesis: There is no difference between weekday and weekend alcohol consumption on G3 grades in the Mathematics dataset, our alpha is 0.05.

Research Hypothesis: At least weekday or weekend alcohol consumption has a significant effect on G3 grades in the Mathematics dataset.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

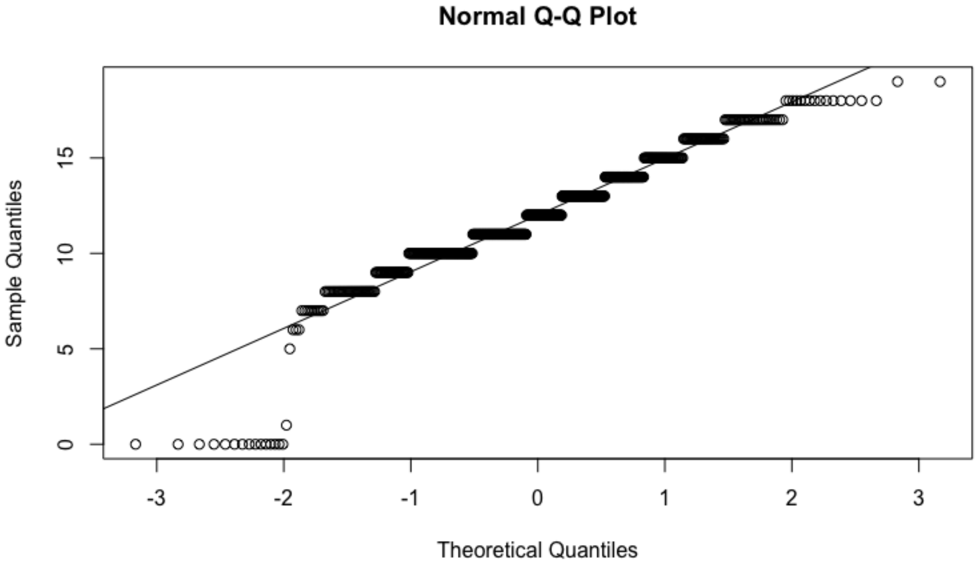
Whether the null was rejected and why: Based of the summary of the Best-Fit model which was the two-way ANOVA additive model the Pr > f values are greater than our alpha(α) of 0.05 which means that we do not reject the null hypothesis.

Conclusion as show in class in words: We see that there is no statistical significance in the G3 grades average in terms of weekday (f(4) = 1.575, p = 0.180) and weekend(f(4) = 0.480, p = 0.751) alcohol consumption. Since we do not reject the null hypothesis, we cannot say that there is statistically significance on the G3 grades in the mathematics dataset when it came to either weekday or weekend alcohol consumption.

Question 2: Perform a 2-way ANOVA examining the effect of weekday and weekend alcohol consumption on G3 grades for the Portuguese dataset.

Assumptions:

Normality



Comment: I used a QQ plot to determine the normality of the data for the G3 column of the Portuguese dataset. After examining the QQ plot we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Independence: We assume independence because the variables seem to be independent of each other and therefore this assumption passes, and we do not need to do any transformation on the data.

Hypotheses in Words:

Omnibus Null Hypothesis: There is no difference between weekday and weekend alcohol consumption on G3 grades in the Portuguese dataset, our alpha is 0.05.

Research Hypothesis: At least weekday or weekend alcohol consumption has a significant effect on G3 grades in the Portuguese dataset.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: Based of the summary of the Best-Fit model which was the two-way ANOVA additive model the weekday (Dalc) Pr > f value is less than our alpha(α) of 0.05 which means that we do reject the null hypothesis.

Conclusion as show in class in words: We see that there is statistical significance in the G3 grades average in terms of weekday (f(4) = 8.193, p = 1.9e-06 (0.0000019)) alcohol consumption. A Tukey post-hoc test revealed significant pairwise differences between the 2-1(Weekday level) pair, with an average difference of -0.9356984 (p = 0.0322973), between the 4-1(Weekday level) pair, with an average difference of -3.3581583 (p = 0.0001916), and between the 4-2 (Weekday level) pair, with an average difference of -2.4224599 (p = 0.0264952). Based on the results we got we can say that weekday alcohol consumption levels of 1 and 2 were statistically significant when it came to G3 grades in the Portuguese dataset compared to the other levels of weekday and weekend alcohol consumption.

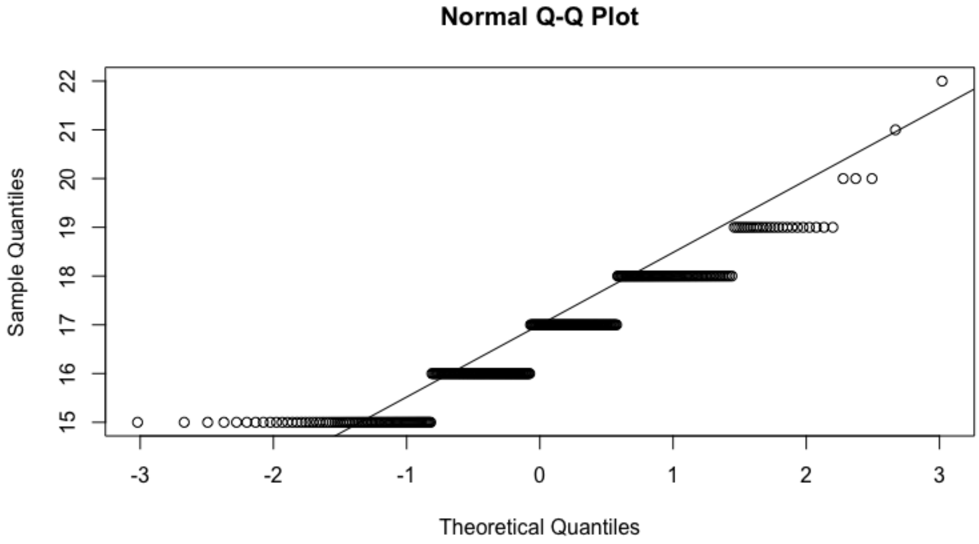
Question 3: Are there any interesting patterns in the data?
An interesting pattern I have noticed in the data is that after performing the two-way ANOVA on both datasets pertaining to the effect of weekday and weekend alcohol consumption on G3 grades, weekday alcohol consumption had lower p-values in both the additive models for both datasets compared to that of the p-values of weekend alcohol consumption. This to me highly suggests that weekday alcohol consumption could have a higher significance when it comes to impacting G3 grades and there was statistical evidence that weekday alcohol consumption had an impact on G3 grades in the Portuguese data set.

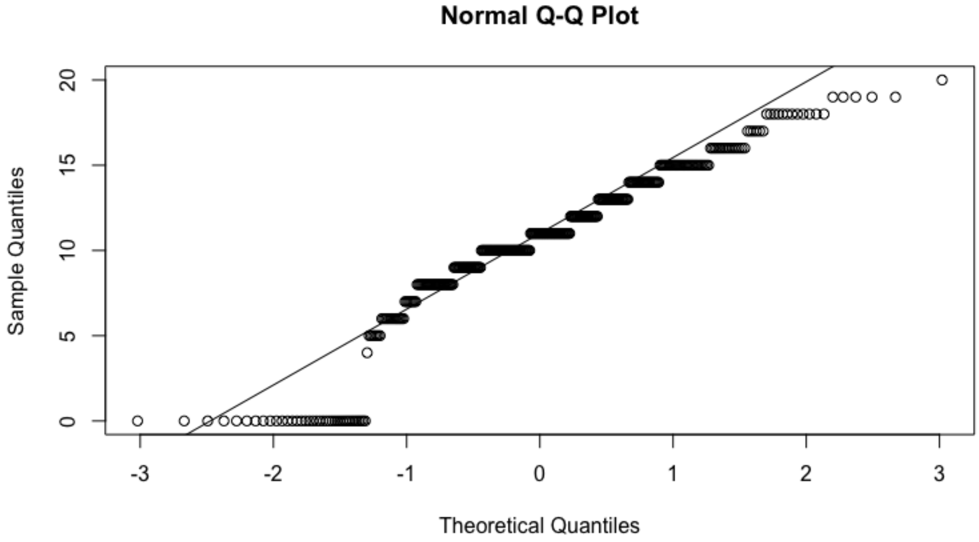
Section 3 Correlational Analysis

Question 1: Is age correlated (using a Pearson R) to G3 grade for Mathematics?

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the age and G3 columns of the Mathematics dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: Age is not correlated to the G3 grade for Mathematics.

Research hypothesis: Age is correlated to the G3 grade for Mathematics.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

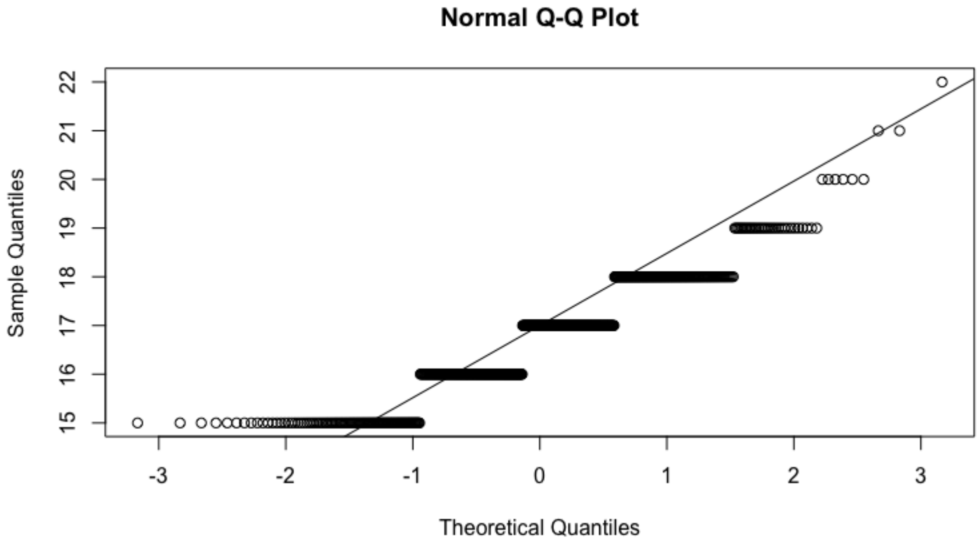
Whether the null was rejected and why: After performing the Pearson R test we got a p-value of 0.001271 which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

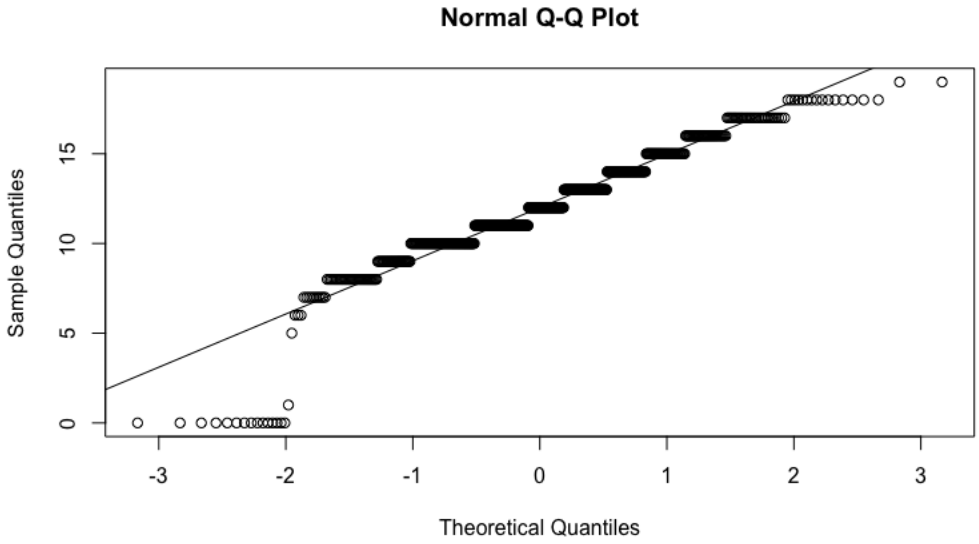
Conclusion/Correlation coefficient results in words: The two variables (Age & G3 grade for Mathematics) were weakly negatively correlated (-0.1615794), and the relationship was statistically significant, r(degrees of freedom) = 393, r (correlation coefficient) = -0.1615794, p = 0.001271. Therefore, we have sufficient statistical evidence to say that the correlation between the two variables is statistically significant.

Question 2: Is age correlated (using a Pearson R) to G3 grade for Portuguese language?

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the age and G3 columns of the Portuguese dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: Age is not correlated to the G3 grade for Portuguese language.

Research hypothesis: Age is correlated to the G3 grade for Portuguese language.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

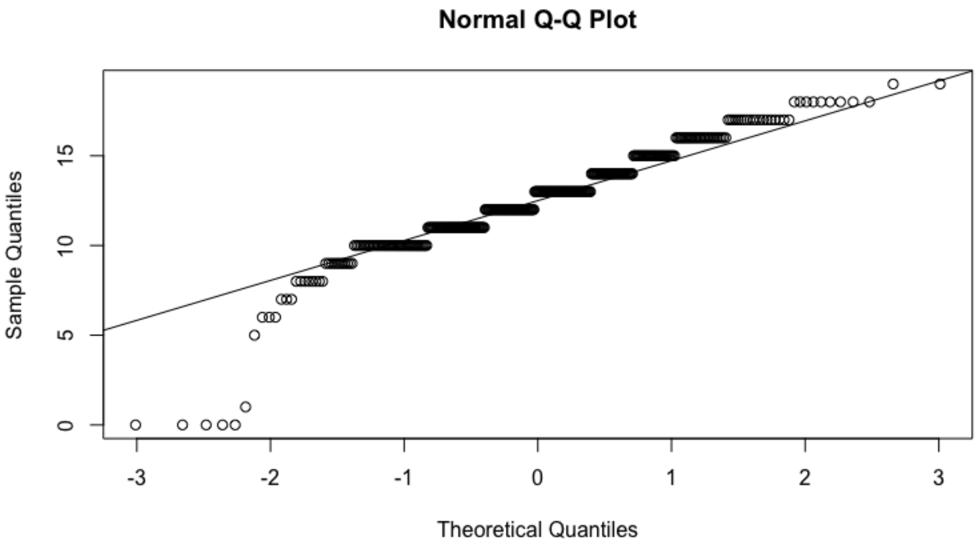
Whether the null was rejected and why: After performing the Pearson R test we got a p-value of 0.006612 which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

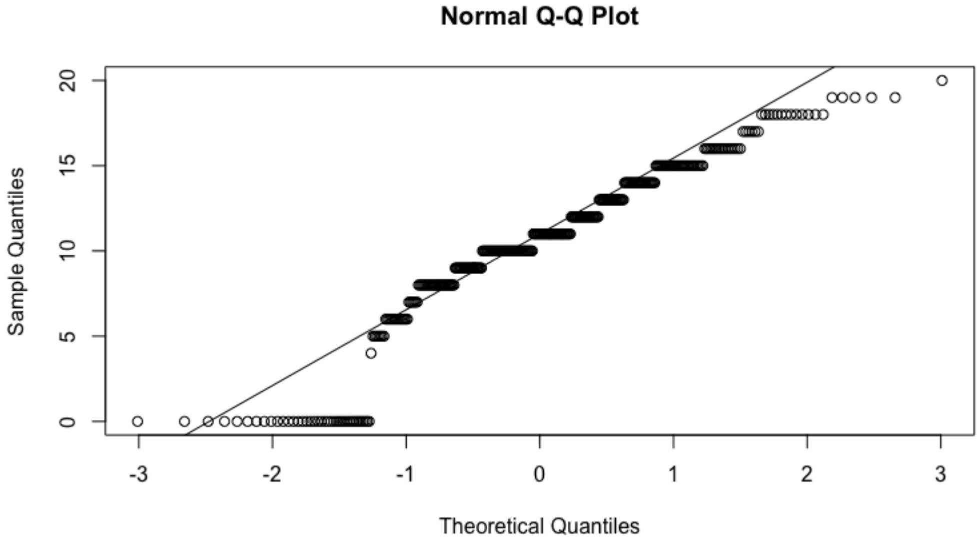
Conclusion/Correlation coefficient results in words: The two variables (Age & G3 grade for Portuguese language) were weakly negatively correlated (-0.1065054), and the relationship was statistically significant, r(degrees of freedom) = 647, r (correlation coefficient) = -0.1065054 , p = 0.006612. Therefore, we have sufficient statistical evidence to say that the correlation between the two variables is statistically significant.

Question 3: For the students appearing in both datasets, does the G3 grade in Portuguese language correlate to the G3 grade in Mathematics (using a Pearson R)?

Assumptions:

Normality





Comment: I used a QQ plot to determine the normality of the data for the G3 columns of the merged Mathematics and Portuguese dataset. After examining the QQ plots we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Hypothesis in words:

Null hypothesis: G3 grade in Portuguese language is not correlated to the G3 grade in Mathematics for the students appearing in both datasets.

Research hypothesis: G3 grade in Portuguese language is correlated to the G3 grade in Mathematics for the students appearing in both datasets.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing the Pearson R test we got a p-value less than 2.2e-16 or p-value < 2.2e-16 which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

Conclusion/Correlation coefficient results in words: The two variables (Portuguese language G3 grade & Mathematics G3 grade for students appearing in both datasets) were weakly positively correlated (0.4803494), and the relationship was statistically significant, r(degrees of freedom) = 380, r (correlation coefficient) = 0.4803494, p = p-value < 2.2e-16. Therefore, we have sufficient statistical evidence to say that the correlation between the two variables is statistically significant.

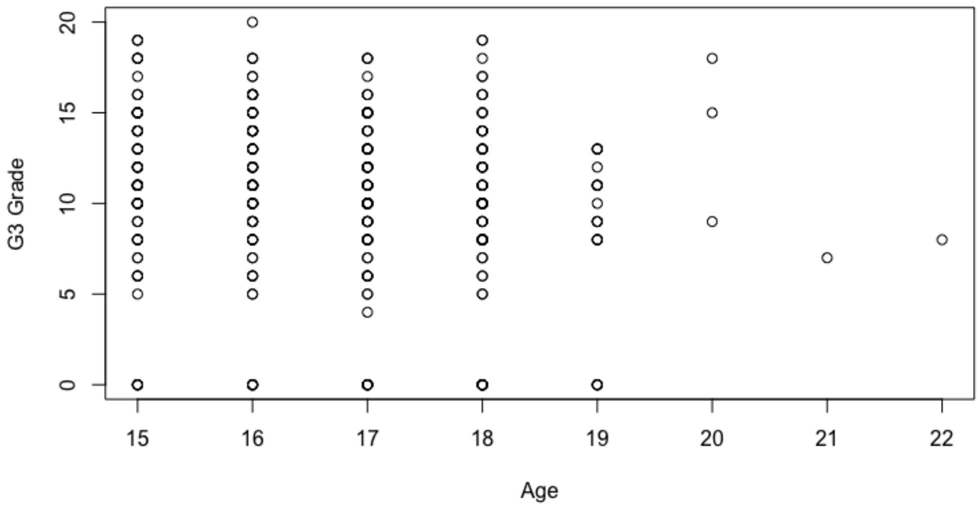
Section 4 Linear Regression Analysis

Question 1: Choose one of the correlational analyses you ran above to use for a linear regression analysis. What is your predictor variable and what is your response variable?

Chosen correlational analysis: Is age correlated to G3 grade for Mathematics?

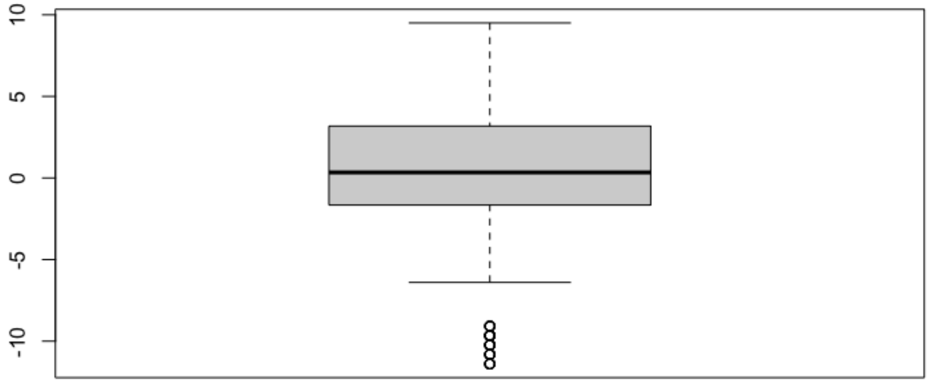
Predictor Variable: Age for Mathematics
Response Variable: G3 grade for Mathematics

Assumptions:

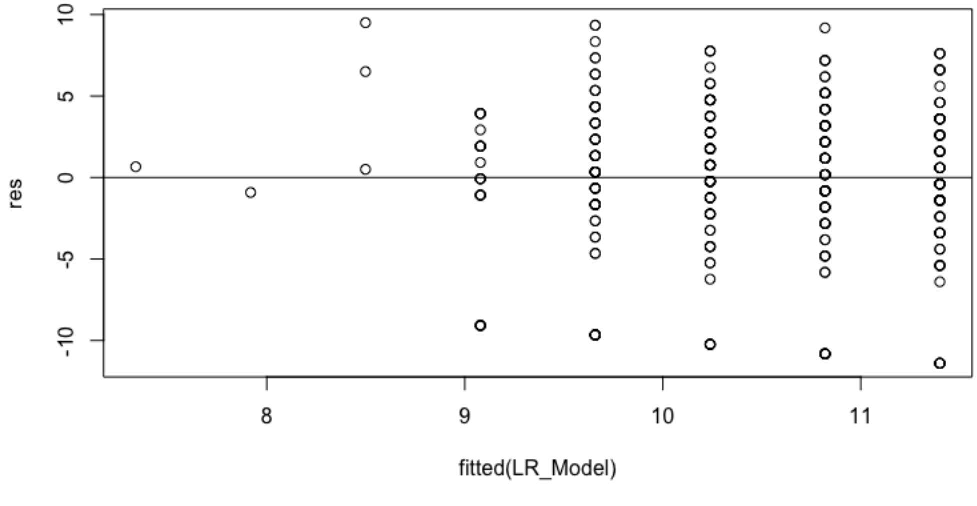


Scatter plot: : After examining the scatter plot, it seems to be that the relationship between the variables do not look linear, therefore this assumption does not pass, and the data has to be transformed.

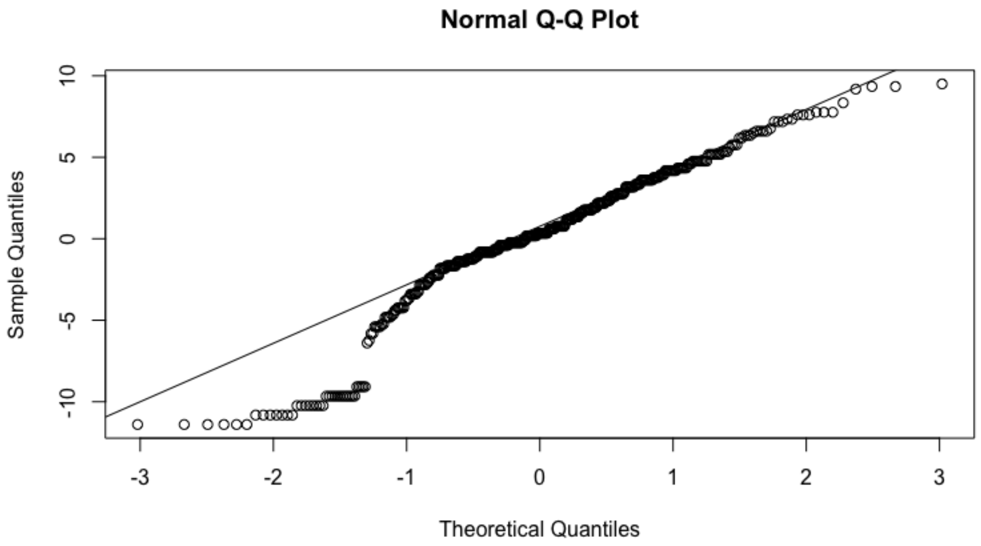
Note: Showed in officer hours, I tried transforming the data by taking the log of the predictor and response variables, but the data still seemed to be non-linear, and I was getting an error when using the transformed data in the linear regression function. I was told to use the non-transformed data for the linear regression and the rest of the assumptions.



Box plot: After running a box plot on the residuals, there are circles that appear with the plot which indicates outliers and therefore this assumption fails.



Checking residuals plots: checking homoscedasticity: After examining the plot we can see that the points form a vertical pattern, so this assumption fail.



Checking for normality using QQ plot: After examining the QQ plot we can say that the data is normally distributed since it clusters close and around the line with the exception of a few points, therefore, this assumption passes.

Checking for independence: We assume independence because the variables seem to be independent of each other and therefore, this assumption passes.

Hypothesis in words:

Null hypothesis: Age is not correlated to the G3 grade for Mathematics.

Research hypothesis: Age is correlated to the G3 grade for Mathematics.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: After performing a linear regression we got a p-value of 0.001271 which is less than our alpha value of 0.05. This means that we can reject the null hypothesis.

Reporting results mathematically:

$\hat{y} = b_0 + b_1x$

- **$\hat{y}$**: The estimated response value = G3 Grade Value
- **b_0**: The intercept of the regression line = 20.1011
- **b_1**: The slope of the regression line = -0.5801

Fitted regression equation: G3 Grade Value = 20.1011 + -0.5801 * (Age Value)

Reporting results in words:
The fitted regression equation for this model is: G3 Grade Value = 20.1011 + -0.5801 * (Age Value). This means that each additional Age value is associated with an average increase in G3 grade value of -0.5801. And the intercept value of 20.1011 tells us the average expected G3 grade for someone with an age of zero.

Section 5 ANOVA or Regression Model

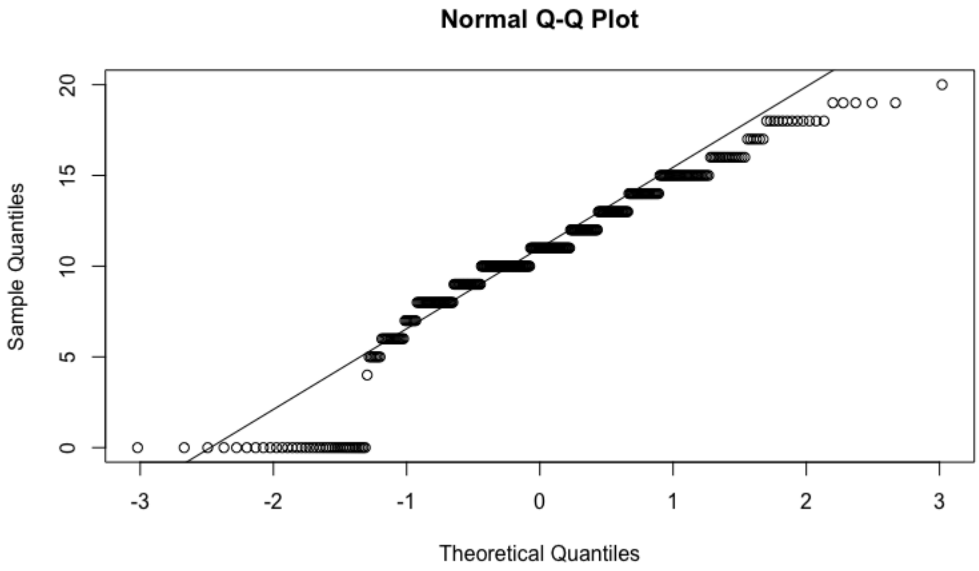
Question 1: Look at the dataset. Choose one or two categorical variable that you think might influence academic performance (remember that ordinal variables are categorical) and perform either an ANOVA or a regression model for qualitative variables. Report your results.

Chosen analysis: ANOVA
Chosen variables: freetime (free time after school), goout(going out with friends)

Perform a 2-way ANOVA examining the effect of free time after school and going out with friends on G3 grades for the Mathematics dataset.

Assumptions:

Normality



Comment: I used a QQ plot to determine the normality of the data for the G3 column of the Mathematics dataset. After examining the QQ plot we can conclude that the data is close enough to be normal with the exception of a couple of data points. Therefore, there is no need to do any transformation on the data and this assumption passes.

Independence: We assume independence because the variables seem to be independent of each other and therefore this assumption passes, and we do not need to do any transformation on the data.

Hypotheses in Words:

Omnibus Null Hypothesis: There is no difference between free time after school and going out with friends on G3 grades in the Mathematics dataset, our alpha is 0.05.

Research Hypothesis: At least free time after school or going out with friends has a significant effect on G3 grades in the Mathematics dataset.

a-level: The alpha(α) level is 0.05.

Rejection criteria: If the p-value is less than the alpha(α) level of 0.05, we reject the null hypothesis.

Whether the null was rejected and why: Based of the summary of the Best-Fit model which was the two-way ANOVA additive model the going out with friends (goout) Pr > f value is less than our alpha(α) of 0.05 which means that we do reject the null hypothesis.

Conclusion as show in class in words: We see that there is statistical significance in the G3 grades average in terms of going out with (f(4) = 4.284, p = 0.0021) friends. A Tukey post-hoc test revealed significant pairwise differences between the 5-2(goout level) pair, with an average difference of -2.24004177 (p = 0.0269973), and between the 5-3 (goout level) pair, with an average difference of -2.27349685 (p = 0.0168248). Based on the results we got we can say that the going out with friends level 2 and 3 were statistically significant when it came to G3 grades in the Mathematics dataset compared to the other levels of going out with friends and free time after school.