Sociological Factors Affecting Number of Days Students Spend Absent from School

TJ Pavaritpong, Vedaant Agarwal, Jacob Razdolsky, Jay Lim
University of Illinois Urbana-Champaign

12 December 2023

1 Introduction

The goal of this project is to investigate whether cultural origin, school level, type of learner, gender along with their interactions are related to the number of days a child is absent from school. In the context of this project, our research question is the following: What is the impact of cultural origin individually on the number of days absent from school?

2 Data Inspection

The data are obtained from a sociological study of Australian Aboriginal and white children reported by Quine (1975) [0].

Since all the predictors are categorical, we cannot perform the usual summary statistics. However, we can look at the frequency plot of the levels of each predictor. It was noted that race, gender and learner have two levels each and school has four levels. Moreover, the frequency distribution of these levels is pretty similar, except for school, which has a comparatively high number of "F1" school observations and comparatively low number of "F0" school observations.

Then, we created side by side box plots to better visualize relationship of individual predictors with the response. The plots can be viewed in Appendix [10] In terms of absences between aboriginal and non-aboriginal students, we see that the median number of absences is much higher for aboriginal students. Male students also seem to have higher median number of absent days compared to female students. Average learners had a slightly higher median number of absent days compared to slow learners, on

average, although there were more outliers in the slow learner group with high number of absent days. As for school levels, students in F1 level generally had a lower median number of absent days compared to students in all other levels. The median number of absent days of the F0 school level group was about the same as F2 and F3, although the IQR range of the F0 group is smaller than the other two. However, as of now we cannot say whether these median differences (or mean differences) are statistically significant.

3 ANOVA Model Fit

Throughout this study, we will use the default significance level of $\alpha = 0.05$.

Before fitting the multi-way ANOVA model, we need to check for the presence of interactions between the predictors. Some of the prominent 2-D interaction plots are shown in Appendix [9.1.2]. Looking at the interaction plots, it is seen that there are interactions between race, learner type, and school. Additionally, there seems to be interactions between gender, learner type, and school. However, there does not seem to be any interactions between gender and race. Furthermore, it is seen that there are interactions between the level of school and all the other predictors: race, learner, and gender. However, these were just 2-D interactions, and there would be more dimensions when all four predictors are considered together rather than just two. Since visualisations of 3D or 4D interactions becomes cumbersome, we will start with a full interaction model.

At this point it is important to note that since we have 4 predictors it is unlikely that every treatment in the dataset has an equal number of observations. Therefore, we will use an Unbalanced ANOVA model with Type III sum of squares. We then proceeded to fit our ANOVA models. We began with the full interactive model, and then began to reduce our model based on the p-value of the highest order interaction term. We removed each highest order variable term that had the highest p-value from the model one by one using ANOVA outputs, leading us to the final optimal ANOVA model fit shown in Table [1].

4 Model Diagnostics

For the model selected in Table [1] from removing the non-significant highest order interactions, we performed model diagnostics. We found that the normality assumption was violated at as seen from the Q-Q plot in Figure[8] as the points were not sitting along a straight line. Wee also performed a ks test, which confirmed that the normality assumption was violated. Additionally, from the residuals vs. fitted plot in Figure [9], it is seen that constant variance assumptions seem to be violated as well due to a non-

uniform, increasing funnel shape of the graph, and the Breusch-Pagan test confirmed this.

4.1 Remedial Measures

To alleviate the departure from normality and constant variance, we performed a Box-Cox transformation. However, before doing that, we must shift the values of the response by adding 1, because the response contains 0 as observations, and the transformation can only happen if the response is positive. The log-likelihood vs λ plot for the Box-Cox transformation is shown in Figure [10]. The λ value that maximised the LLF was 0.22, but for the sake of interpretability, we chose $\lambda = 0.25$, which is included in the 95 percent confidence interval. This was done for the sake of interpretability as it corresponds to transforming the response by a quartic root. After performing this transformation, the model was fit with the transformed response and the model assumptions were checked again. As seen in the new residuals vs. fitted plot in Figure [9]residuals vs. fitted plot in Figure [12], as well as the Breusch-Pagan test output of $p = 0.2908$, we conclude that the constant variance assumption is now met. However, as seen in the new Q-Q plot, the normality assumption is still not met, however, it has become much better from the original model. However, at this point we decided not to introduce any changes to the predictors to try and better the normality assumption to preserve the interpretability of our final model.

4.2 Unusual Observations

Additionally, checked for unusual observations could see that there were 50 high leverage points. After the plus one shift to the response variable, we could see that there were only 4 bad high-leverage points. Moreover, there were no highly influential points, and no outliers in the transformed model.

Finally, we checked the transformed model using ANOVA to test whether the higher order interaction is still significant, and the result stayed the same. Thus, we were not able to remove any further interaction terms.

5 Estimation of Treatment Means

The model we fit waas a factor effects model with sum constraints, to compute the estimators for effects and interaction terms. Since interaction terms were present, a point wise estimate might not be possible, but this can be checked using the Tukey's family confidence intervals.

5.1 95% Family-wise Confidence Interval

In order to establish a statistical confidence for statistical difference, we performed a family test of all pairwise differences for each factor separately from our final model using Tukey's test. From the Tukey's test, we could observe the p-value to diagnose whether the level is statistically different or not.

5.2 Individual Factors

For individual predictors in our final model, we could observe that, for all individual factors except race, the means of the factor levels were statistically the same, while the means for the factor levels were statistically different for the race factor. Note that this was for individual factors only and should not be extended to the entire model.

5.3 Interaction Terms between Factors

For interaction terms between two factors, most of the levels in the interaction terms between factors were observed to be statistically the same with some exceptions.

For example, for the interaction term between race and learner, the levels such as white and slow students and Aboriginal and average students; Aboriginal and slow students and white and average students; and white and slow and Aboriginal and slow students were observed to be statistically different.

For the interaction term between race and gender, levels like white and female student and Aboriginal and female student; Aboriginal and male student and white and female student; and white and male student and Aboriginal and male student were observed to be statistically different.

For the interaction term between race and school, levels such as Aboriginal and F2 school students and white and F0 school students; Aboriginal and F2 school students and white and F1 school students; and white and F2 school students and Aboriginal and F2 school students were observed to be statistically different.

However, for the interaction terms between learner and gender, between learner and school, and between gender and school, all of the levels were observed to have no statistical difference.

For interaction terms between three factors, race, learner, and school, all of the levels were observed to be statistically same, except for some levels like Aboriginal, slow, and F2 school students and white, average, and F0 school students; Aboriginal, slow, and

F2 students and white, slow, and F1 students; Aboriginal, slow, and F2 school students and white, average, and F2 school students; and white, slow, and F2 school students and Aboriginal, slow, and F2 school students which were observed to be statistically different.

5.4 Differences in Attendance due to Cultural Origin

Since interaction terms involving 'race' (cultural origin) and other factors are present, it is hard to quantify precisely the difference in attendance due to cultural origin alone. However, we can look at the means of specific factor level combinations involving 'race' that are significant. We also avoid interpreting the two-factor interaction terms because a three-factor interaction term is statistically significant here.

In the significant three-factor interaction term involving race, learner, and school, we observed that slow-learning aboriginal students in second year of secondary school had higher mean number of absent days compared to both slow and average-learning non-aboriginal students in second year of secondary school. The slow-learning aboriginal students in second year of secondary school also had a higher mean number of absent days compared average-learning non-aboriginal students in both primary school and the first year of secondary school.

5.5 Optimal Combination of Factor Levels

The optimal selection of factor levels was a bit complicated to do because our highest order interaction between all four predictors was not significant, and was excluded from the model. Due to this, we did not have an interaction term that included the levels of all four predictors, which meant we could not compare the treatment levels of all 4 predictors together. However, for the sake of finding an optimal combination of factor levels, we run a Tukey's test to check for which 4 predictor interaction had the greatest difference in factor level mean. Since we define an optimal combination as one which minimises the absence rate for this dataset, we will look for the lowest difference of means for factor levels in the Tukey's test. Doing this, we found that the optimal selection of factor levels to get the lowest absence rate was non-aboriginal, average learners, females, and school F2 for this dataset and model.

6 Additional Research Question

From the final model above, we observed the difference in attendance due to all the predictors. Throughout the statistical analysis, we were able to see that cultural origin did have the largest individual impact on the number of days absent, but due to all the

other interaction terms, we were not able to properly observe whether the cultural origin individually has an impact on the attendance.

So to answer the research question regarding the impact of cultural origin individually on the number of days absent from school, we created a single linear regression model where we had race as the predictor and the number of days absent as the response to observe whether the race as a predictor is statistically significant or not. We also performed a log transformation on this model to ensure that normality assumptions were met, which was seen with a normal QQ-plot with all points falling closely on the line.

After fitting the model, checking model assumptions, applying a transformation to fix the Normality assumption, we conducted a one-way ANOVA test. We observed that there is a difference between the Aboriginal students and white students in how they affect the number of days absent for this data set. Also, from conducting a Tukey's test for the difference of factor means, we were able to witness that Aboriginal students tend to have higher number of absent days from school than white students for this data set.

7 Conclusion

We observed that Australian aboriginal students tend to have a higher number of absences when compared to Australian non-aboriginal (white) students in this data set, when only the race was taken into consideration.

We also expect that on average, the group that would have the lowest absence rate for this study when taking everything into consideration, were females, non-aboriginal students in second year of secondary school who were average learners.

For the overall study, when considering the race, school, gender, and type of learner of an Australian student in this dataset, we could not say anything about how race individually affects the absence rate because how race affects the absence rate depends on school, gender and type of learner.

8 References

S. Quine (1975), Achievement Orientation of Aboriginal and White Adolescents. Doctoral Dissertation, Australian National University, Canberra.

9 Appendix

9.1 Data Visualization

9.1.1 Box-plots

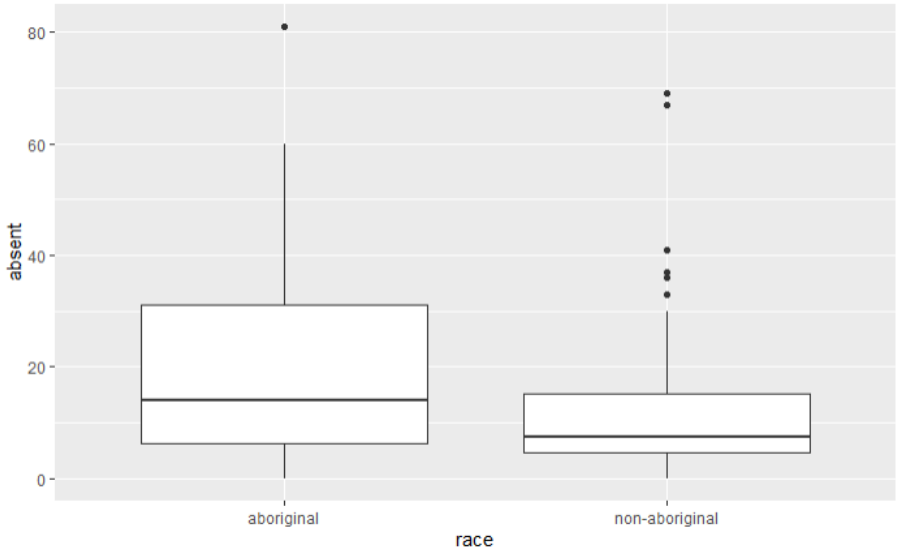

Figure 1: The Box-plot for race predictor vs. absent response.

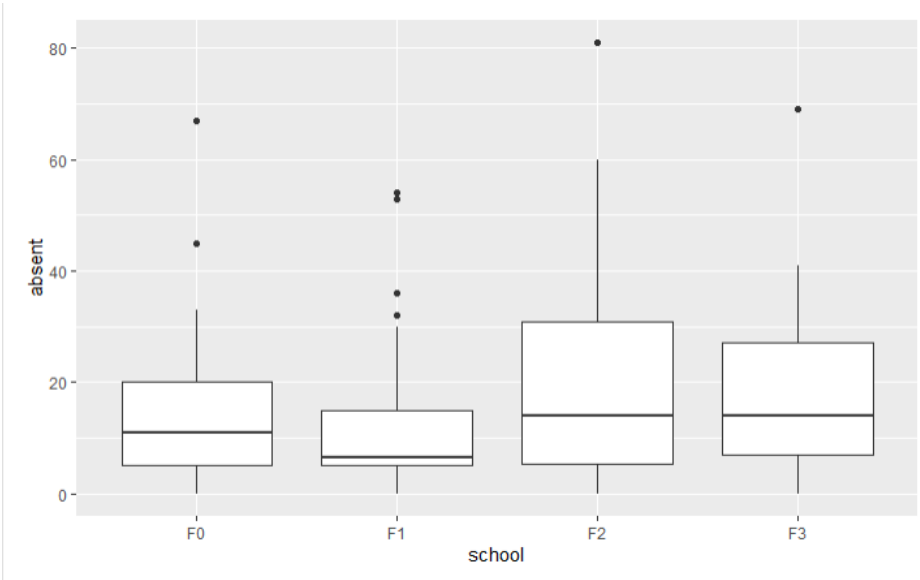

Figure 2: The side by side Box-plot for school predictor vs. absent response.

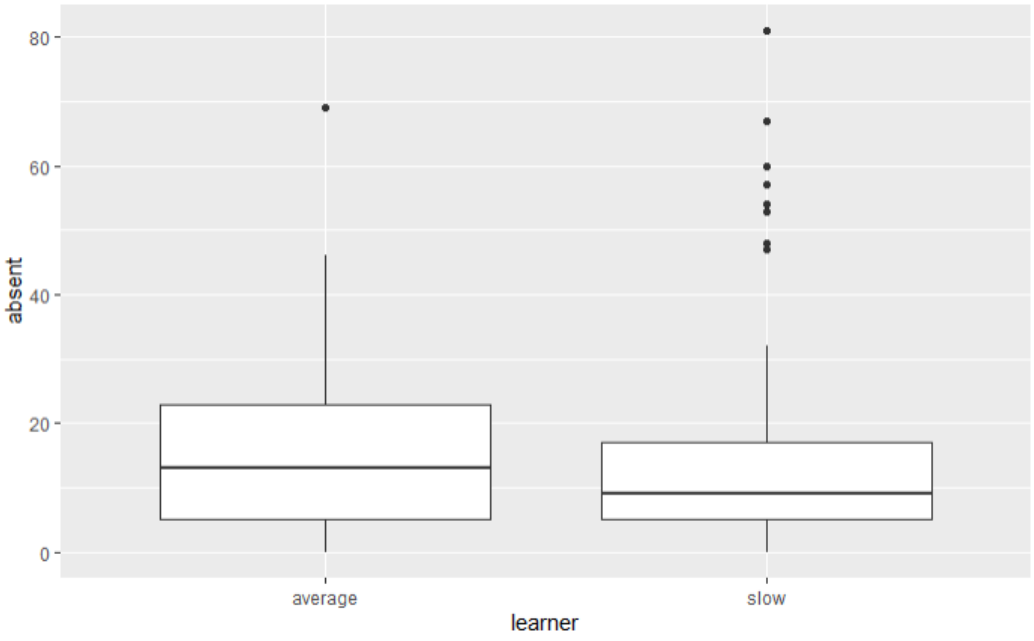

Figure 3: The side by side Box-plot for learner predictor vs. absent response.

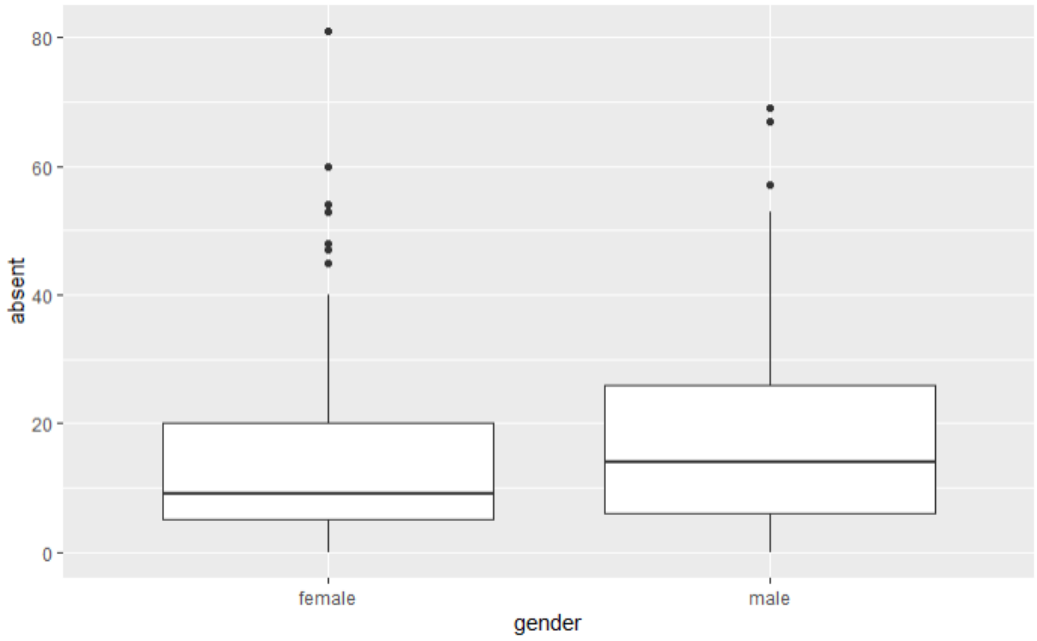

Figure 4: The side by side Box-plot for gender predictor vs. absent response.

9.1.2 Interaction Plots

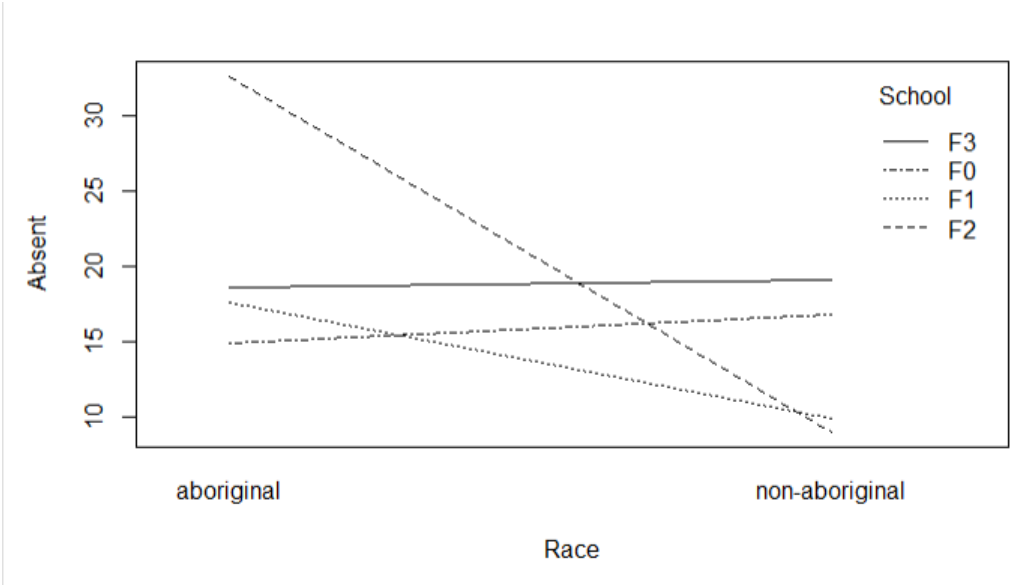

Figure 5: The interaction plot for race and school.

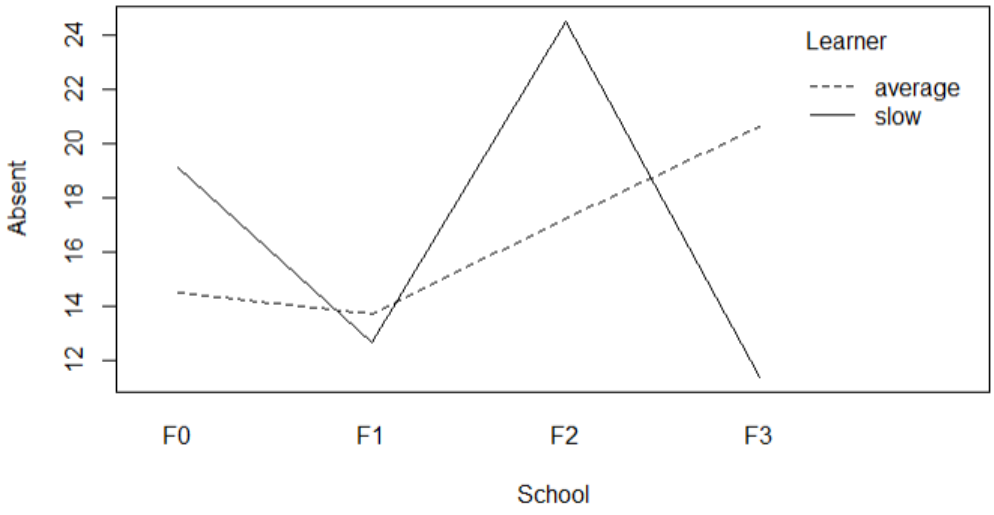

Figure 6: The interaction plot for school and learner.

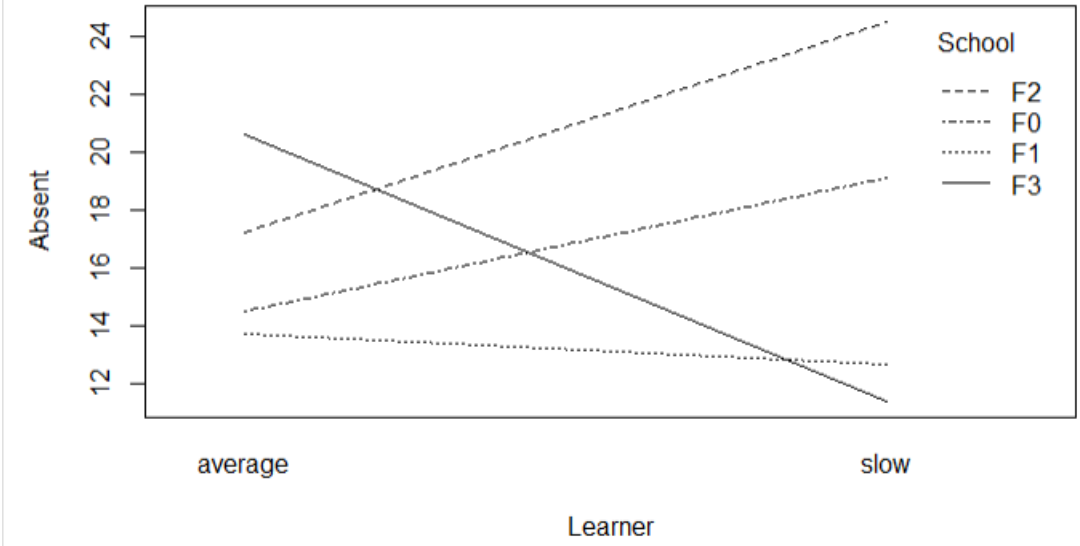

Figure 7: The interaction plot for learner and school.

9.2 Model Selection

9.2.1 Before Box-Cox transformation

```
Call:
lm(formula = absent ~ race * learner * gender * school - race:learner:gender:school -
    race:gender:school - learner:gender:school - race:learner:gender,
    data = notinschool)

Residuals:
    Min      1Q  Median      3Q     Max
-34.504  -8.505  -1.694   6.098  46.377

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                               20.960      5.835   3.592 0.000461 ***
racenon-aboriginal                        -7.986      7.026  -1.137 0.257795
learnerslow                               -4.542      9.290  -0.489 0.625730
gendermale                                -7.729      6.381  -1.211 0.228004
schoolF1                                  -9.460      7.788  -1.215 0.226639
schoolF2                                  -5.498      9.463  -0.581 0.562205
schoolF3                                  -5.639      6.875  -0.820 0.413591
racenon-aboriginal:learnerslow            26.408     11.759   2.246 0.026379 *
racenon-aboriginal:gendermale              3.771      5.061   0.745 0.457523
learnerslow:gendermale                    -4.163      5.578  -0.746 0.456822
racenon-aboriginal:schoolF1                9.277      9.477   0.979 0.329424
racenon-aboriginal:schoolF2              -11.439      9.564  -1.196 0.233815
racenon-aboriginal:schoolF3                5.826      8.057   0.723 0.470920
learnerslow:schoolF1                      13.752     10.940   1.257 0.210939
learnerslow:schoolF2                      23.703     11.132   2.129 0.035092 *
learnerslow:schoolF3                      -3.849     11.165  -0.345 0.730836
gendermale:schoolF1                        6.478      7.443   0.870 0.385706
gendermale:schoolF2                       17.772      8.607   2.065 0.040903 *
gendermale:schoolF3                       18.566      7.154   2.595 0.010523 *
racenon-aboriginal:learnerslow:schoolF1  -41.716     14.545  -2.868 0.004810 **
racenon-aboriginal:learnerslow:schoolF2  -34.771     15.593  -2.230 0.027441 *
racenon-aboriginal:learnerslow:schoolF3  -26.475     16.339  -1.620 0.107548
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.9 on 132 degrees of freedom
Multiple R-squared:  0.3409,    Adjusted R-squared:  0.2361
F-statistic: 3.252 on 21 and 132 DF,  p-value: 1.934e-05
```

Table 1: The model after deleting the highest order non-significant interactions.

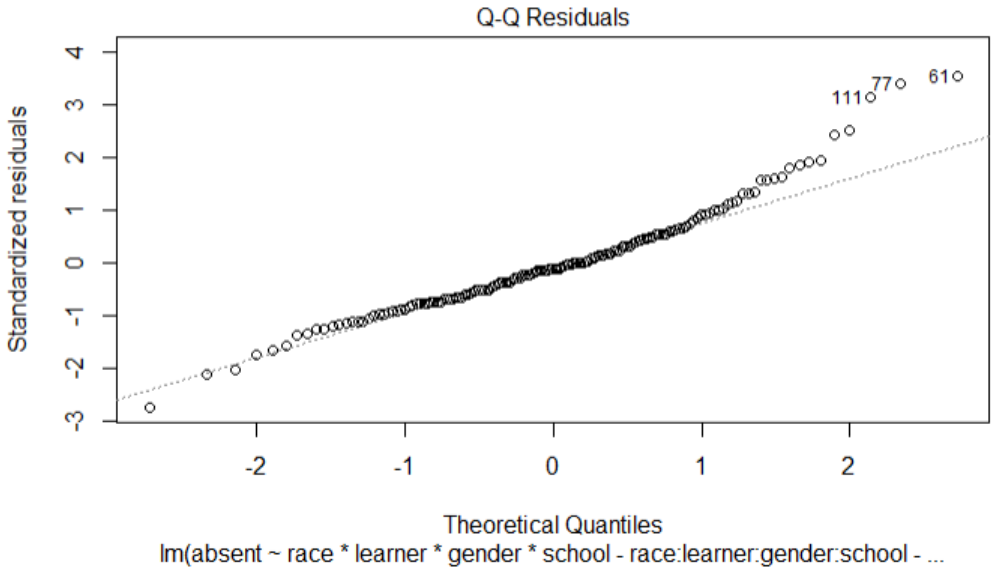

Figure 8: QQ-plot before the Box-Cox transformation.

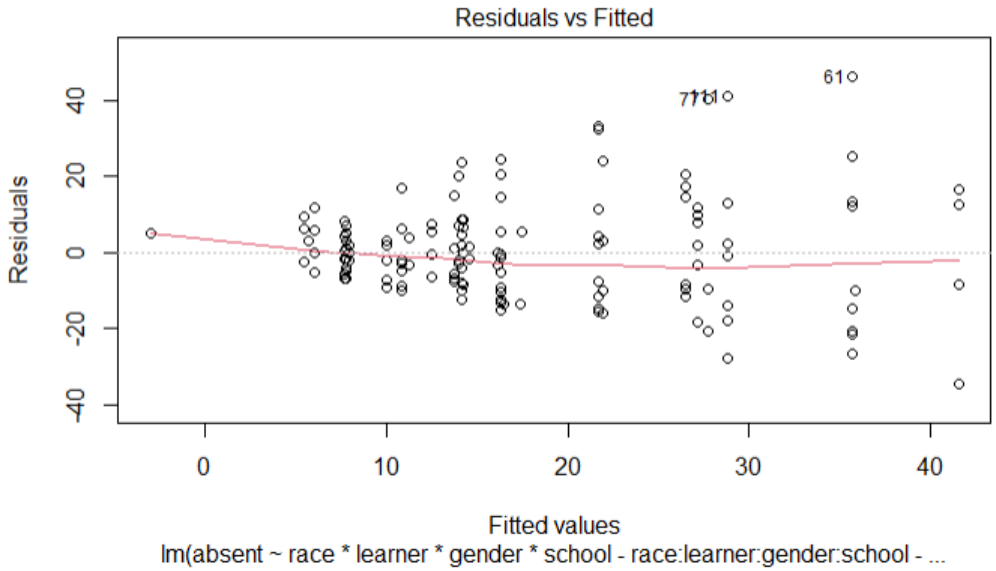

Figure 9: The residual vs. fitted plot before the Box-Cox transformation.

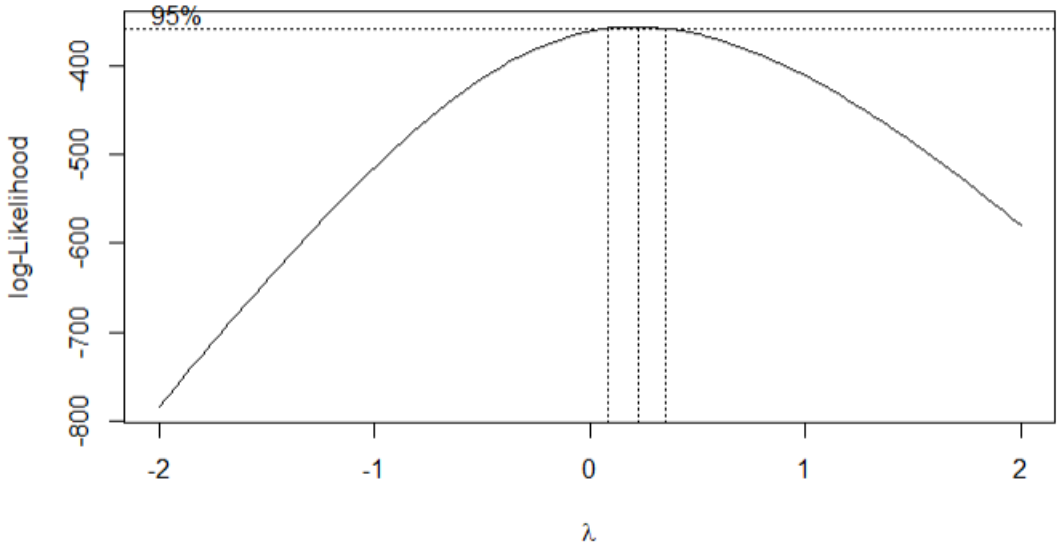

Figure 10: The Box-Cox Plot along with the 95 percent confidence interval for λ. The maximum value was about 0.22

9.2.2 Final Model

```
Call:
lm(formula = absent^(1/5) ~ race * learner * gender * school -
    race:learner:gender:school - race:gender:school - learner:gender:school -
    race:learner:gender, data = notinschool)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84785 -0.18907 -0.00007  0.20403  0.51793

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1.80281    0.11964  15.068  < 2e-16 ***
racenon-aboriginal                        -0.17097    0.14409  -1.187  0.23752
learnerslow                               -0.13234    0.19050  -0.695  0.48846
gendermale                                -0.16573    0.13086  -1.266  0.20758
schoolF1                                  -0.18718    0.15970  -1.172  0.24330
schoolF2                                  -0.24164    0.19404  -1.245  0.21523
schoolF3                                  -0.18464    0.14099  -1.310  0.19260
racenon-aboriginal:learnerslow             0.58047    0.24113   2.407  0.01745 *
racenon-aboriginal:gendermale             -0.02110    0.10378  -0.203  0.83917
learnerslow:gendermale                    -0.08564    0.11439  -0.749  0.45535
racenon-aboriginal:schoolF1                0.18333    0.19435   0.943  0.34726
racenon-aboriginal:schoolF2               -0.20547    0.19612  -1.048  0.29671
racenon-aboriginal:schoolF3                0.14735    0.16522   0.892  0.37411
learnerslow:schoolF1                       0.27286    0.22434   1.216  0.22604
learnerslow:schoolF2                       0.47963    0.22828   2.101  0.03753 *
learnerslow:schoolF3                       0.02726    0.22896   0.119  0.90540
gendermale:schoolF1                        0.18267    0.15263   1.197  0.23353
gendermale:schoolF2                        0.47650    0.17650   2.700  0.00785 **
gendermale:schoolF3                        0.44013    0.14670   3.000  0.00323 **
racenon-aboriginal:learnerslow:schoolF1 -0.90077    0.29827  -3.020  0.00304 **
racenon-aboriginal:learnerslow:schoolF2 -0.66586    0.31976  -2.082  0.03924 *
racenon-aboriginal:learnerslow:schoolF3 -0.52730    0.33506  -1.574  0.11794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2851 on 132 degrees of freedom
Multiple R-squared:  0.3337,    Adjusted R-squared:  0.2277
F-statistic: 3.148 on 21 and 132 DF,  p-value: 3.278e-05
```

Table 2: The final model after Box-Cox transformation.

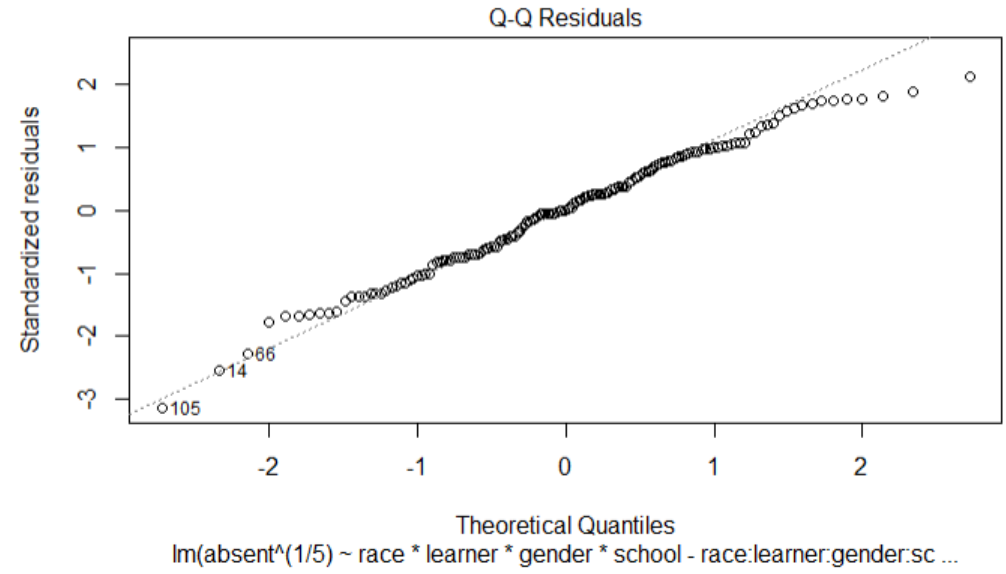

Figure 11: QQ-plot after the Box-Cox transformation.

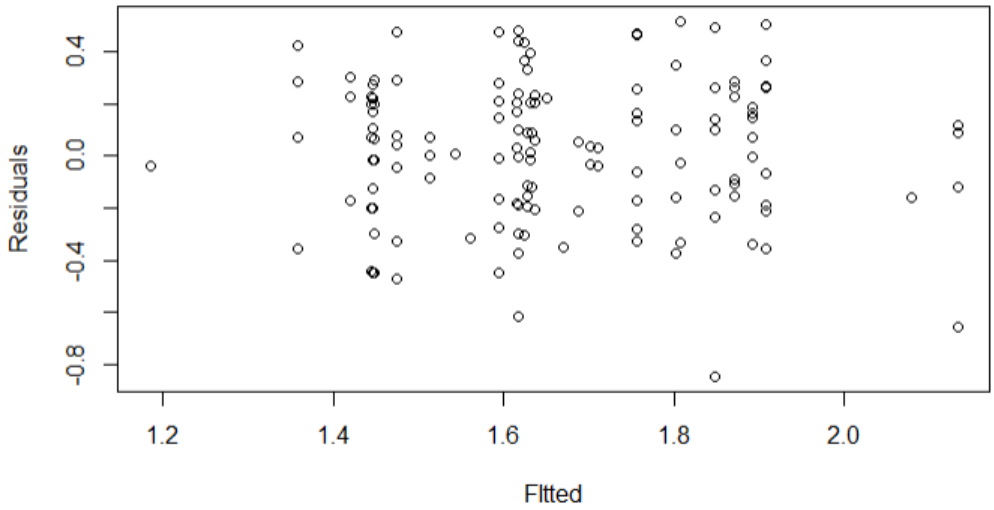

Figure 12: The residual vs. fitted plot after the Box-Cox transformation.

9.3 Model Diagnostic

9.3.1 Some 95% Family-wise Confidence Level

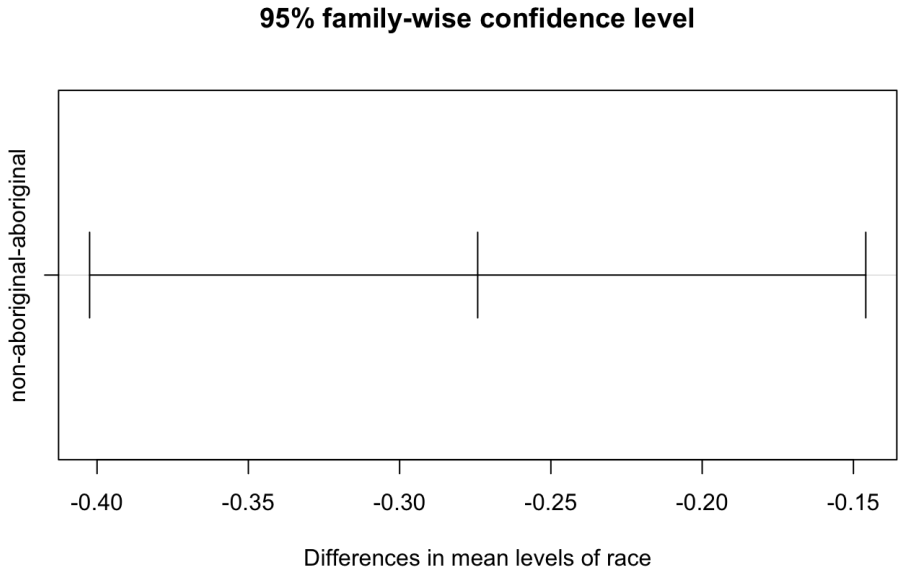

Table 3: The 95% Family-wise Confidence Level for race factor.

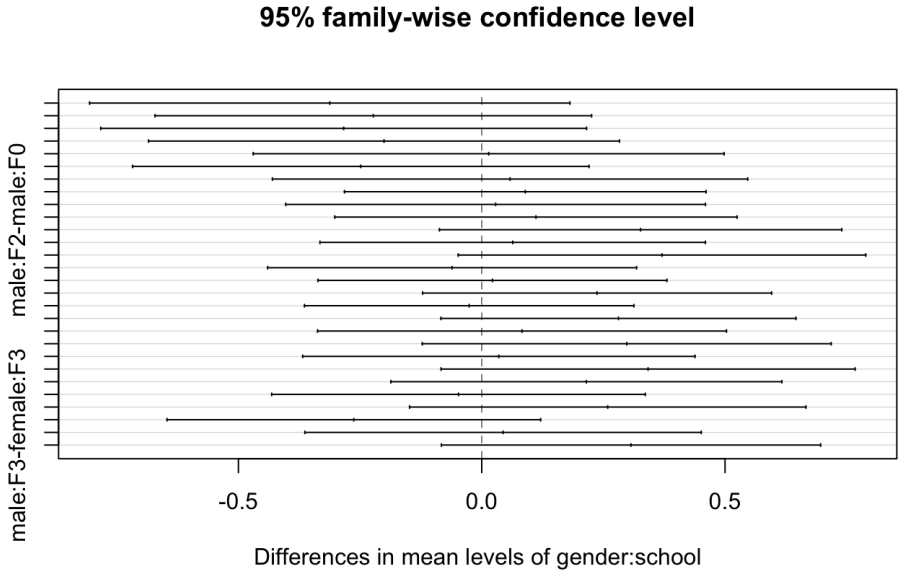

Table 4: The 95% Family-wise Confidence Level for gender and school factors.

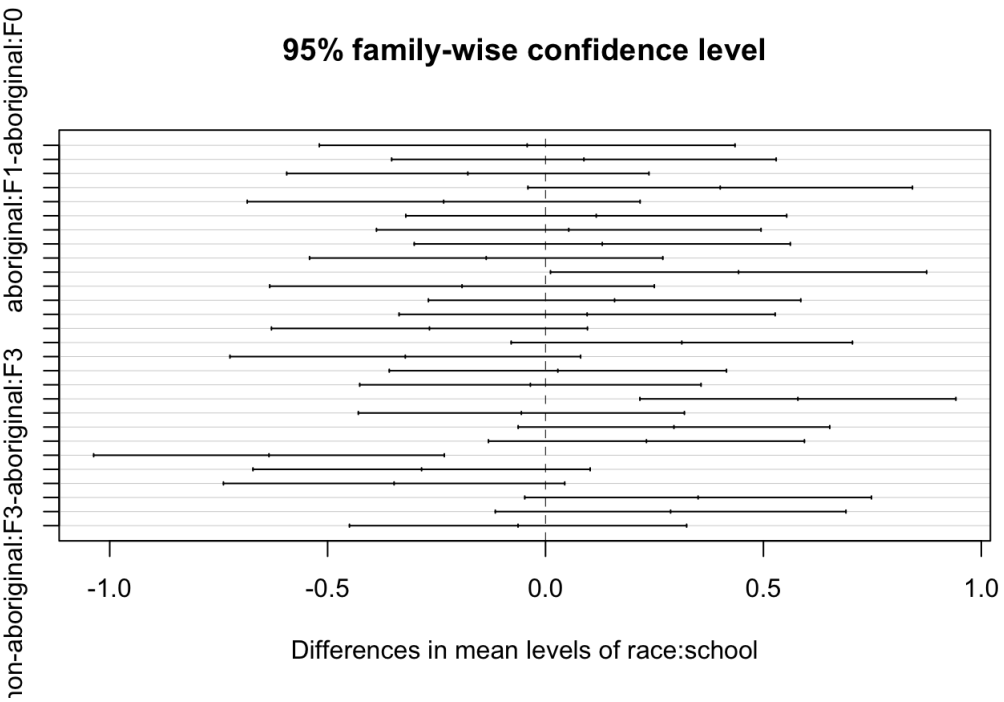

Table 5: The 95% Family-wise Confidence Level for race and school factors.

9.4 Additional Research Question

9.4.1 Model Selection

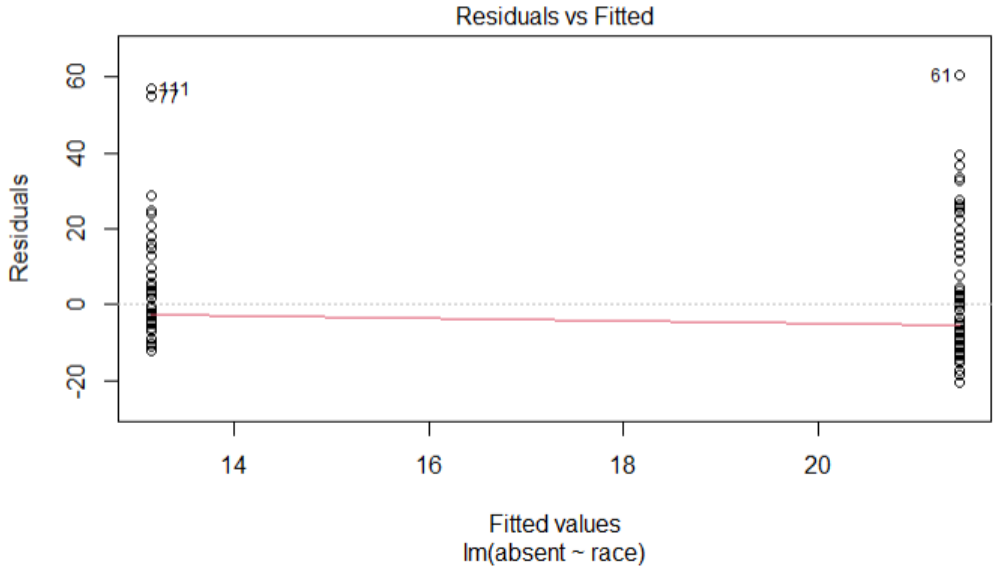

Table 6: The residual vs. fitted plot before the Log transformation for the single linear regression.

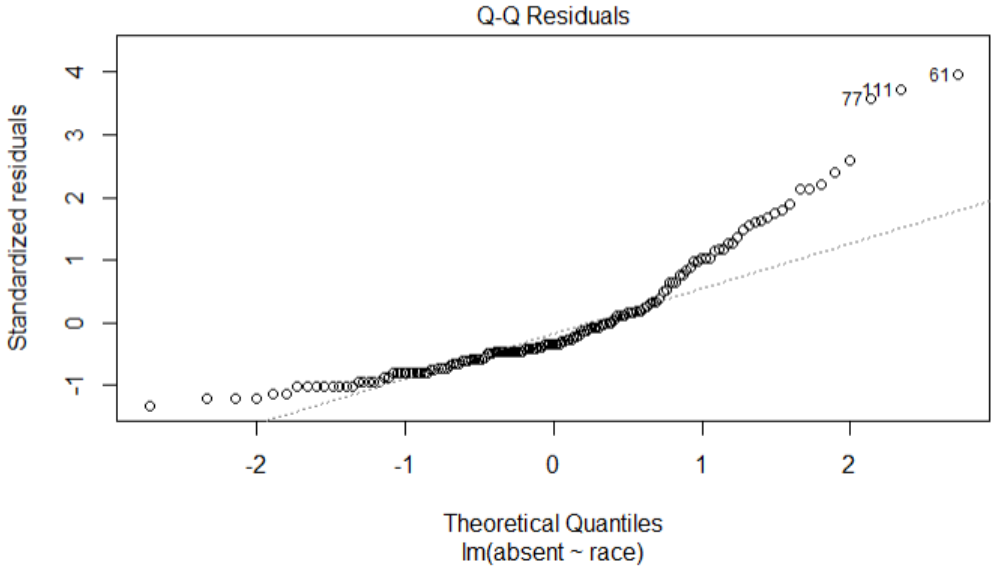

Table 7: QQ-plot before the Log transformation for the single linear regression.

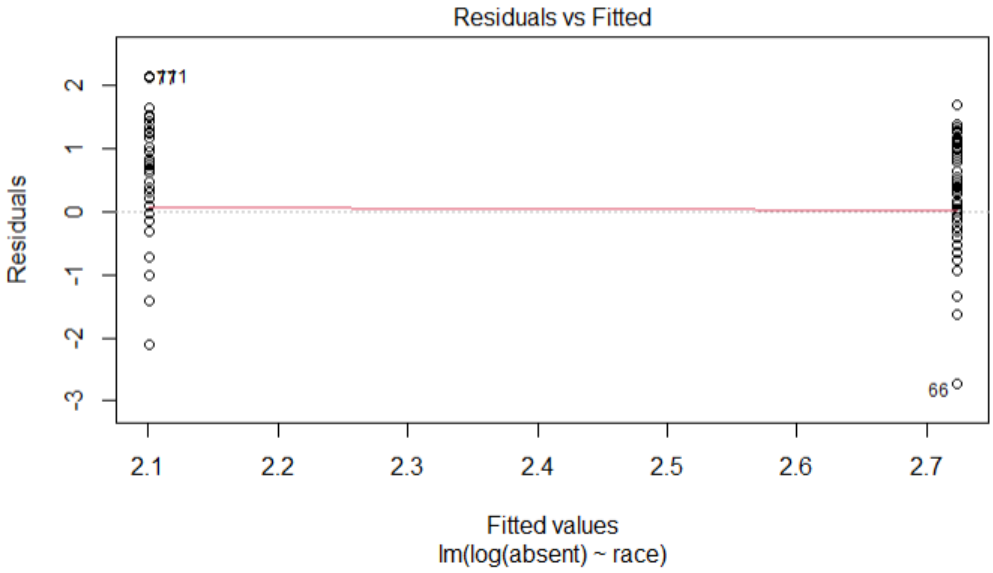

Table 8: The residual vs. fitted plot after the Log transformation for the single linear regression.

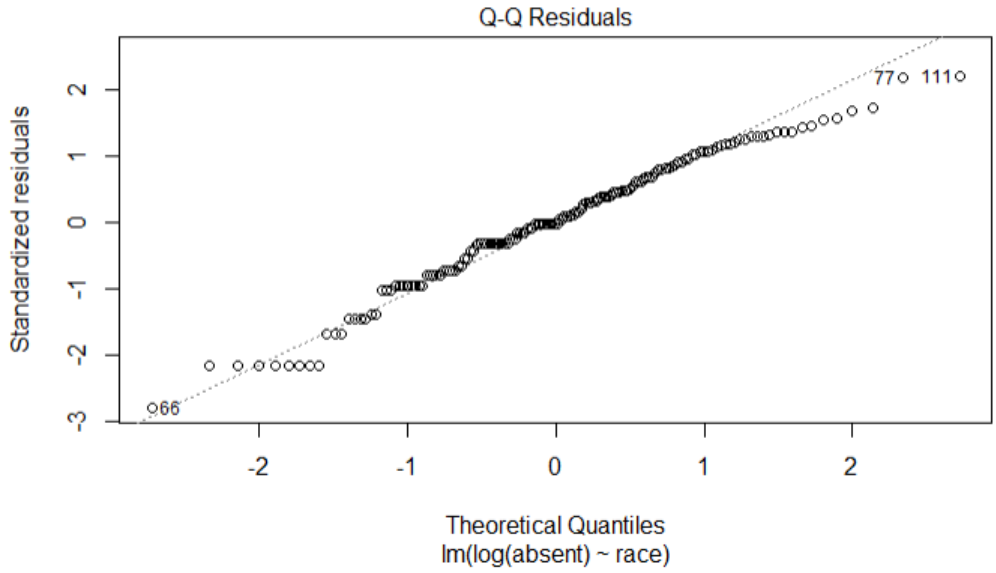

Table 9: QQ-plot after the Log transformation for the single linear regression.

9.4.2 Final Model

```
Analysis of Variance Table

Response: log(absent)
           Df  Sum Sq Mean Sq F value    Pr(>F)
race        1  14.889 14.8889  15.521 0.000124 ***
Residuals 152 145.814  0.9593
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 10: One-Way Anova test, Race as the predictor and Absent as the response.

9.4.3 95% Family-wise Confidence Level

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = absent ~ race, data = notinschool)

$race
                                diff       lwr       upr     p adj
non-aboriginal-aboriginal -8.295946 -13.20404 -3.387849 0.0010561
```

Table 11: The 95% Family-wise Confidence Level for individual race factor.