Cameron Kerkemeyer and RJ Burjek

Alexandra Chronopoulou

STAT 425

8 December 2023

School Absence Case Study

When it comes to human beings, a primary education is vital during the early stages of life. Children need to learn things like basic functions and human behavior along with ways to communicate with others. Sometimes this learning is done at home, but for many, this primary education comes within a classroom setting. For those who do their learning in schools, consistent attendance is very important or else children may face the risk of falling behind in their education, thus putting them at a disadvantage in comparison to their peers. However, maintaining consistent attendance can be challenging for some. Children may have difficulties commuting to and from school, and extraordinary cases can cause unexpected absences as well. This idea is what leads us to our study today. External factors like those mentioned prior are commonly known to prevent children from attending school, but we would like to study whether personal and cultural traits influence that attendance as well. To answer this question, we explored data from a study done by Susan Quine (1973) on the academic attendance during a given school year of Australian aboriginal and white children. For each student represented in the data set, we are given their cultural orientation, gender, level of school, and the type of learner they are along with the number of days they were absent from school throughout the year. Using those first four predictors, we would like to see if those qualities have any influence on the number of days that a child is unable to attend school.

Beginning our study, we looked at the effect that a child's race had on their number of absences. Within the variable "race," children fell into one of two levels: aboriginal and non-aboriginal. We started off by creating a box plot of the number of days a student was absent in each respective cultural group. We noticed a similar mean number of absences between our aboriginal and non-aboriginal groups, but we also found there to be a slight difference in the distribution and variation of the two groups. Following this, we fit a linear model by regressing race against our absent variable. After deriving a linear model, we began to check for any influential points within our model. We started out by checking for any high leverage points and found there to be none. Next, we checked for outliers. Using a Bonferroni correction value and comparing them to our residual values, we found there to be three outliers at observations 61, 77, and 111. Lastly, we used Cook's Distances to check for highly influential points and found there to be none of those as well. Finishing our model exploration, we test for equality of means using Scheffe's method of comparisons. With our test yielding negative lower and upper bounds for a 95% confidence interval and a p-value that is less than our significance level of 0.05, we conclude that non-aboriginal students are expected to have a lower number of absences when compared to the aboriginal students.

Next, we looked at the effect that the interaction between a child's race and their gender had on the number of times they were absent from school. Each student in our study was categorized as either male or female. This comes with the child's race still being decided by aboriginal and not aboriginal. As in the first model, we started off with the box plots of each predictor. The box plot for race has not changed since the first model. However, when viewing the boxplot for gender, it could be seen that the mean value was approximately the same, with males having a greater distribution, and females having a greater number of outliers. Following the box plots, we then found the interaction plot for each possible interaction. In the interaction plots, we found no intersecting lines, thus indicating there should not be interactions present. It could then be seen in the four plots that each level for the predictors matched each other either positively or negatively. With that in mind, it could also be seen that the distributions are relatively different between levels such as aboriginal versus non-aboriginal, and female versus male. We then transformed the response variable in the model to absence plus one due to zeros in our data preventing us from further analysis of the model being studied. We then found the summary statistics and the type-three ANOVA table of the model. This allowed us to find the Normal Q-Q Plot as well as the fitted versus residuals plot. We could see that the Normal Q-Q Plot was not quite linear, and that the fitted-residuals plot did not have a constant variance. We then ran another Box-Cox plot with the transformed model to find that the model satisfied the necessary assumptions. To improve the normality and linearity, we then attempted an additive model of gender and race versus absence. We then conducted the same summary statistic and ANOVA summaries as the transformed mode and then ran the Normal Q-Q Plot and fitted versus residual plots. We found that both the normality and linearity had not noticeably improved and that could be seen as well in the Box-Cox plot for the additive model as well. We then moved on to finding the 95% interaction confidence interval which we used the Tukey comparison test to find. It was calculated that the confidence interval for the additive model was statistically significant due to the p-value being less than our 0.05 threshold.

After exploring the interacting effect of race and gender, we studied the effect that a student's level in school has on their absences as well. Again, we created a box plot to look at the variation in absences for each age group. The students in the data were split into four different grade levels denoted by the terms F0, F1, F2, and F3. When analyzing the box plot for levels of school versus the number of absences, each group had similar means and distributions of days absent from school. After this, we regressed the interaction of our predictors race, gender, and level in school against the number of absences and derived an ANOVA table for our model. Using a significance level of 0.05, we removed the predictor of least significance one by one until all our predictors were statistically significant. This left us with a linear model containing the predictor race and the interaction term between level of school and race. Next, we checked our model for normality. We did this by creating a Q-Q plot and running a Shapiro-Wilk test. In our Q-Q plot, the residuals did not fall along a linear path, and our Shapiro-Wilk test yielded a p-value less than our 0.05 threshold, so we determined that our model was not normally distributed. In order to fix this, we used a Box-Cox transformation to normalize the distribution of our data. Going forward, this transformed model took place as our final model for testing. Next, we tested for any influential points within our new model, starting off by looking for any high leverage points. After finding the leverages of each point, we found there to be no high leverage points within our transformed model. Then, we checked for outliers within our model, and when sorting the residuals of each point and comparing them to our Bonferroni correction value, we found there to be two outliers at observations 77 and 111. Lastly, we checked for highly influential points by finding the Cook's Distances of each given observation. Following our calculations, we found there to be no highly influential points within this model. The last thing we did in our model exploration was testing the comparisons between each of our model predictors. Using a Tukey comparison test, we derived 95% confidence intervals for each pairwise comparison of predictors, but none of those confidence intervals yielded a p-value below our 0.05 threshold and were therefore not statistically significant.

Lastly, we investigated the effect that all four of our predictors had on the number of times a student was absent from school. Each child in the data set was categorized as either an average learner or a slow learner, and this data was stored under the variable "learner." In the box plot of our variable learner and our variable absent, we observe that the two groups within our learner category display similar means and distributions of absent values. Our next step is creating interaction plots for each pairwise combination of our four predictors. Doing this, we observe interaction between school level and race, school level and gender, and school level and type of learner. Next, we fit a linear model of the interaction between all four of our predictors. Subsequently, we used a stepwise shrinkage method in order to reduce the model to a form that was most statistically significant. Our final linear model included the predictors gender, learner, race, school, learner:race, gender:school, learner:school, race:school, and learner:race:school. After observing a non-linear pattern in our Q-Q plot and a p-value less than our 0.05 significance level in our Shapiro-Wilk test, we determined that our data was not normally distributed and used a Box-Cox transformation to adjust our model. After deriving this new transformed model, we tested for any influential points within our model. We started off by testing for high leverage points. Upon calculating the leverages of each of our data points, we found there to be no high leverage points in our model. Next, we checked for outliers. When comparing our residuals to our Bonferroni correction value, we found there to be two outliers at observations 66 and 105. Lastly, we looked at the Cook's Distances of each of our points and found there to be no highly influential points in our model. The last step in our model exploration was testing the comparison of means between each of our model's predictors. Using a Tukey comparison test, we found 95% confidence intervals for each of the pairwise comparison of predictors, but none of the confidence intervals yielded a p-value below our 0.05 threshold and were therefore not statistically significant.

Overall, through this case study, we were able to successfully investigate the relationship of the number of days a child is absent from school with cultural origin, type of learner, gender, and school level along with their interactions. We investigated using Box-Cox plots, Kolmogorov-Smirnov Tests, Breusch-Pagan Tests, interaction plots, 95% confidence and interaction intervals, Tukey Tests, and ANOVA methods. By using these methods, we found that our final model to be a transformed model of (absence + 1) ^ (-1) with predictors race + gender + learner + school. In the context of the research question, we found that the amount of school days that a child missed, plus one and to the negative first power, is directly related to the race, gender, type of learner, and school level of the child. This model serves as a good prediction of the number of school days that a child missed but could be validated and improved through continued testing and greater data volume.

References

Quine, S. (1973). *Achievement Orientation of Aboriginal and White Australian Adolescents*. Doctoral Dissertation, Australian National University, Canberra.