Projects - Analyses of Two Independent Datasets

Hospital Glove Use & Elementary School Attendance

Elijah Castro

March 24, 2023

Contents

Hospital Glove Use - Cardiology Department

Summary

Data containing information about a study involving 23 members of a hospital's cardiology department nursing staff are here analyzed using generalized linear model techniques to investigate an experiment conducted on the effect of an educational program on compliance with glove use. Our goal focuses on determining whether the educational program on the importance of using gloves improve glove use in heart valve surgeries and whether it depends on the years of experience. Using a quasi-binomial regression model, we found that the different periods of observation and a nurse's experience are significant predictors of improving glove use. However, our findings also suggest interaction terms do not play a role in improving glove use, indicating that the amount of years of experience for a nursing staffer is not dependent for improving glove use. These insights hold valuable information for those in the hospital looking to improve the glove use rates for their nursing staff.

Introduction

Data involving the first couple members of the cardiology department's nursing staff are displayed below along with a description of each variable included in the experiment.

Table 1: Data of First Two Observed Nurses

	Period	Observed	Gloves	Experience
1	1	2	1	15
2	2	7	6	15
3	3	1	1	15
5	1	2	1	2
6	2	6	5	2
7	3	11	10	2
8	4	9	9	2

Variable Description:

1. *Period*: Observation period (1 = before intervention, 2 = one month after intervention, 3 = two months after, 4 = 5 months after intervention)
2. *Observed*: Number of times the nurse was observed
3. *Gloves*: Number of times the nurse used gloves
4. *Experience*: Years of experience of nurse

Our objective here is to produce a framework that accurately describes the relationship between the independent variables, *Period* and *Experience*, and the response, (*Gloves*, *Observed* - *Gloves*), for the nursing staff in this experiment. We will achieve this by uncovering the regression model that best fits the data provided. Some of the methods employed to complete this objective include exploratory data analysis, model-building, diagnostics, and when relevant, identifying and accounting for outliers and influential points.

Procedure

To begin, a few preliminaries are explored. This includes exploratory analyses of nursing staff's glove use across each observation period, fitting an initial binomial regression model, exploring possible interaction terms with the predictors, and examining diagnostic plots. This is used as a starting point to see what can be improved.

Next, we use our preliminaries to investigate possible signs of over-dispersion. This describes observing variation that is higher than would be expected in our data set. If the data itself has an over-dispersion problem, we should consider alternative models, such as quasi-likelihood methods. We will compare different models using Chi-squared test to ensure we are building the best model to fit this data set.

After fitting the final model, we again investigate interaction terms and diagnostics. Lastly, we check for potential influential points in the final model. If any exist, findings will be reported without the influential observation present in the data set.

Note: To view the entire process of finding the final model, click on the link below:

In-Depth Approach to Find Best Model

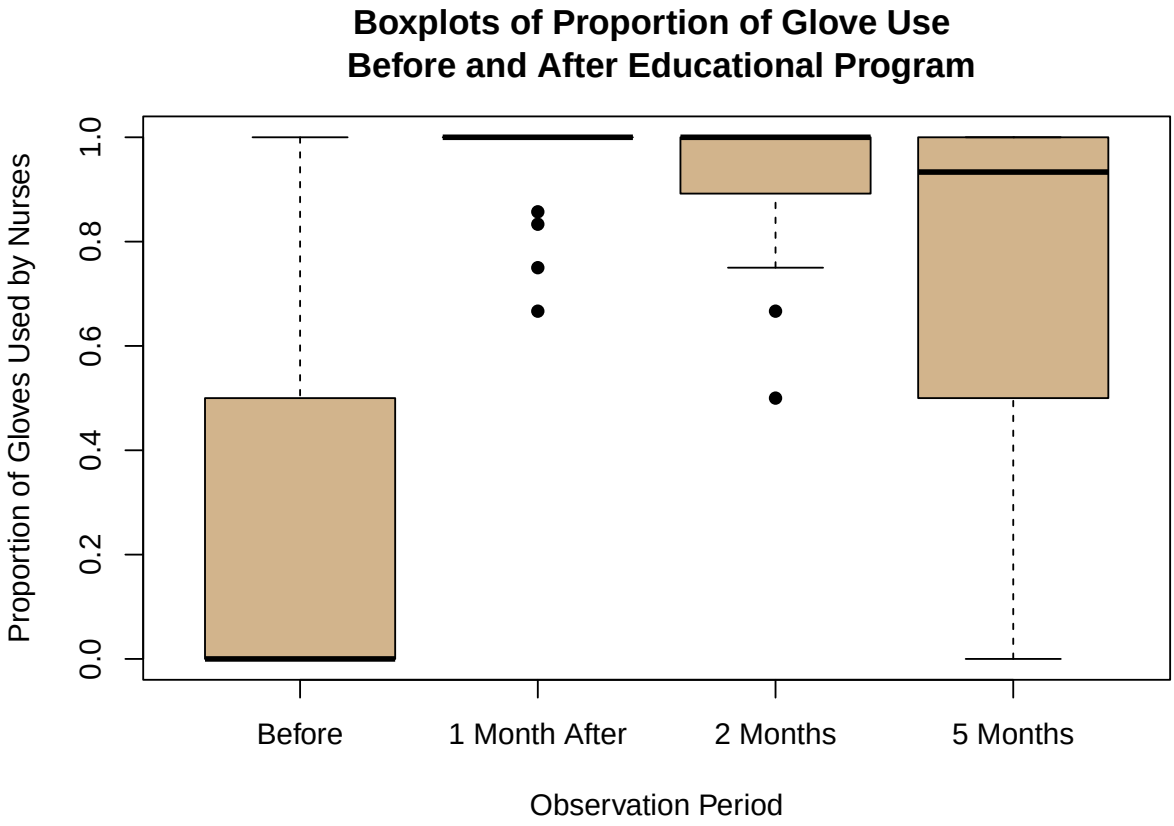
Findings

Here, we examine exploratory relationships in the data set and discuss findings from the final model. The following changes were made from the initial model in the final model:

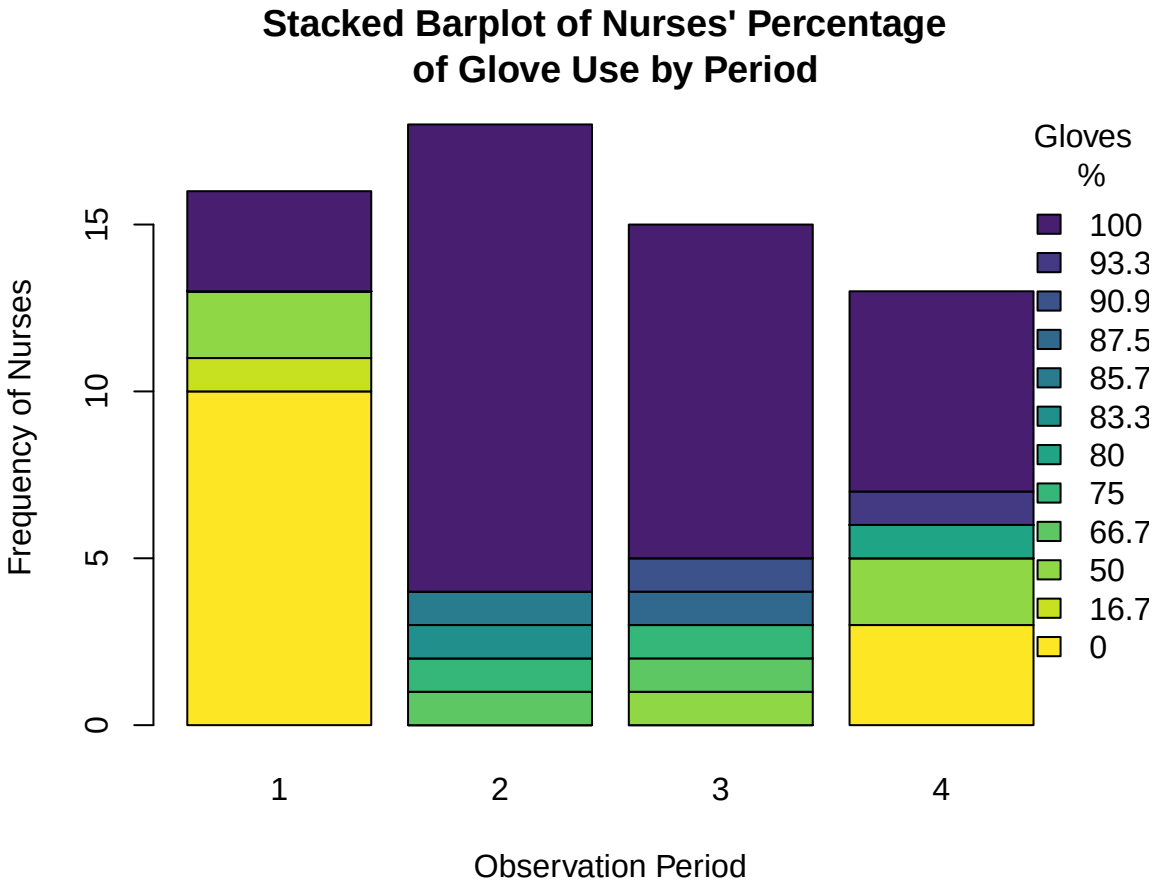
- Fit a quasi-binomial regression model after discovering signs of over-dispersion in the data.
- Removed the interaction term since it was no longer significant in the quasi-binomial model.
- Identified an influential point, so findings will be reported without this point present.

Exploratory Relationships

First, we display boxplots that show the proportion of gloves worn during each observation period. This visual informs us that before the educational program, the mean number of proportion of gloves used in heart valve surgeries was close to 0. Then, a month to 2 months after the program, we see that mean number jump to 1, indicating all nursing staff start using gloves. By 5 months after the program was presented to the nurses, we note that the mean dips slightly, and the size of the box increases dramatically. This indicates that unlike in periods 2 and 3, where the vast majority of nurses had their proportion of gloves used close to 1, in period 4, 25% to 75% of nurses had their proportion of gloves used range from 0.5 to 1. This implies by this period, more nurses have began using less gloves in heart valve surgeries.



Now, we focus on displaying a stacked barplot of nurses based on percentage of glove use per observation period. In this visual below, we find that before the educational program, a huge majority of nurses never wore gloves in heart valve surgeries. However, by periods 2 and 3 (1-2 months after the educational program was presented), the vast majority of nurses now wear gloves 100% of the time. Then, by period 4, 5 months after the educational program, the number of nurses with high percentages of glove use falls significantly. We also note that there is a return of some nurses who completely stop wearing gloves for the first time since period 1.



In summary, we can see from both exploratory plots that the mean number of times nurses used gloves increased from before the program (Period 1) to after the program (Periods 2, 3, and 4). However, it is still unclear whether this increase is due to the educational program or some other factor.

Final Model

Next, we display the model summary for the final model.

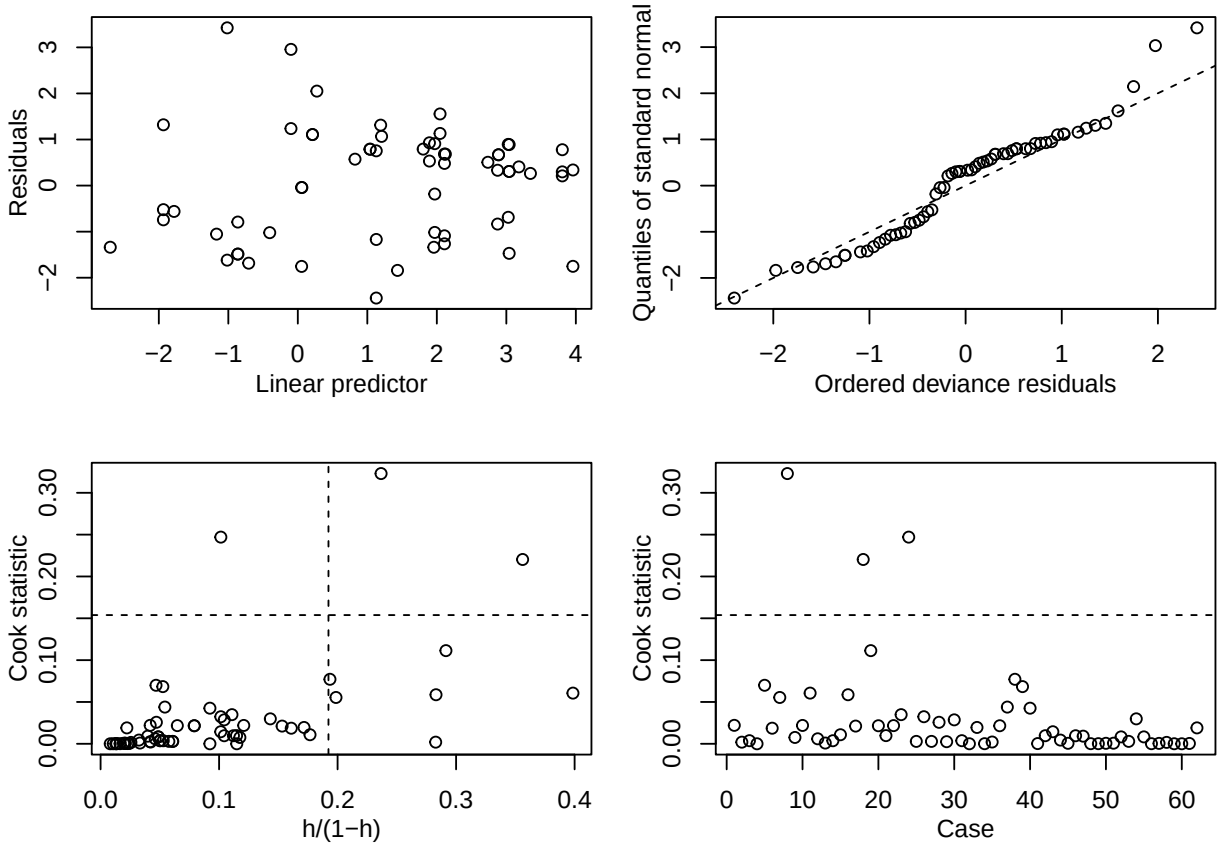
Table 2: Model Summary - Quasi-Binomial Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.3605	0.5229	0.6894	0.4934
Period2	3.904	0.7342	5.317	1.833e-06
Period3	2.973	0.6544	4.543	0.00002931
Period4	1.992	0.5751	3.463	0.001019
Experience	-0.153	0.04255	-3.595	0.0006787

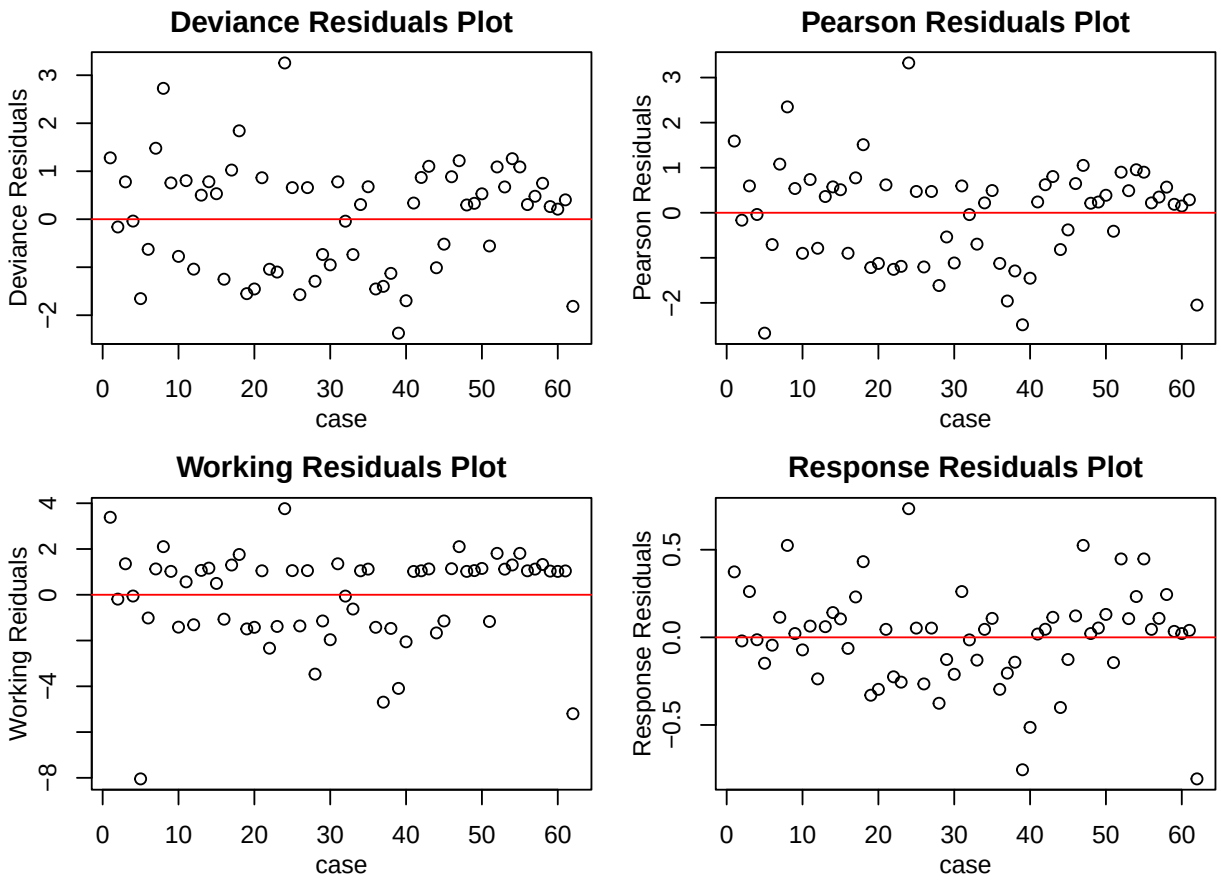
Without regarding the intercept, we notice that all the independent variables are significant at the 5%

level. We also note the effect of the educational program does not depend on the amount of years of experience since there is no interaction term present. This means regardless of the years of experience of the nurse, all of them are positively impacted by the educational program on the importance of glove use.

Diagnostics



From the diagnostic plots, we can confirm that this final model fits well with the data set. Residuals show random scatter (homoscedasticity), ordered deviance residuals follow a straight line, and cook's statistic values are very low.



Additionally, looking at all four different residual plots, we note that they all show random scatter (homoscedasticity). This further confirms the notion that this model represents this experiment well.

Analysis

Instead of trying to interpret the coefficient estimates on the logit scale from the model summary, we can instead take the exponential of the estimates to get the odds ratio and its respective 95% confidence interval. We note that the 95% confidence intervals for all odds ratios are statistically significant since none of the intervals include 1 (not including `intercept`).

Table 3: Odds Ratios of Coefficient Estimates with Corresponding 95% Confidence Interval

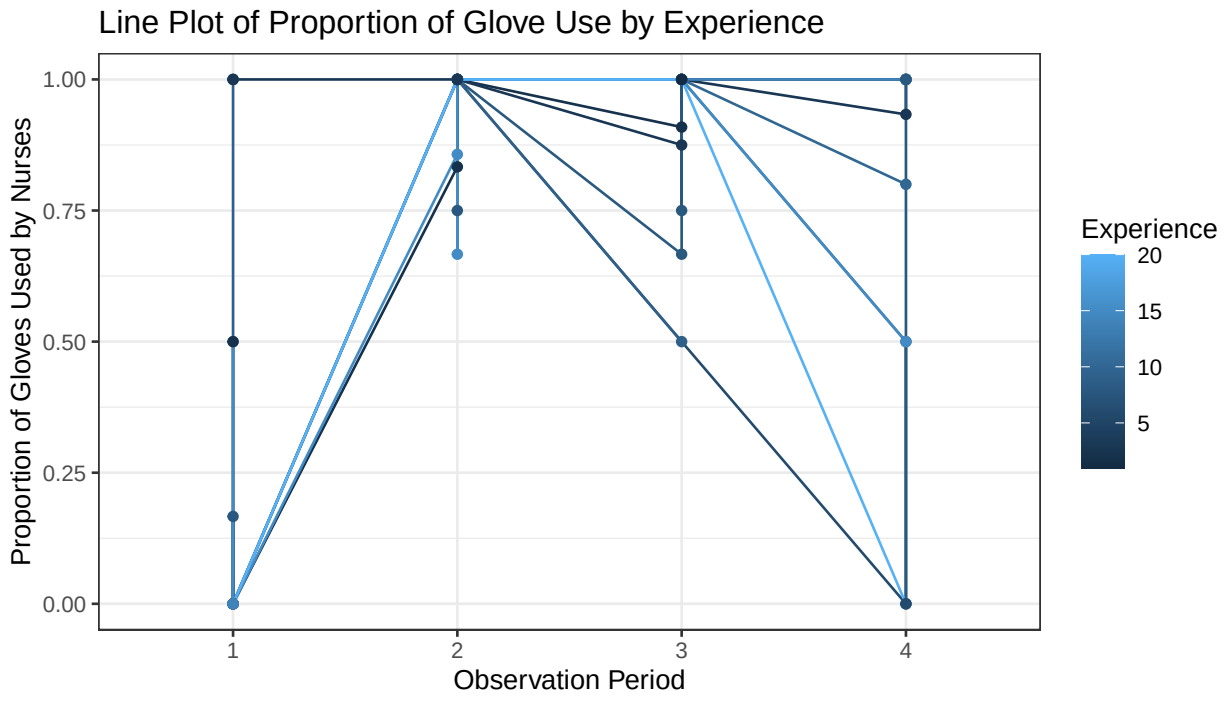
	Estimate	Lower_Limit	Upper_Limit
(Intercept)	1.435	0.5148	3.995
Period2	49.6	11.76	209.1
Period3	19.55	5.419	70.53
Period4	7.33	2.375	22.62
Experience	0.8581	0.7898	0.9324

The interpretation of the odds ratio for the model's coefficient estimates are as follows:

- **Period 2**: The estimated difference in odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries one month after intervention is 49.6 times higher than before the intervention.
- **Period 3**: The estimated difference in odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries two months after intervention is 19.55 times higher than before the intervention.
- **Period 4**: The estimated difference in odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries five months after intervention is 7.33 times higher than before the intervention.
- **Experience**: For a one unit increase in years of experience, the odds of a nurse successfully using gloves when they are observed without their knowledge during heart valve surgeries decreases by 0.8581. In terms of the change in odds, this means that each additional increase of one year in experience is associated with about a 14.2% decrease in the odds of a nurse successfully using gloves when they are observed without their knowledge during heart valve surgeries, holding all other variables fixed.

As previously discussed, we know the effect of the educational program does not depend on the years of experience. That is, no matter the experience the nurse has in the cardiology department, we know the educational program causes them to increase their glove use during heart valve surgeries. However, we still do not know how to quantify the effect the educational program has based on experience-level of the nursing staff over time.

Thus, we will display the effect of the observation period on glove use by experience level using a line plot. This will give us a visual representation of how the effect of the educational program differs based on years of experience over time:



From the line plot, we first note that nurses with less years of experience have darker blue lines, and nurses with more years of experience have lighter blue lines.

Then, we can see that the more experienced nurses start with essentially 0% of gloves used before the educational program and increase their proportion to 1 for periods 2 and 3. However, by period 4, we find that proportion falls back down to 0, indicating the educational program does not necessarily have any long-term effects for experienced nurses.

On the other hand, for less experienced nurses, we also see the similar jump from period 1 to 2 and 3. By period 4, however, the drop in proportion of glove use is nowhere near as dramatic as more experienced nurses, indicating that long-term effects of glove use for less experienced nurses are more noticeable.

In summary, even though all nurses are affected by the educational program on the importance of glove use, we found out the long-term effects of the program indeed differs based on years of experience.

Conclusions

In conclusion, educational programs can result in a clinically significant increase in glove use by cardiology department registered nurses. However, long-term improvement was less pronounced for the group of more experienced registered nurses. If we were to extend this research further, some possibilities to remedy this issue could be a continual use of these programs to constantly reeducate and remind nurses of the importance of using gloves after a designated span of time.

Elementary School Attendance Behavior

Summary

Data containing a random selection of 316 6th graders from two elementary schools in the district is here analyzed using generalized linear model techniques to investigate how the number of days of absence depends on the remaining variables provided. Our goal focuses on identifying the components that influence absent numbers, including the school they attend, their gender, and their standardized test scores (math and language). Using a negative binomial regression model, we found that the student's elementary school they attend, gender, and math scores are all significant predictors of attendance behavior. Additionally, our findings also indicate that multiple interaction terms help explain a more comprehensive attendance pattern among students, providing valuable information for educators in this district looking to improve attendance rates.

Introduction

Data on the first few 6th graders are shown below along with a description of each variable included in this data set.

Table 4: Data of First Five 6th Graders

school	gender	math	language	absence
1	M	56.4	44.1	4
1	M	36.3	48.2	4
1	F	31.9	42.7	2
1	F	28.9	42.2	3
1	F	6.3	28.6	3

Variable Description:

1. *school*: An indicator of two schools
2. *gender*: Male ("M") or Female ("F")
3. *math*: Standardized math test score
4. *language*: Standardized language test score
5. *absence*: Number of days of absence

Our objective here is to produce a framework that accurately describes the relationship between the independent variables, *school*, *gender*, *math*, and *language*, and the response, *absence*, for the 6th graders in these elementary schools. We will achieve this by uncovering the regression model that best fits the data provided. Some of the methods employed to complete this objective include exploratory data analysis, model-building, and diagnostics.

Procedure

To begin, a few preliminaries are explored. This includes examining descriptive statistics, exploratory analyses of 6th graders absences across gender and school, fitting an initial poisson regression model, exploring possible interaction terms with the predictors, and examining diagnostic plots. This is used as a starting point to see what can be improved.

Next, we use our preliminaries to investigate possible signs of over-dispersion. This describes observing variation that is higher than would be expected in our data set. If the data itself has an over-dispersion problem, we should consider alternative models, such as negative binomial regression. We will compare different models using likelihood ratio tests to ensure we are building the best model to fit this data set.

After fitting the final model, we again investigate interaction terms and diagnostics. Lastly, we attempt to interpret coefficient estimates to help paint the entire picture of attendance behavior for these two elementary schools.

Note: To view the entire process of finding the final model, click on the link below:

In-Depth Approach to Find Best Model

Findings

Here, we examine exploratory relationships in the data set and discuss findings from the final model. The following changes were made from the initial model in the final model:

- Fit a negative binomial regression model after revealing signs of over-dispersion in the data.
- Included two interaction terms to the final model.
- Removed the language predictor since it was no longer significant in the negative binomial model.

Descriptive Statistics

We begin with some descriptive statistics to get an overview of the data:

Table 5: Math Summary Data

school	gender	Mean Scores	Variance of Scores
1	F	42.03	291.5
1	M	42.36	387.1
2	F	55.98	257.5
2	M	54.56	169.1

Table 6: Language Summary Data

school	gender	Mean Scores	Variance of Scores
1	F	44.38	232.6
1	M	41.88	323.5
2	F	60.16	245.2
2	M	53.6	270.4

Table 7: Absences Summary Data

school	gender	Mean Scores	Variance of Scores
1	F	10.42	99.41
1	M	6.186	42.76
2	F	3.64	30.82
2	M	3.221	22.86

From the summary statistics, we can see that for school 1, mean scores for math and language are much lower for both genders than school 2. Additionally, mean absences for both genders are higher in school 1 than in school 2.

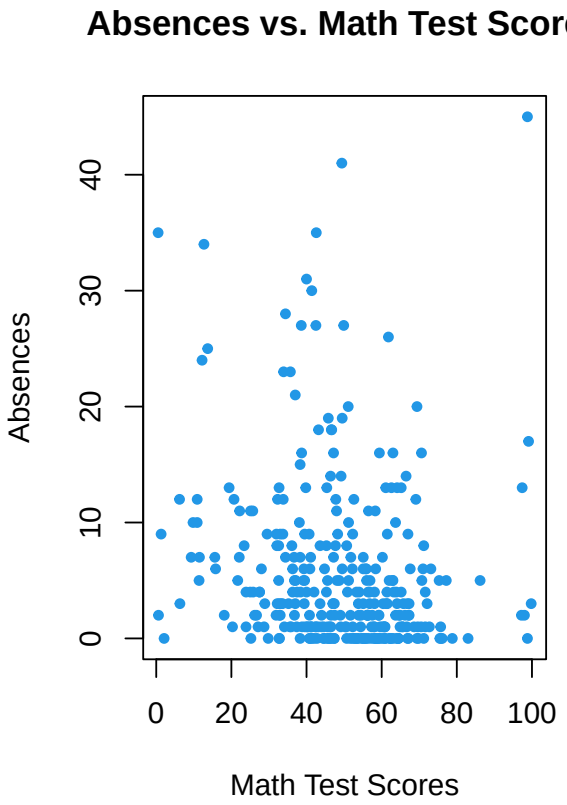
Looking across genders, we note that regardless of school, males tend to have slightly lower test scores than females, especially in language. Additionally, mean absences for males are lower than females across both schools.

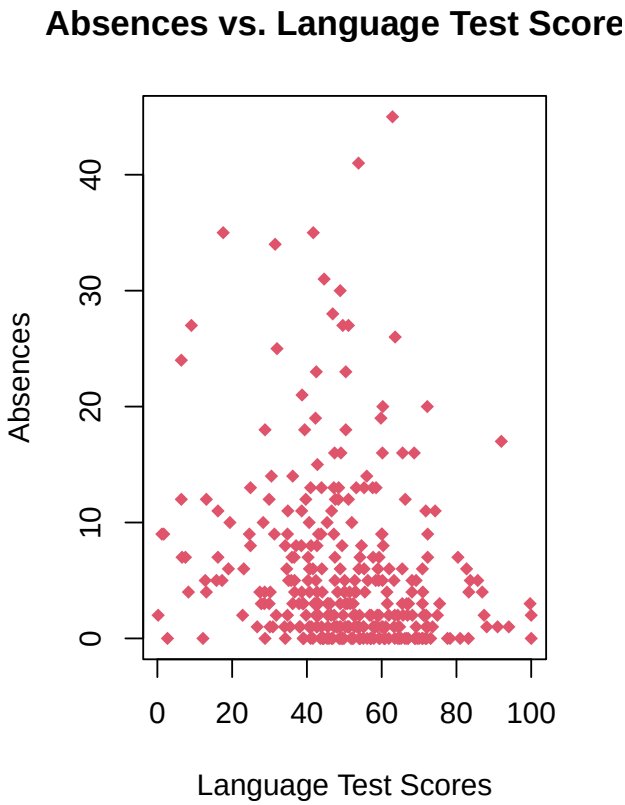
Also, we find that for each type of summary data, the variance is significantly larger than the mean. When variation is higher than would be expected, this indicates signs of over-dispersion.

Exploratory Relationships

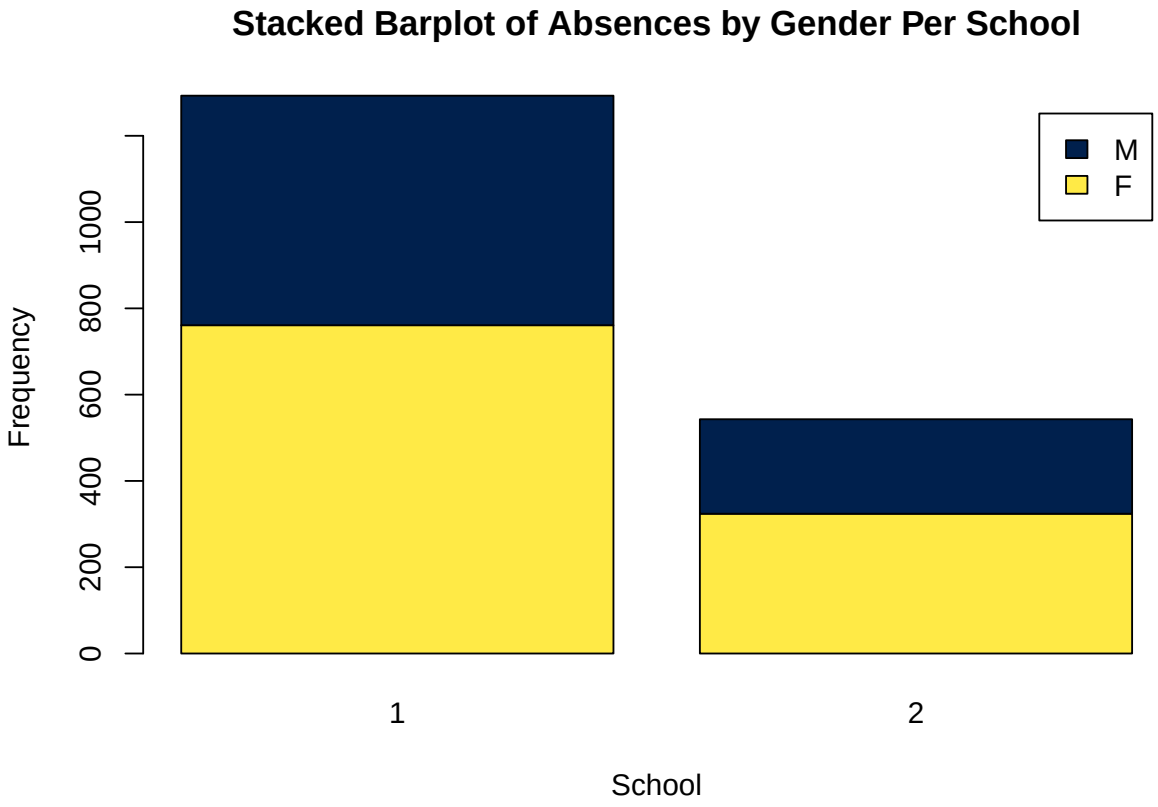
Now, we examine exploratory visualizations to confirm what we recognized from the descriptive statistics.

Below are scatterplots of absences vs. each standardized test score. We find that there are slightly negative relationships between both math and language scores and the number of days of absence. This implies students with higher test scores tend to have fewer absences.

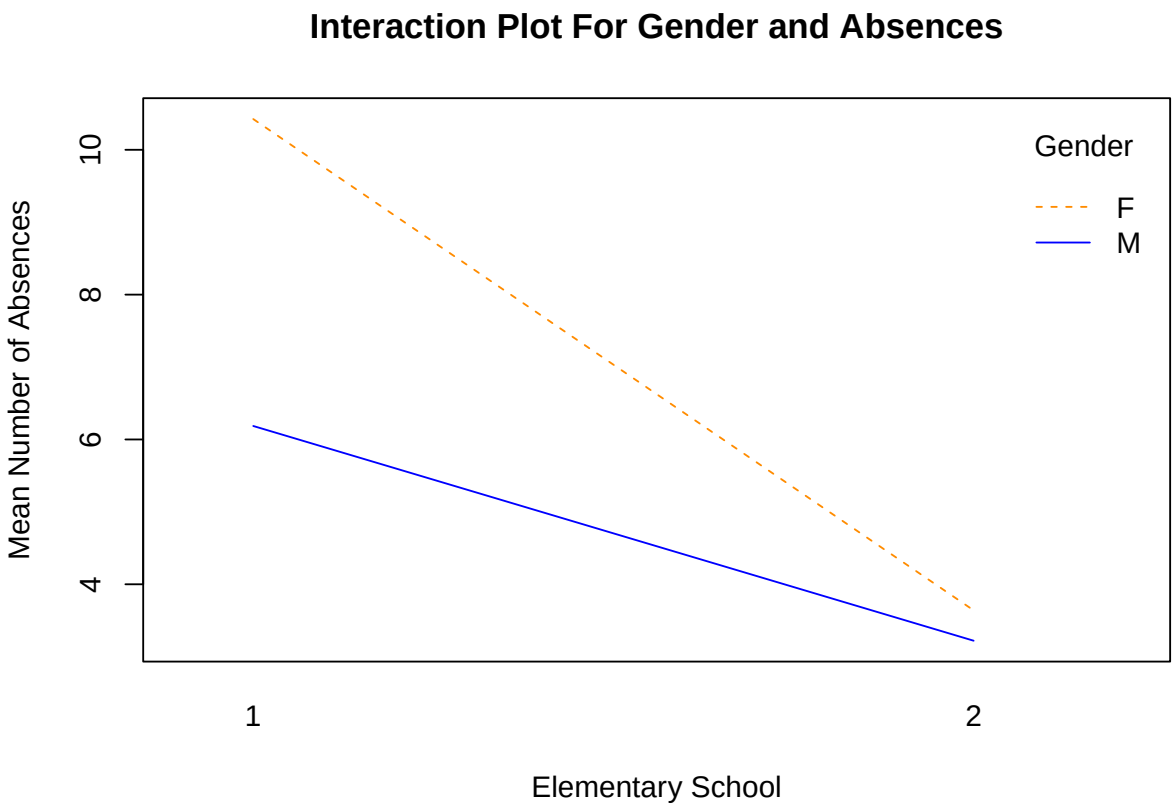




We also display a stacked barplot of gender absences by school. Here, we visually confirm that for school 1, male and female absences are much higher than in school 2.



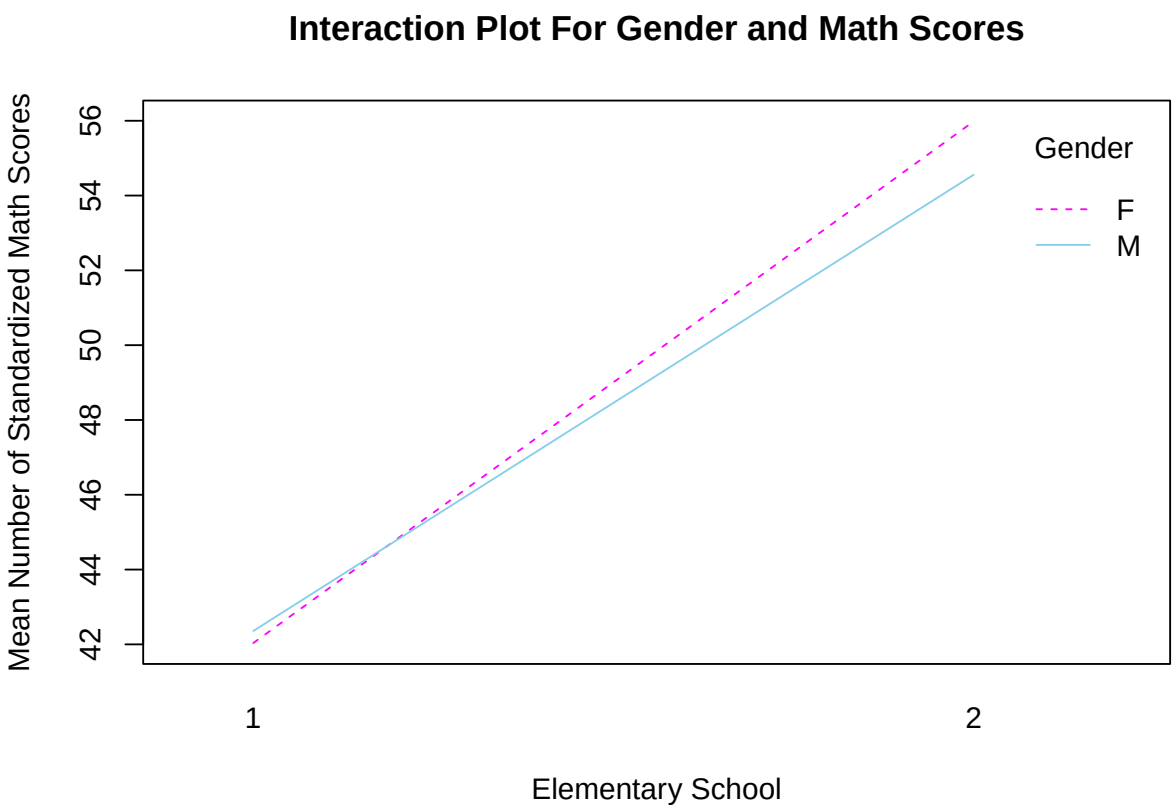
From the exploratory plots we can understand the way absences are affected by test scores and which school a student attends. However, it appears that depending on which school male and female 6th graders attend, it leads to a significant increase or decrease in their number of absences. Therefore, we will check for possible interaction between gender and absences by school using an interaction plot.



If the lines on the interaction plot are parallel, then there's no interaction between the factors. If the lines intersect, then there's likely an interaction between them.

We can see that our lines are not intersecting, but our lines are also not parallel, which means we should still investigate if an interaction between school and gender on absence is present.

Additionally, from our descriptive statistics, the mean of math scores seems to dependent on what school these 6th graders attend as well. Therefore, we will check for possible interaction between gender and math by school using an interaction plot.



We can see that our lines are intersecting, which means we should investigate if an interaction between gender and math is present.

Final Model

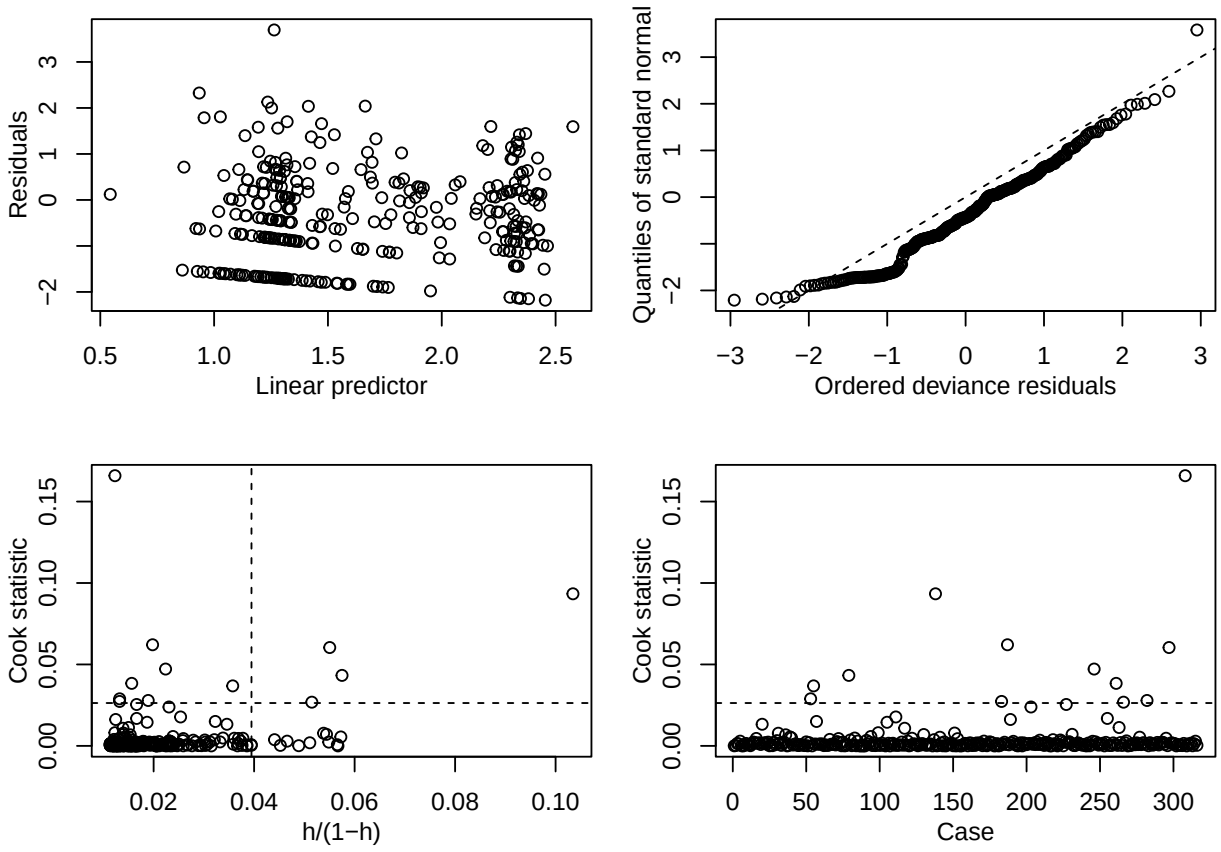
Next, we display the model summary for the final model.

Table 8: Model Summary - Negative Binomial Model

term	estimate	std.error	statistic	p.value
(Intercept)	2.162	0.2635	8.208	2.246e-16
school2	-1.106	0.1955	-5.656	1.552e-08
genderM	0.2132	0.3696	0.5769	0.564
math	0.004188	0.005451	0.7682	0.4423
school2:genderM	0.6718	0.2811	2.39	0.01683
genderM:math	-0.01846	0.007733	-2.387	0.017

Here, we find that the interaction terms are both significant at the 5% level, but the main effects are not. We still keep them in the final model, however, since their interaction is significant.

Diagnostics



From the diagnostic plots, we can confirm that this final model fits well with the data set. Residuals show decent random scatter (homoscedasticity), ordered deviance residuals decently follow a straight line, and cook's statistic values are very low.

Additionally, looking at all four different residual plots, we note that they all show random scatter (homoscedasticity). This further confirms the notion that this model fits this data set well.

Analysis

To better conceptualize the interpretation of the coefficient estimates, we can take the exponential of the estimates to get the incidence rate ratios (IRR) and its respective 95% confidence interval. We note that the 95% confidence intervals for all incidence rate ratios are statistically significant when the intervals do not include 1.

The interpretation of the incidence rate ratios for the model's coefficient estimates are as follows (including non-significant variables):

Note: The interpretation for coefficients that are not statistically significant in our final model may conflict with what was analyzed in the exploratory analysis. We will proceed with caution on accepting these particular interpretations as truth. Additionally, the inclusion of interaction terms makes the interpretation of the effect of each variable quite challenging to fully explain. Thus, we will use this interpretation more for reference than as a guide for us to completely describe attendance behavior between these two elementary schools.

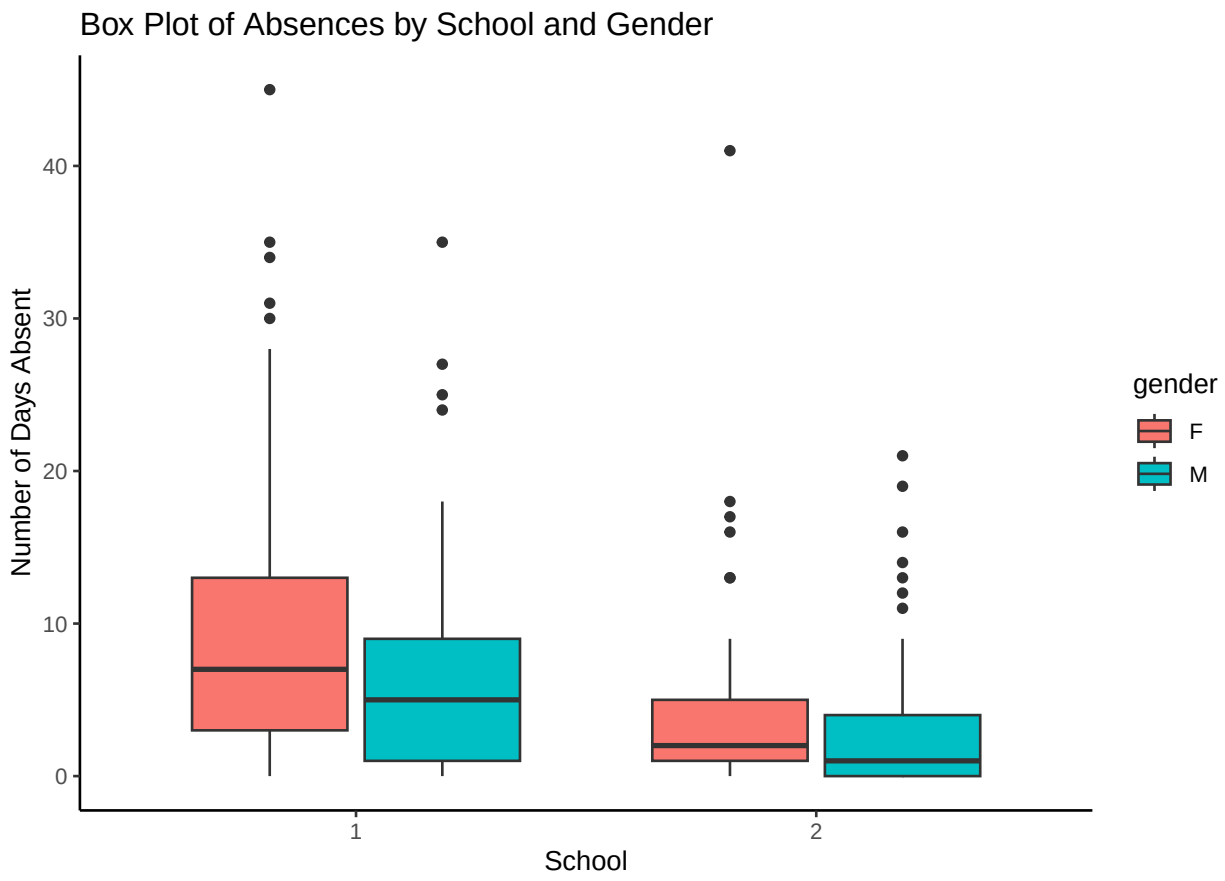
Table 9: Incidence Rate Ratios with 95% C.I.

	Estimate	Lower_Limit	Upper_Limit
(Intercept)	8.688	5.186	14.57
school2	0.3309	0.2256	0.4858
genderM	1.237	0.5999	2.555
math	1.004	0.994	1.015
school2:genderM	1.958	1.129	3.397
genderM:math	0.9822	0.9666	0.997

- **Intercept**: Represents the expected IRR for the reference group, which is the baseline category for all categorical variables (i.e., school1, female gender, and zero math score). This means that, on average, the expected count of absences for this group is 8.688 times higher than the count of absences for the group with zero predictor values.
- **school2**: Represents the expected change in the count of absences between the second school and the first school, holding all other variables constant. Specifically, the expected count of absences for the second school is lower by a factor of 0.3309 than that for the first school.
- **genderM**: Represents the expected change in the count of absences between male and female students, holding all other variables constant. Specifically, the expected count of absences for male students is higher by a factor of 1.237 than that for female students.
- **math**: Represents the expected change in the count of absences associated with a one-unit increase in the standardized math score, holding all other variables constant. Specifically, the expected count of absences increases by a factor of 1.004 for each one-unit increase in math score
- **school2:genderM**: Represents the expected difference in the IRR of absences between male and female students in the second school compared to the first school. Specifically, the expected IRR of absences for male students in the second school is 1.958 times higher than that for female students in the first school, after controlling for other variables.
- **genderM:math**: Represents the expected difference in the IRR of absences between male and female students for a one-unit increase in math score. Specifically, the IRR of absences decreases by a factor of 0.9822 more for male students than for female students, for each one-unit increase in math score.

In light of the struggles in suitably interpreting the incidence rate ratios for the interaction terms, we attempt to visually inspect absences based on multiple predictors (`school` and `gender`) to gather more information on how all these variables connect with one another.

Below, we display boxplots of absences against each gender and school. This visual shows us that females tend to have more mean absences than males across both schools, and the mean absences in school 1 is greater than the mean absences in school 2.



Overall, what we can gather is that the biggest impact of attendance behavior seems to be based on which school the 6th grader attends. Then, within each school, there appears to be an influence in the number of absences based on what gender the student is. While it is tricky to quantify each predictor's effect individually, it is important to visualize the relationships between the model's predictors and understand how they may be dependent with one another.

Conclusions

In conclusion, it is important to highlight the difficulties in fully comprehending the interpretation of the effects each variable has in explaining attendance behavior for these 6th graders in the district. However, without a doubt, we can confirm that the first elementary school consists of a much higher absence rate than the second elementary school, regardless of gender.

Therefore, if we were to extend this research out further, one possibility would be to have the district devote more attention and offer more support to the first elementary school to see if it helps improve student attendance rates, particularly among females, who showed the highest amount of absences in the data set. Another option would be to collect more information about these 6th graders, such as their parents income. Perhaps there are other factors outside of school that we were not able to consider in this analysis that could have a major effect in attendance behavior for these students.