PSTAT 220B Project Scratch Work

Elijah Castro

March 24, 2023

Contents

In a good data analysis, one usually needs to go through several exploratory "iterations" before reaching at the final results. In the process, tentative models should be evaluated and reevaluated by both statistical analytical tools and by common sense. In the presentation of the final results, however, one should avoid tedious reporting, but instead focus on the important findings. When appropriate, do use plots in your exploration, and do include good ones in your presentation.

1. Nursing Staff Glove Use

An experiment was conducted to investigate the effect of an educational program on compliance with glove use in a cardiology department of a hospital. The participants in the study were all 23 members of the department's nursing staff. They all attended an educational program addressing the importance of wearing gloves. Without their knowledge, the nurses were observed during heart valve surgeries before and one, two, and five months after an educational program to see how often they wore gloves. Each procedure by a nurse was counted as a separate observation.

The dataset `glove.dat` contains the following variables:

Period Observation period (1 = before intervention, 2 = one month after intervention, 3 = two months after, 4 = 5 months after intervention)

Observed Number of times the nurse was observed

Gloves Number of times the nurse used gloves

Experience Years of experience of nurse

Investigate 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.

To investigate whether the educational program on the importance of using gloves improved glove use in heart valve surgeries and whether it depended on the years of experience, we first load and examine the data set.

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

EDA / Descriptive Statistics

Now, we begin with some descriptive statistics to get an overview of the data:

Period	Observed	Gloves	Experience
1:16	Min. : 1.0	Min. : 0.0	Min. : 1.0
2:19	1st Qu.: 1.0	1st Qu.: 1.0	1st Qu.: 3.0
3:15	Median : 2.0	Median : 1.0	Median : 8.0
4:13	Mean : 3.6	Mean : 2.6	Mean : 8.7
NA	3rd Qu.: 4.0	3rd Qu.: 3.5	3rd Qu.:14.0
NA	Max. :15.0	Max. :14.0	Max. :20.0

Period	mean_gloves	sd_gloves	var_gloves
1	0.8125	1.515	2.296
2	3.368	3.004	9.023
3	3.333	2.944	8.667
4	3	4.183	17.5

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

Additionally, we find that for each observation period, the variance in gloves used by nurses is significantly larger than the mean. When variation is higher than would be expected, this indicates signs of over-dispersion. We will keep this information in mind when fitting our model of interest for this data set.

Table 4: Count of Nurses Based on Gloves Used by Period

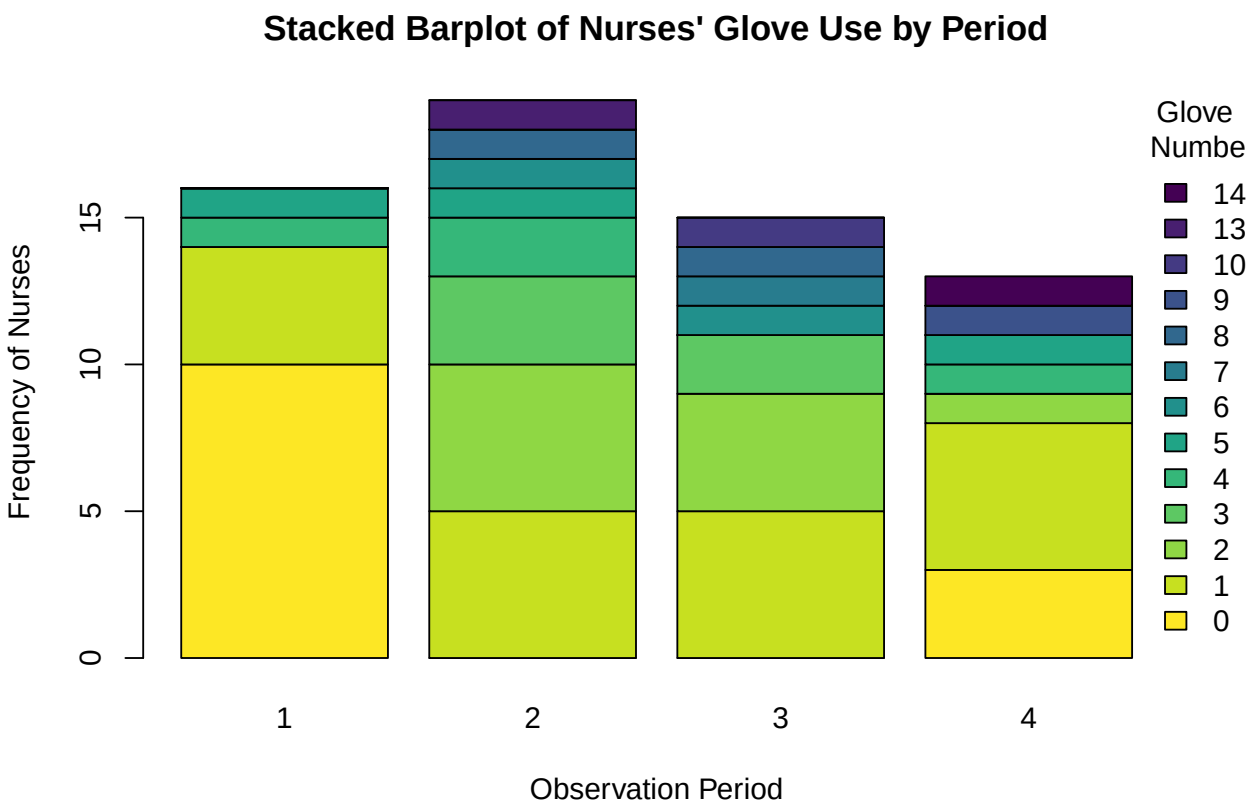
	1	2	3	4
0	10	0	0	3
1	4	5	5	5
2	0	5	4	1
3	0	3	2	0
4	1	2	0	1
5	1	1	0	1
6	0	1	1	0
7	0	0	1	0
8	0	1	1	0
9	0	0	0	1
10	0	0	1	0
13	0	1	0	0
14	0	0	0	1

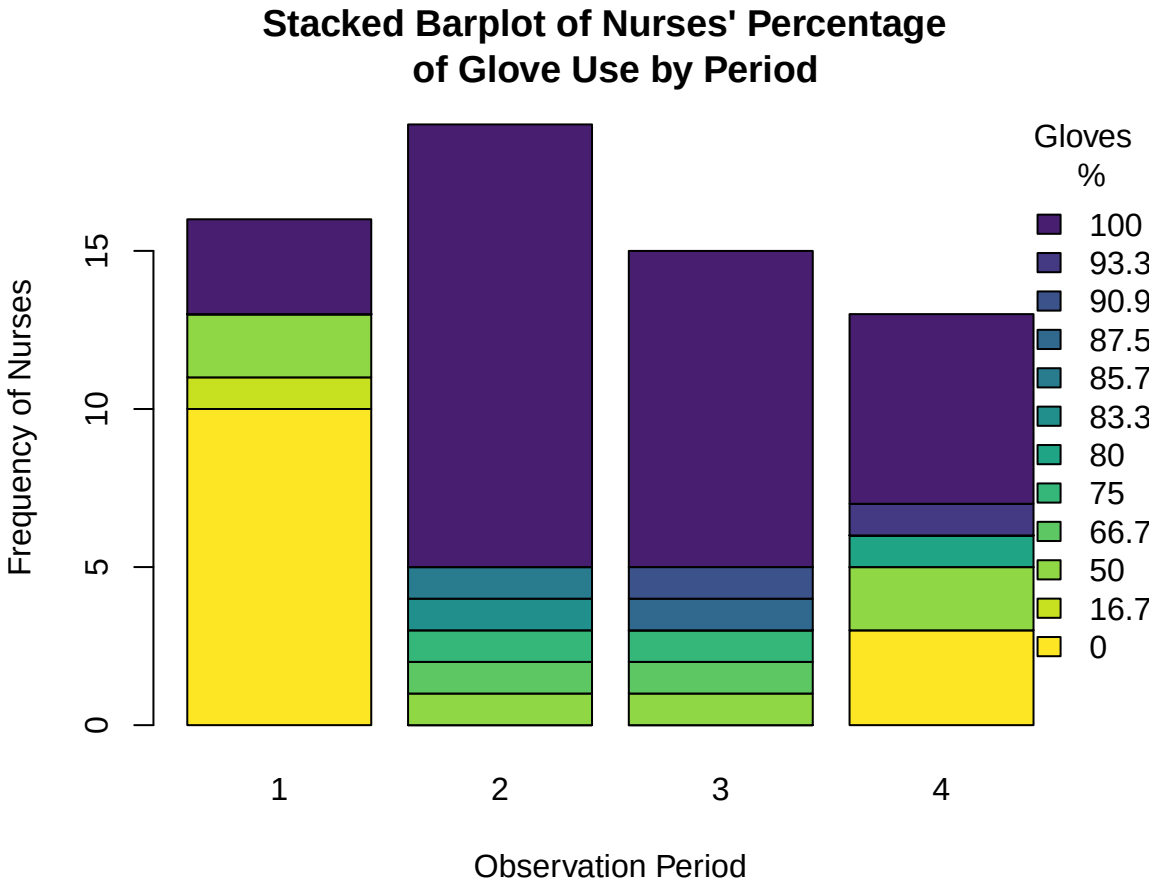
Table 5: Count of Nurses Based on Percentage of Gloves Used by Period

	1	2	3	4
0	10	0	0	3
16.7	1	0	0	0
50	2	1	1	2
66.7	0	1	1	0
75	0	1	1	0
80	0	0	0	1
83.3	0	1	0	0
85.7	0	1	0	0
87.5	0	0	1	0
90.9	0	0	1	0
93.3	0	0	0	1
100	3	14	10	6

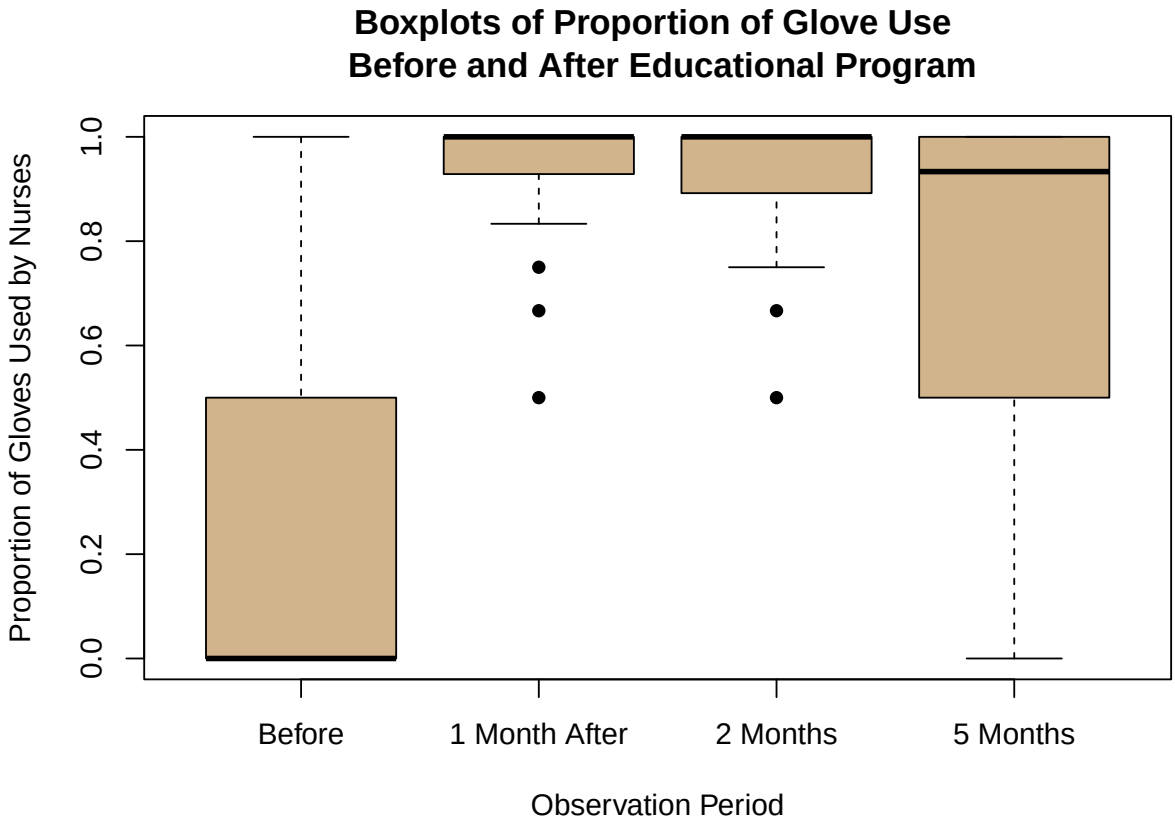
With these two tables, we can find out the count of nurses based on how many gloves they used during that particular observation period, or the count of nurses based on the percentage of gloves used per observation period. Using the percentage of gloves will be a better indicator of telling us the behavior of the nurses regarding consistent glove use. Therefore, we will focus on creating a stacked barplot of nurses based on percentage of glove use per observation period.

In the visual below, we find that before the educational program, a huge majority of nurses never wore gloves in heart valve surgeries. However, in periods 2 and 3, we see that shortly 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 wear gloves 0% of the time since period 1.

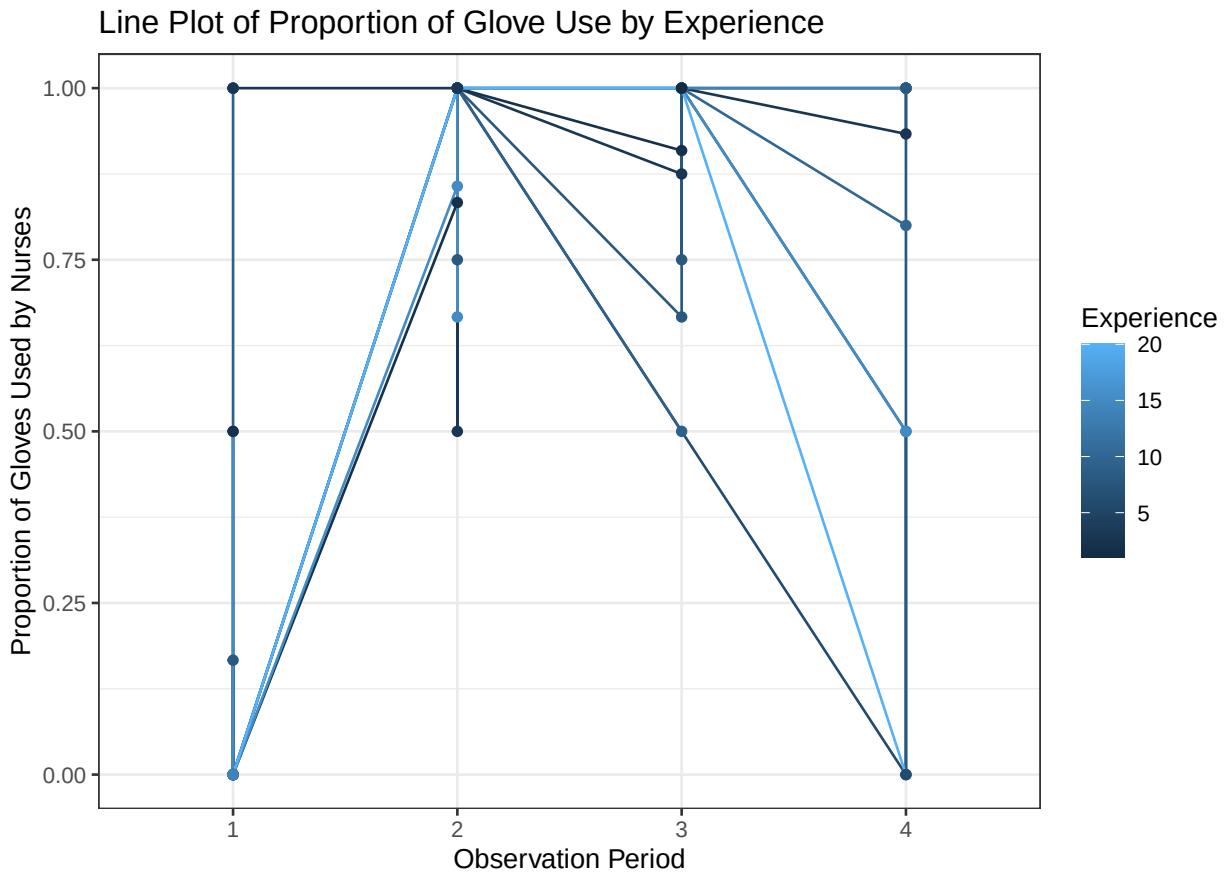

Stacked Barplot of Nurses' Glove Use by Period
Frequency of Nurses
0
5
10
15
1
2
3
4
Observation Period
Glove
Numbe
14
13
10
9
8
7
6
5
4
3
2
1
0


Stacked Barplot of Nurses' Percentage
of Glove Use by Period
Gloves
%
100
93.3
90.9
87.5
85.7
83.3
80
75
66.7
50
16.7
0
Frequency of Nurses
0
5
10
15
1
2
3
4
Observation Period

Now, we will display boxplots that will show the proportion of gloves worn during each observation period. Similar to the stacked barplot, this visual tells 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 jumpe to almost 1. By 5 months after the program was presented to the nurses, we note that the mean dips slightly, and the size of the boxplot has increased dramatically. This means 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 indicates more nurses are using less gloves in heart valve surgeries.



Our last plot will visualize the effect of Period on glove use by Experience using a line plot. This will give us a visual representation of whether the effect of the intervention differs based on years of experience:



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.

Model-Building

Binomial Regression Model To investigate whether the educational program on the importance of using gloves improved glove use in heart valve surgeries and whether it depended on the years of experience, we first fit a binomial regression model. In this model, we column bind to specify the number of successes (`Gloves`) and failures (`Observed - Gloves`) for each observation as the response, and we include `Period` and `Experience` as the predictors. We also take a look at the output of the model summary.

```
model.b <- glm(cbind(Gloves, Observed - Gloves) ~ Period + Experience,
               family = "binomial", data = nurses_gloves)

summary(model.b)
```

```
##
## Call:
## glm(formula = cbind(Gloves, Observed - Gloves) ~ Period + Experience,
##     family = "binomial", data = nurses_gloves)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.375  -1.015   0.349   0.822   3.248
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2734     0.4517    0.61     0.55
## Period2       3.6334     0.5927    6.13  8.8e-10 ***
## Period3       2.9595     0.5661    5.23  1.7e-07 ***
## Period4       1.9963     0.4982    4.01  6.1e-05 ***
## Experience   -0.1424     0.0359   -3.96  7.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 179.401  on 62  degrees of freedom
## Residual deviance:  85.162  on 58  degrees of freedom
## AIC: 124.3
##
## Number of Fisher Scoring iterations: 5
```

We also use `Anova()` to check the significance of each term individually in the model:

```
Anova(model.b, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cbind(Gloves, Observed - Gloves)
##            LR Chisq Df Pr(>Chisq)
## Period         63.6  3    1.0e-13 ***
```

```
## Experience     17.3  1    3.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To investigate whether the effect of the educational program depends on the years of experience, we update our model to include an interaction term for Period and Experience. We again use `Anova()` to investigate whether the interaction term is significant for our model:

```
model.b2 <- update(model.b, ~. + Period:Experience)

Anova(model.b2, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cbind(Gloves, Observed - Gloves)
##                   LR Chisq Df Pr(>Chisq)
## Period                4.23  3    0.23789
## Experience           12.06  1    0.00052 ***
## Period:Experience     8.54  3    0.03603 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we find that the interaction term is significant at the 5% level for our model, but it does make the original `Period` predictor insignificant. Therefore, we will compare the two nested models using Chi-squared test.

```
anova(model.b, model.b2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Gloves, Observed - Gloves) ~ Period + Experience
## Model 2: cbind(Gloves, Observed - Gloves) ~ Period + Experience + Period:Experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        58       85.2
## 2        55       76.6  3     8.54    0.036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At the 5% level, we find that the model containing the interaction term is a better fit for the data set.

Quasi-Binomial Regression Model After the inspection of our preliminary binomial regression model, we note that there are signs of possible over-dispersion present in the model. A rule of thumb is that for a Binomial GLM, the ratio between the residual deviance (85.162 in this case) and the degrees of freedom (58) should be close to 1. In this example we can see the rule is clearly violated, indicating a Binomial GLM does not model the data set well.

Instead, we can fit a quasi-binomial model. Quasi-binomial models are suitable when the response variable is binary and the variation is higher than would be expected.

In this model, we also column bind to specify the number of successes (`Gloves`) and failures (`Observed - Gloves`) for each observation. The formula includes this as the response and `Period` and `Experience` as

the predictors. Again, we will also update our model with an interaction term for `Period` and `Experience`, which allows us to investigate whether the effect of the educational program depends on the years of experience in the quasi-binomial scenario.

```
# quasi-binomial glm
model.qb <- glm(cbind(Gloves, Observed - Gloves) ~ Period + Experience,
             data = nurses_gloves, family = quasibinomial(link = "logit"))


# summary of model
summary(model.qb)
```

```
##
## Call:
## glm(formula = cbind(Gloves, Observed - Gloves) ~ Period + Experience,
##     family = quasibinomial(link = "logit"), data = nurses_gloves)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.375  -1.015   0.349   0.822   3.248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2734     0.5511    0.50   0.6217
## Period2       3.6334     0.7230    5.03  5.1e-06 ***
## Period3       2.9595     0.6906    4.29  7.0e-05 ***
## Period4       1.9963     0.6077    3.29   0.0017 **
## Experience   -0.1424     0.0438   -3.25   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.5)
##
##     Null deviance: 179.401  on 62  degrees of freedom
## Residual deviance:  85.162  on 58  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

We also use `Anova()` to check the significance of each term individually in the model:

```
Anova(model.qb, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cbind(Gloves, Observed - Gloves)
##            LR Chisq Df Pr(>Chisq)
## Period         42.7  3    2.8e-09 ***
## Experience     11.6  1    0.00066 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To investigate whether the effect of the educational program depends on the years of experience, we update our model to include an interaction term for Period and Experience. We again use `Anova()` to investigate whether the interaction term is significant for our model:

```
model.qb2 <- update(model.qb, ~. + Period:Experience)

Anova(model.qb2, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cbind(Gloves, Observed - Gloves)
##                    LR Chisq Df Pr(>Chisq)
## Period                3.08  3      0.380
## Experience            8.78  1      0.003 **
## Period:Experience     6.22  3      0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, we actually find that the interaction term is **not** significant at the 5% level for our model, and it still makes the original `Period` predictor insignificant. Unlike in the binomial regression model, the addition of an interaction term in the quasi-binomial model is not necessary. We will confirm this notion by comparing these two nested models using Chi-squared test.

```
anova(model.qb, model.qb2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Gloves, Observed - Gloves) ~ Period + Experience
## Model 2: cbind(Gloves, Observed - Gloves) ~ Period + Experience + Period:Experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        58       85.2
## 2        55       76.6  3     8.54      0.1
```

At the 5% level, we find that the model not containing the interaction term is a better fit for the data set.

To conclude this model building portion, we emphasize the importance of finding the correct model fit for this data set because as we have shown, picking a different model can have vastly different results. In the binomial regression model, we concluded that including the interaction term was beneficial to our model fit, but in the case of the quasi-binomial model, the interaction term was not necessary. Therefore, in terms of investigating whether the effect of the educational program depends on the years of experience, we have found this to **not** be the case. As the quasi-binomial model has shown us, the effect of the educational program does not depend on the years of experience. No matter the years of experience of the nurse, all of them are affected by the educational program on the importance of glove use.

```
tidy(model.qb) %>%
  mutate("exp(estimate)" = exp(estimate)) %>%
  relocate(c(1,2,6)) %>%
  pander()
```

Summary Data for Final Model (Quasi-Binomial Without Interaction)

term	estimate	exp(estimate)	std.error	statistic	p.value
(Intercept)	0.2734	1.314	0.5511	0.4961	0.6217
Period2	3.633	37.84	0.723	5.025	5.124e-06
Period3	2.96	19.29	0.6906	4.286	0.00006957
Period4	1.996	7.362	0.6077	3.285	0.001732
Experience	-0.1424	0.8673	0.04385	-3.246	0.001944

```
se <- sqrt(diag(vcov(model.qb)))

# table of estimates with 95% CI
tab <- cbind(Estimate = round(model.qb$coefficients, 3),
             Lower_Limit = round(model.qb$coefficients - 1.96 * se, 3),
             Upper_Limit = round(model.qb$coefficients + 1.96 * se, 3))

tab %>%
  pander()
```

	Estimate	Lower_Limit	Upper_Limit
(Intercept)	0.273	-0.807	1.353
Period2	3.633	2.216	5.051
Period3	2.96	1.606	4.313
Period4	1.996	0.805	3.187
Experience	-0.142	-0.228	-0.056

```
# odds ratios instead of coefficients on the logit scale
exp(tab) %>%
  pander(caption = "Odds Ratios of Coefficient Estimates \n with Corresponding 95% Confidence Interval")
```

Table 8: Odds Ratios of Coefficient Estimates with Corresponding 95% Confidence Interval

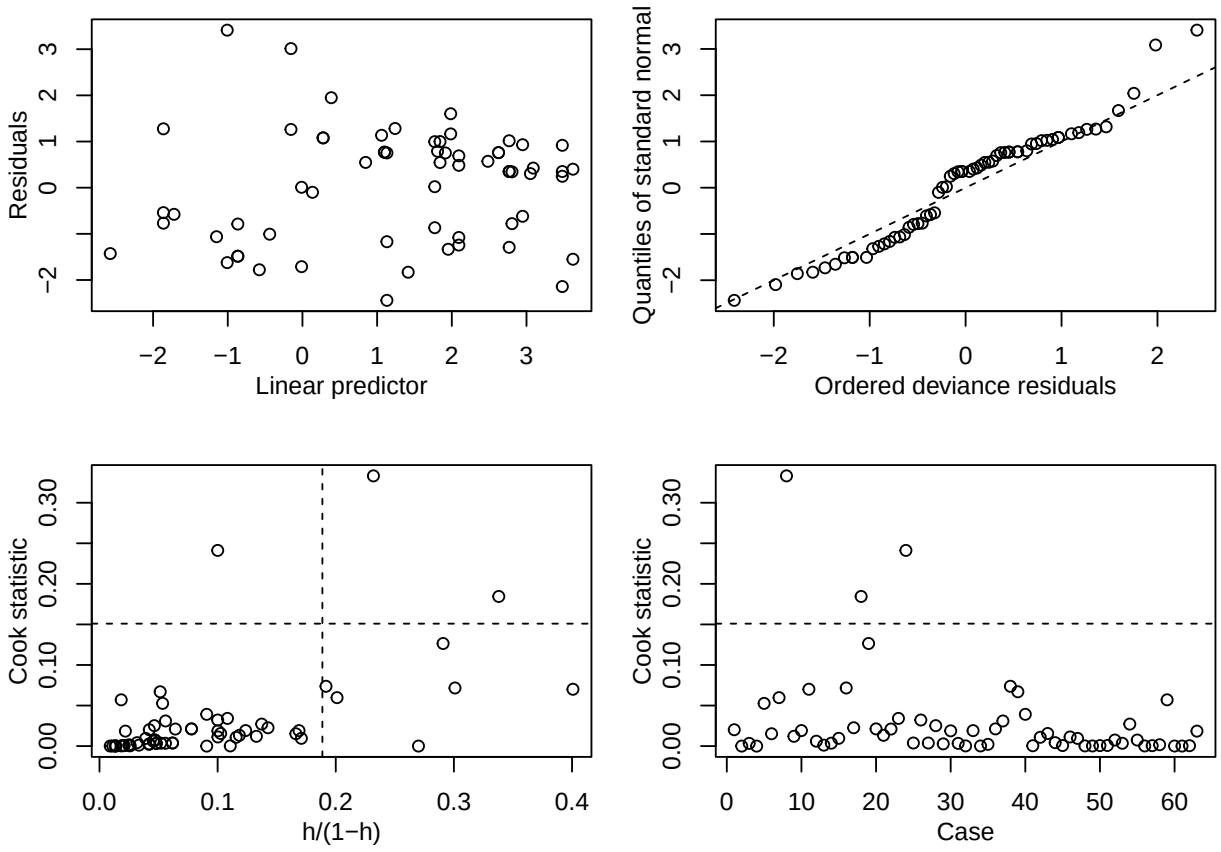
	Estimate	Lower_Limit	Upper_Limit
(Intercept)	1.314	0.4462	3.869
Period2	37.83	9.171	156.2
Period3	19.3	4.983	74.66
Period4	7.36	2.237	24.22
Experience	0.8676	0.7961	0.9455

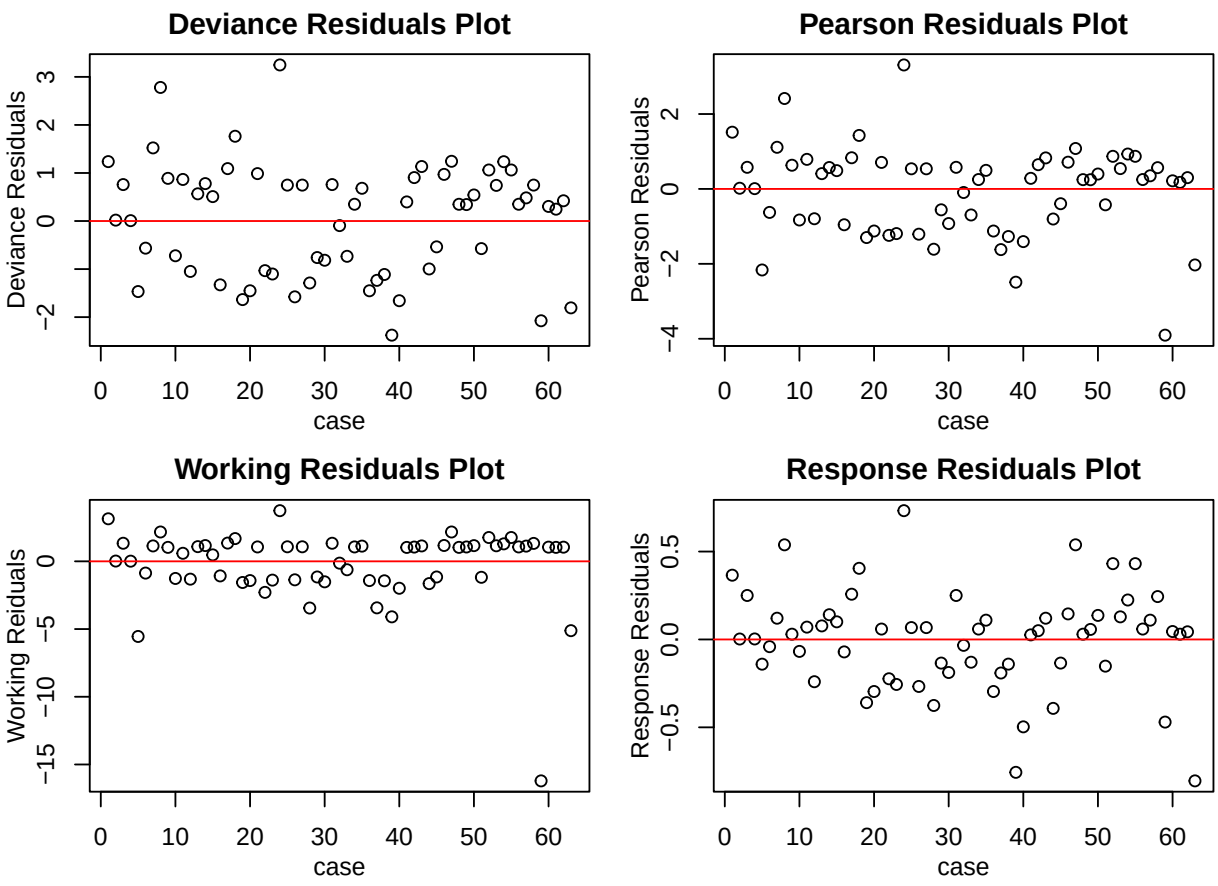
Instead of trying to interpret the coefficient estimates on the logit scale, we can take the exponential of the estimates to get the odds ratio and its respective 95% confidence interval. We note that the 95% confidence intervals for all odds ratios are statistically significant since none of the intervals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

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

- **Period 2**: The odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries are 37.83 times higher one month after the educational program on the importance of using gloves.
- **Period 3**: The odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries are 19.3 times higher two months after the educational program on the importance of using gloves.
- **Period 4**: The odds of nurses successfully using gloves when they are observed without their knowledge during heart valve surgeries are 7.36 times higher five months after the educational program on the importance of using gloves.
- **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 0.8676. In terms of the change in odds, this means that each additional increase of one year in experience is associated with about a 13.2% 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.

Diagnostic Plots




Deviance Residuals Plot
Deviance Residuals
case
Pearson Residuals Plot
Pearson Residuals
case
Working Residuals Plot
Working Reiduals
case
Response Residuals Plot
Response Residuals
case

Fitting Model Without Observation 59

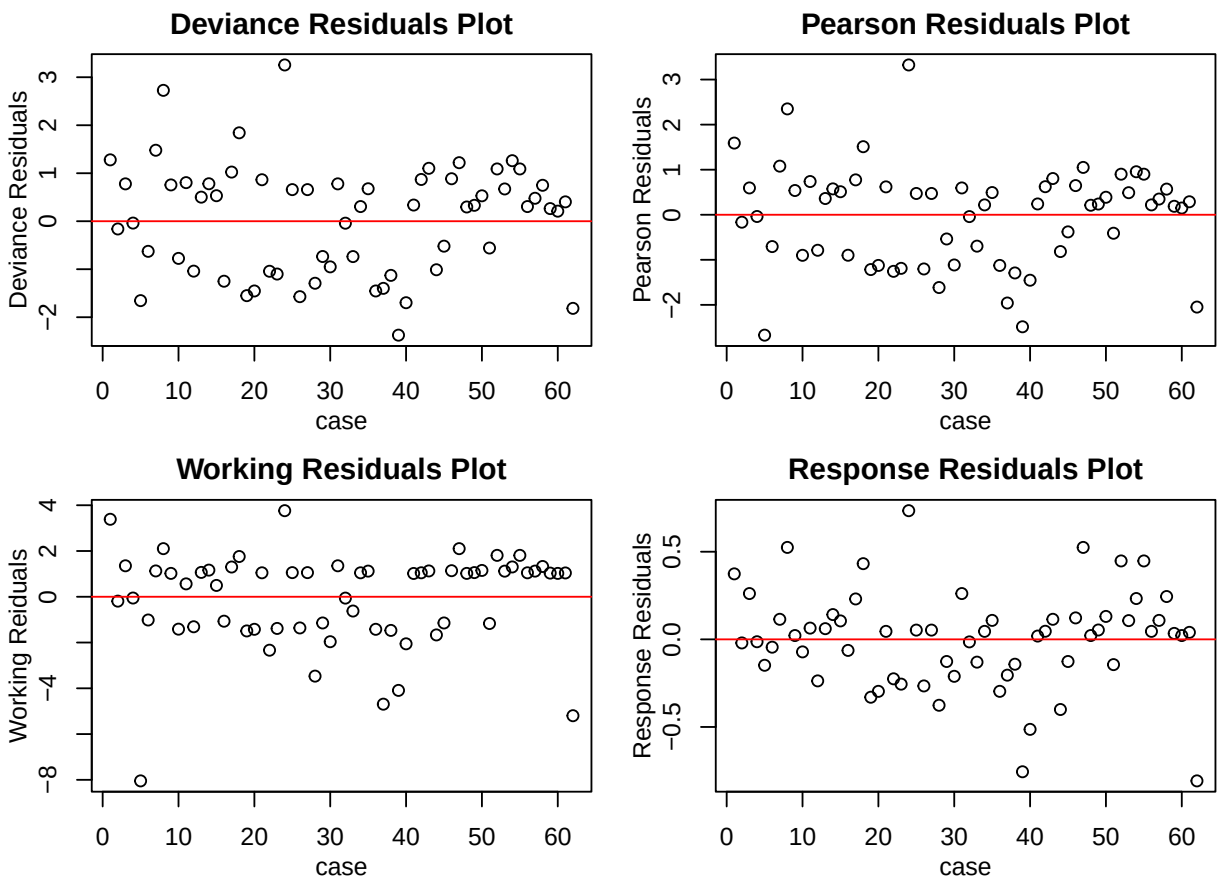
Upon inspection of our diagnostic plots, we find that observation 59 appears to be influential. Therefore, we will the same final quasi-binomial model without this point and investigate whether we should remove it from the model. We will use `Anova()` to check.

```
##
## Call:
## glm(formula = cbind(Gloves[-59], Observed[-59] - Gloves[-59]) ~
##     Period[-59] + Experience[-59], family = quasibinomial(link = "logit"),
##     data = nurses_gloves)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.374  -0.996   0.318   0.780   3.256
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.3605     0.5229    0.69  0.49335
## Period[-59]2      3.9038     0.7342    5.32  1.8e-06 ***
## Period[-59]3      2.9729     0.6544    4.54  2.9e-05 ***
## Period[-59]4      1.9918     0.5751    3.46  0.00102 **
## Experience[-59]  -0.1530     0.0426   -3.59  0.00068 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.3)
##
##     Null deviance: 178.899  on 61  degrees of freedom
## Residual deviance:  80.557  on 57  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
Anova(model.qb.final, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cbind(Gloves[-59], Observed[-59] - Gloves[-59])
##                 LR Chisq Df Pr(>Chisq)
## Period[-59]         50.8  3    5.4e-11 ***
## Experience[-59]     14.5  1    0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, diagnostic plots, specifically, residual plots look much better without this observation present.



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.

2. School Attendance Behavior

A school district is concerned with school attendance. They randomly selected 316 6th graders from two elementary schools in the district and collected their school attendance, standardized test scores and gender. The dataset `school.txt` includes the following variables:

school an indicator of two schools.

gender Male ("M") or Female ("F").

math standardized math test score.

language standardized language test score.

absence number of days of absence.

The goal is to study the attendance behavior. Specifically, investigate how the number of days of absence depends on other variables.

To investigate how the number of days of absence depends on other variables, we first load and examine the data set.

Table 9: 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

EDA / Descriptive Statistics

Now, we begin with some descriptive statistics to get an overview of the data:

school	gender	math	language	absence
1:159	F:162	Min. : 0	Min. : 0	Min. : 0
2:157	M:154	1st Qu.: 38	1st Qu.: 41	1st Qu.: 1
NA	NA	Median : 49	Median : 50	Median : 3
NA	NA	Mean : 49	Mean : 50	Mean : 6
NA	NA	3rd Qu.: 61	3rd Qu.: 61	3rd Qu.: 8
NA	NA	Max. :100	Max. :100	Max. :45

Table 11: Math Summary Data

school	gender	mean_math	sd_math	var_math
1	F	42.03	17.07	291.5
1	M	42.36	19.67	387.1
2	F	55.98	16.05	257.5
2	M	54.56	13	169.1

Table 12: Language Summary Data

school	gender	mean_language	sd_language	var_language
1	F	44.38	15.25	232.6
1	M	41.88	17.99	323.5
2	F	60.16	15.66	245.2
2	M	53.6	16.44	270.4

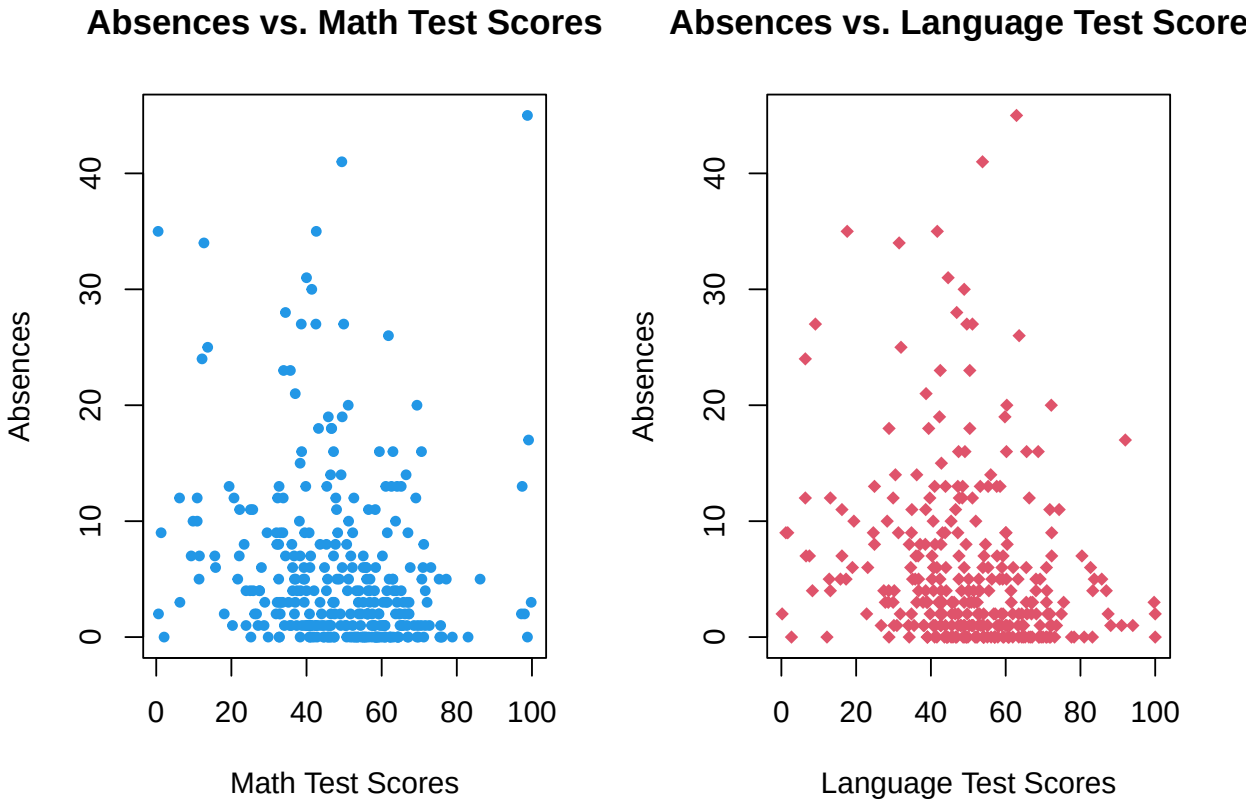
Table 13: Absences Summary Data

school	gender	mean_absence	sd_absence	var_absence
1	F	10.42	9.971	99.41
1	M	6.186	6.539	42.76
2	F	3.64	5.552	30.82
2	M	3.221	4.781	22.86

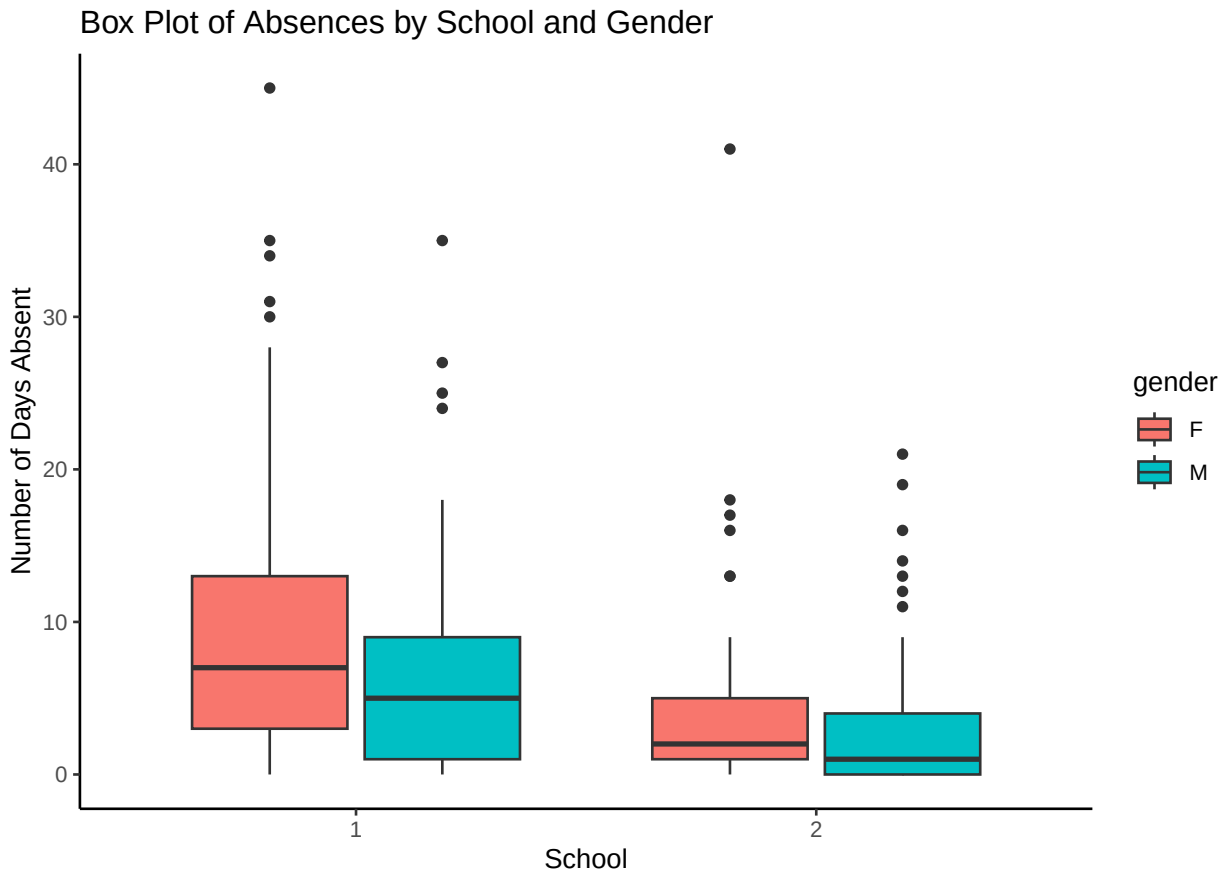
From the summary statistics, we can see that for school 1, mean scores for math and language are much lower for both genders than school 2. Additionally, mean absences for both genders are higher in school 1 than in school 2. Looking across genders, we note that regardless of school, males tend to have 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. We will keep this information in mind when fitting our model of interest for this data set.

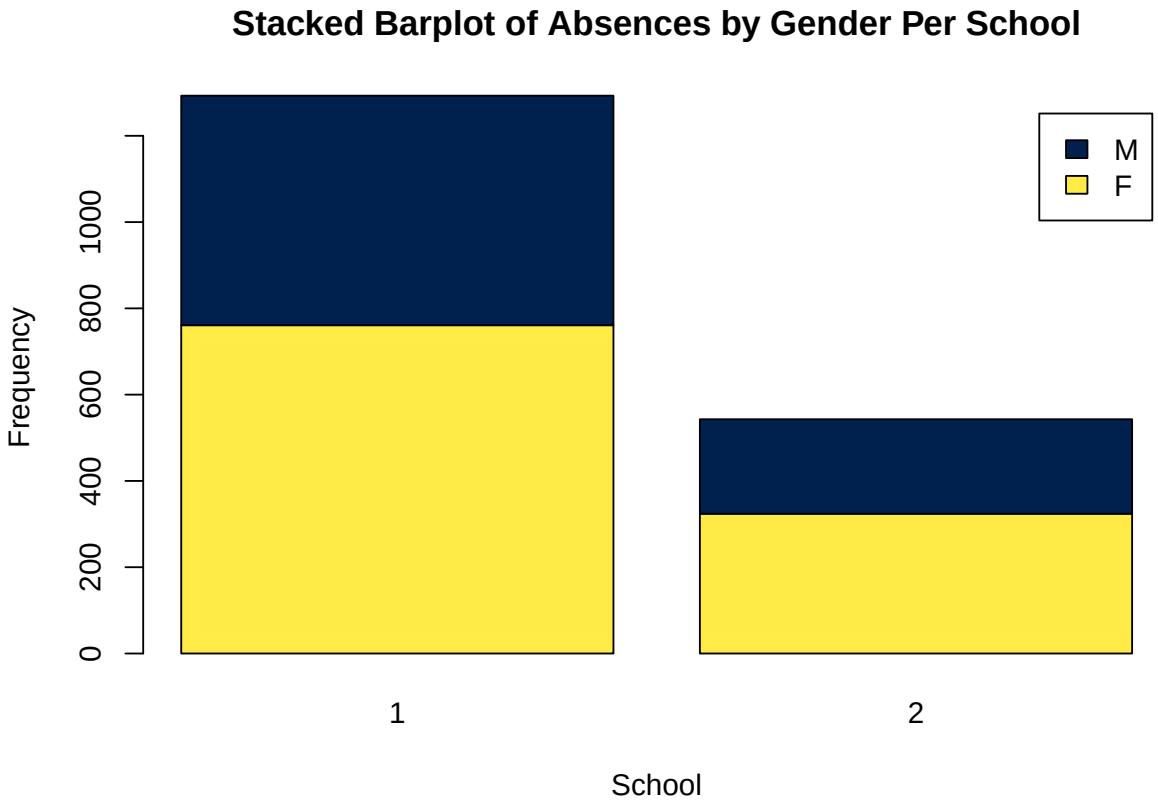
Below are scatterplots of absences vs. each standardized test score. We find that there are slightly negative relationships between both math and language scores and the number of days of absence. Students with higher test scores tend to have fewer absences.



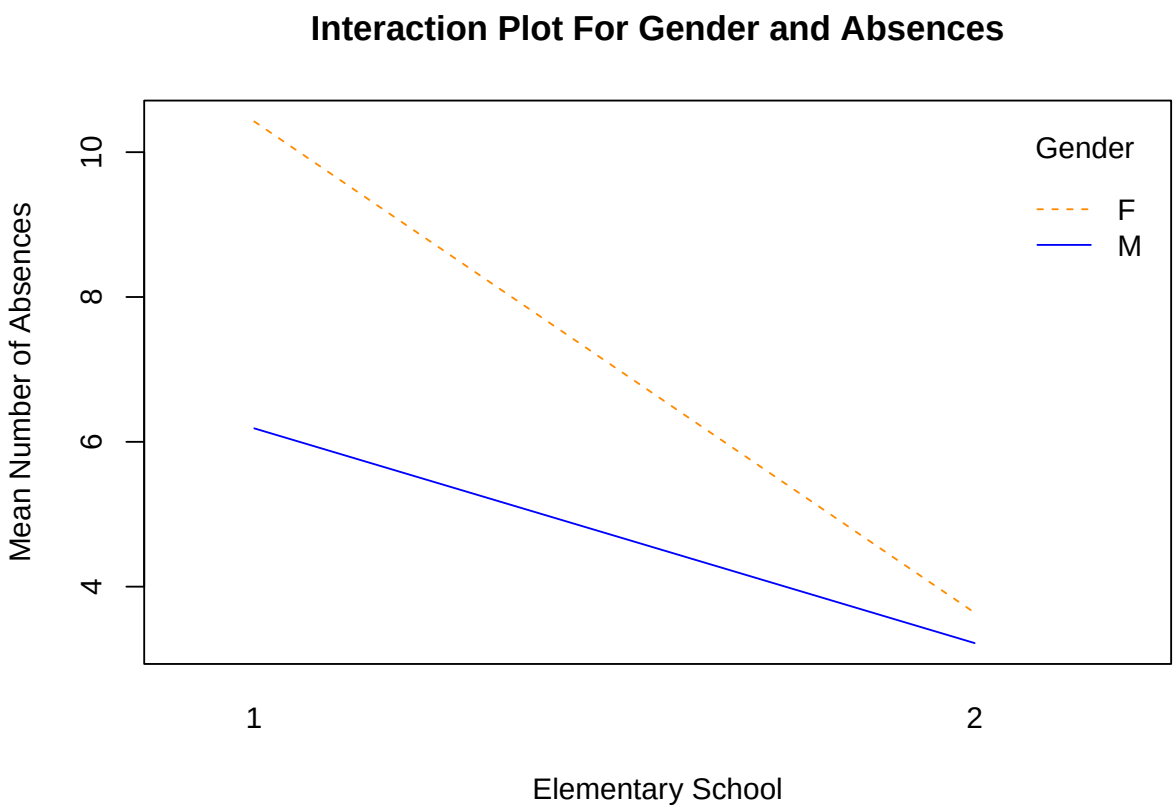
Below, we display boxplots of absences against each gender and school. This visual shows us that females tend to have more mean absences than males across both schools, and the mean absences in school 1 is greater than the mean absences in school 2.



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



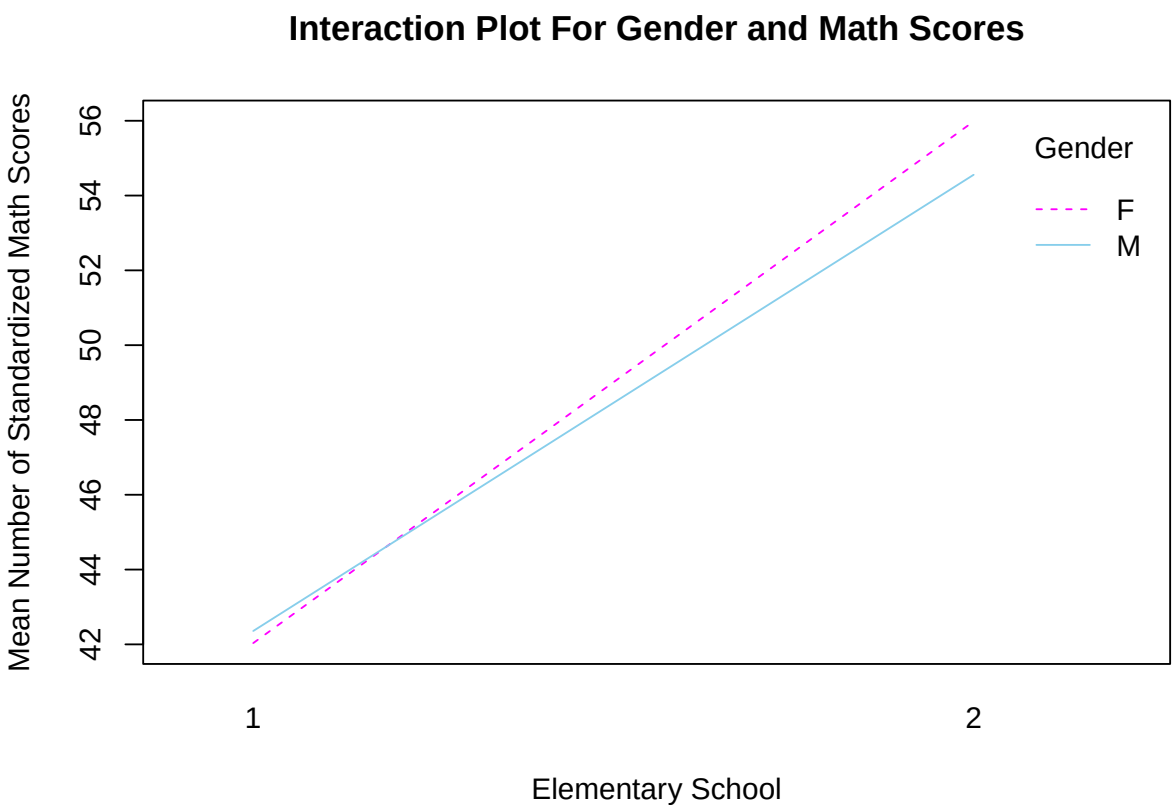
It appears that depending on which school male and female 6th graders attend, it leads to an increase or decrease in their number of absences. Therefore, we will check for possible interaction between gender and absences by school using an interaction plot.



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.



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 intersecting, which means we should investigate if an interaction between gender and math is present.

Model-Building

Poisson Regression Model To investigate how the number of days of absence depends on other variables, we first fit a poisson regression model. In this model, we include `absence` as the response, and we include `school`, `gender`, `math`, and `language` as the predictors. We also take a look at the output of the model summary.

```
model.p <- glm(absence ~ school + gender + math + language,
               family = "poisson", data = school_absences)


summary(model.p)
```

```
##
## Call:
## glm(formula = absence ~ school + gender + math + language, family = "poisson",
##     data = school_absences)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.520  -2.297  -1.121   0.811  10.674
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.577871   0.073171   35.23  < 2e-16 ***
## school2     -0.806338   0.055622  -14.50  < 2e-16 ***
## genderM     -0.435447   0.048188   -9.04  < 2e-16 ***
## math         0.000906   0.001801    0.50  0.61489
## language    -0.007261   0.001901   -3.82  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2409.8  on 315  degrees of freedom
## Residual deviance: 2007.4  on 311  degrees of freedom
## AIC: 2879
##
## Number of Fisher Scoring iterations: 6
```

We also use `Anova()` to check the significance of each term individually in the model:

```
Anova(model.p, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##          LR Chisq Df Pr(>Chisq)
## school      223.0  1    < 2e-16 ***
```

```
## gender       83.1  1    < 2e-16 ***
## math          0.3  1    0.61513
## language     14.5  1    0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To investigate whether the effect of the absences depends on the school and gender, we update our model to include an interaction term for `school` and `gender`. We again use `Anova()` to investigate whether the interaction term is significant for our model:

```
model.p2 <- update(model.p, ~. + school:gender)

Anova(model.p2, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##               LR Chisq Df Pr(>Chisq)
## school           199.3  1    < 2e-16 ***
## gender            92.6  1    < 2e-16 ***
## math               0.2  1    0.68836
## language          13.1  1    0.00030 ***
## school:gender     12.7  1    0.00037 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we find that the interaction term is significant at the 5% level for our model. Therefore, we will compare the two nested models using Chi-squared test.

```
anova(model.p, model.p2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: absence ~ school + gender + math + language
## Model 2: absence ~ school + gender + math + language + school:gender
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       311       2007
## 2       310       1995  1     12.7  0.00037 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At the 5% level, we find that the model containing the interaction term is a better fit for the data set.

Now, we try to add an interaction term for `gender` and `math` predictor since `math` on its own is highly insignificant, and print the model summary:

```
model.p3 <- update(model.p2, ~. + gender:math)

summary(model.p3)
```

```
##
## Call:
```

```
## glm(formula = absence ~ school + gender + math + language + school:gender +
##     gender:math, family = "poisson", data = school_absences)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.131  -2.375  -0.938   0.794  11.237
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.19892    0.09405   23.38  < 2e-16 ***
## school2         -1.09183    0.07229  -15.10  < 2e-16 ***
## genderM          0.31342    0.12148    2.58   0.0099 **
## math             0.00918    0.00203    4.52  6.2e-06 ***
## language        -0.00559    0.00190   -2.95   0.0032 **
## school2:genderM  0.67687    0.11189    6.05  1.5e-09 ***
## genderM:math    -0.02113    0.00269   -7.86  4.0e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2409.8  on 315  degrees of freedom
## Residual deviance: 1933.6  on 309  degrees of freedom
## AIC: 2809
##
## Number of Fisher Scoring iterations: 6
```

Including interaction terms, all the variables are now highly significant in our model.

We also use `Anova()` to check the significance of each term individually in the model, and compare the two nested models using Chi-squared test

```
Anova(model.p3, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##               LR Chisq Df Pr(>Chisq)
## school           245.9  1    < 2e-16 ***
## gender             6.6  1     0.0102 *
## math              19.8  1    8.4e-06 ***
## language           8.7  1     0.0032 **
## school:gender     36.0  1    1.9e-09 ***
## gender:math       61.2  1    5.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model.p2, model.p3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: absence ~ school + gender + math + language + school:gender
## Model 2: absence ~ school + gender + math + language + school:gender +
```

```
##     gender:math
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       310       1995
## 2       309       1934  1     61.2  5.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At the 5% level, we find that the model containing two interaction terms is a better fit for the data set.

Negative Binomial Regression Model

After the inspection of our preliminary poisson regression model, we note that there are signs of over-dispersion present in the model. A rule of thumb is that for a Poisson GLM, the ratio between the residual deviance (2007.6 in this case) and the degrees of freedom (312) should be close to 1. In this example we can see the rule is clearly violated, indicating a Poisson GLM does not model the data set well.

Instead, we can fit a negative binomial model. Negative binomial models are suitable to test for connections between confounding and predictor variables on a count response variable (in this case `absence`).

In this model, we include `absence` as the response, and we include `school`, `gender`, and `language` as the predictors. Again, we will also update our model with interaction terms for `school` and `gender` and `gender` and `math` since we discovered the effect of the absences depends on the school and gender and the school males and females attend affect their math scores.

```
# negative binomial model
model.nb <- glm.nb(absence ~ school + gender + language + math,
                   data = school_absences)


# summary of model
summary(model.nb)
```

```
##
## Call:
## glm.nb(formula = absence ~ school + gender + language + math,
##     data = school_absences, init.theta = 0.8705311227, link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -2.10   -1.06   -0.44    0.28    3.20
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.62207    0.22336   11.74  < 2e-16 ***
## school2     -0.76526    0.14461   -5.29  1.2e-07 ***
## genderM     -0.39802    0.13390   -2.97    0.003 **
## language    -0.01005    0.00524   -1.92    0.055 .
## math         0.00211    0.00509    0.41    0.679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for Negative Binomial(0.87) family taken to be 1)
##
##     Null deviance: 410.54  on 315  degrees of freedom
## Residual deviance: 357.38  on 311  degrees of freedom
## AIC: 1746
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.8705
##           Std. Err.:  0.0863
##
##  2 x log-likelihood:  -1733.7890
```

We also use `Anova()` to check the significance of each term individually in the model:

```
Anova(model.nb, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##          LR Chisq Df Pr(>Chisq)
## school      28.77  1    8.2e-08 ***
## gender       8.74  1     0.0031 **
## language     3.55  1     0.0596 .
## math         0.19  1     0.6603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To again investigate whether the effect of the absences depends on the school and gender, we update our model to include an interaction term for `school` and `gender`. We again use `Anova()` to double check whether the interaction term is significant for our model:

```
model.nb2 <- update(model.nb, ~. + school:gender)

Anova(model.nb2, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##               LR Chisq Df Pr(>Chisq)
## school           25.15  1    5.3e-07 ***
## gender           10.13  1     0.0015 **
## language          3.20  1     0.0735 .
## math              0.10  1     0.7466
## school:gender     2.14  1     0.1432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we find that the interaction term is not significant at the 5% level for our model. Now, we will compare the two nested models using likelihood ratio tests.

```
anova(model.nb, model.nb2)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: absence
##                                                 Model theta Resid. df
## 1                 school + gender + language + math  0.87       311
## 2 school + gender + language + math + school:gender  0.88       310
##      2 x log-lik.   Test    df LR stat. Pr(Chi)
## 1           -1734
## 2           -1732 1 vs 2     1      2.1    0.14
```

According to the results, we should prefer the model without the interaction term. This contrasts when we were fitting a simple Poisson regression model. Perhaps the addition of the interaction term for `gender` and `math` will affect the significance of our other predictors. We again update our model and use `Anova()` to double check whether the interaction term and other terms change their significance for our model:

```
model.nb3 <- update(model.nb2, ~. + gender:math)


# summary of model
summary(model.nb3)
```

```
##
## Call:
## glm.nb(formula = absence ~ school + gender + language + math +
##     school:gender + gender:math, data = school_absences, init.theta = 0.9007544654,
##     link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.219  -1.069  -0.404   0.265   3.557
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.33415    0.28306    8.25  < 2e-16 ***
## school2         -1.05684    0.19895   -5.31  1.1e-07 ***
## genderM          0.14862    0.37003    0.40    0.688
## language        -0.00807    0.00519   -1.55    0.120
## math             0.00867    0.00616    1.41    0.159
## school2:genderM  0.63087    0.28121    2.24    0.025 *
## genderM:math    -0.01736    0.00772   -2.25    0.025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9) family taken to be 1)
##
##     Null deviance: 420.50  on 315  degrees of freedom
## Residual deviance: 357.86  on 309  degrees of freedom
## AIC: 1742
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  0.9008
##           Std. Err.:  0.0904
##
##  2 x log-likelihood:  -1726.0080
```

```
Anova(model.nb3, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: absence
##               LR Chisq Df Pr(>Chisq)
## school           30.06  1    4.2e-08 ***
## gender            0.18  1      0.675
## language          2.32  1      0.128
## math              2.31  1      0.128
## school:gender     5.08  1      0.024 *
## gender:math       5.71  1      0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we find that the interaction terms are now both significant at the 5% level for our model. However, the main effects are not. Now, we will compare the two nested models using likelihood ratio tests.

```
anova(model.nb2, model.nb3)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: absence
##                                                                 Model theta
## 1               school + gender + language + math + school:gender  0.88
## 2 school + gender + language + math + school:gender + gender:math  0.90
##   Resid. df    2 x log-lik.   Test    df LR stat. Pr(Chi)
## 1       310           -1732
## 2       309           -1726 1 vs 2     1      5.6   0.017
```

According to the results, we should now prefer the model with the interaction terms present.

Lastly, since `language` was insignificant at the 5% level and does not have an interaction present with any other predictor in the model, we check if we should remove it from the model by comparing the two nested models using likelihood ratio tests.

```
model.nb.final <- update(model.nb3, ~. - language)


anova(model.nb3, model.nb.final)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: absence
##                                                             Model theta
```

```
## 1            school + gender + math + school:gender + gender:math  0.89
## 2 school + gender + language + math + school:gender + gender:math  0.90
##   Resid. df    2 x log-lik.   Test    df LR stat. Pr(Chi)
## 1       310           -1728
## 2       309           -1726 1 vs 2     1      2.3    0.13
```

According to the results, the model which includes `language` is not significantly different than the model without it. Therefore, we choose the simpler model without `language`, and we have arrived at our final model.

```
# final model summary
summary(model.nb.final)
```

```
##
## Call:
## glm.nb(formula = absence ~ school + gender + math + school:gender +
##     gender:math, data = school_absences, init.theta = 0.8930857462,
##     link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.172  -1.065  -0.431   0.284   3.560
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.16249    0.26346    8.21  2.2e-16 ***
## school2         -1.10557    0.19548   -5.66  1.6e-08 ***
## genderM          0.21325    0.36962    0.58    0.564
## math             0.00419    0.00545    0.77    0.442
## school2:genderM  0.67180    0.28105    2.39    0.017 *
## genderM:math    -0.01846    0.00773   -2.39    0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.89) family taken to be 1)
##
##     Null deviance: 417.99  on 315  degrees of freedom
## Residual deviance: 358.11  on 310  degrees of freedom
## AIC: 1742
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.8931
##           Std. Err.:  0.0894
##
##  2 x log-likelihood:  -1728.3220
```

```
tidy(model.nb.final) %>% pander()
```

Summary Data for Final Model (Negative Binomial)

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

```
se <- sqrt(diag(vcov(model.nb.final)))

# table of estimates with 95% CI
tab <- cbind(Estimate = round(model.nb.final$coefficients, 3),
             Lower_Limit = round(model.nb.final$coefficients - 1.96 * se, 3),
             Upper_Limit = round(model.nb.final$coefficients + 1.96 * se, 3))

tab %>%
  pander()
```

	Estimate	Lower_Limit	Upper_Limit
(Intercept)	2.162	1.646	2.679
school2	-1.106	-1.489	-0.722
genderM	0.213	-0.511	0.938
math	0.004	-0.006	0.015
school2:genderM	0.672	0.121	1.223
genderM:math	-0.018	-0.034	-0.003

Above are the coefficient estimates for the final model and its respective 95% confidence interval. We note that the 95% confidence intervals that are statistically significant do not include 0.

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

Note: We will also include interpretations for coefficients that are not statistically significant in our final model. They may conflict with what was discovered in the exploratory analysis.

- **Intercept**: Represents the expected log count of absences for the reference group, which is the baseline category for all categorical variables (i.e., school1, female gender, and zero math score).
- **school2**: Represents the difference in the expected log count of absences between the second elementary school and the first, holding all other variables constant. Specifically, the expected log count of absences for the second school is lower by 1.106 than that for the first school.
- **genderM**: Represents the difference in the expected log count of absences between male and female students, holding all other variables constant. Specifically, the expected log count of absences for male students is higher by 0.213 than that for female students.

- **math**: Represents the change in the expected log count of absences associated with a one-unit increase in the standardized math score, holding all other variables constant. Specifically, the expected log count of absences increases by 0.004 for each one-unit increase in math score.
- **school2:genderM**: Represents the difference in the effect of being in school 2 (versus school 1) on the log count of absences between males and females. Specifically, the expected log count of absences for male students in the second school is higher by 0.672 than that for female students in the first school.
- **genderM:math**: Represents the change in the effect of math score on the expected log count of absences between male and female students. Specifically, for each one-unit increase in math score, the expected log count of absences decreases by 0.018 more for male students than for female students.

Incidence Rate Ratios

To better conceptualize the interpretation of the coefficient estimates, we can take the exponential of the estimates to get the incidence rate ratios 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 information more for reference than as a guide for us to completely describe attendance behavior between these two elementary schools.

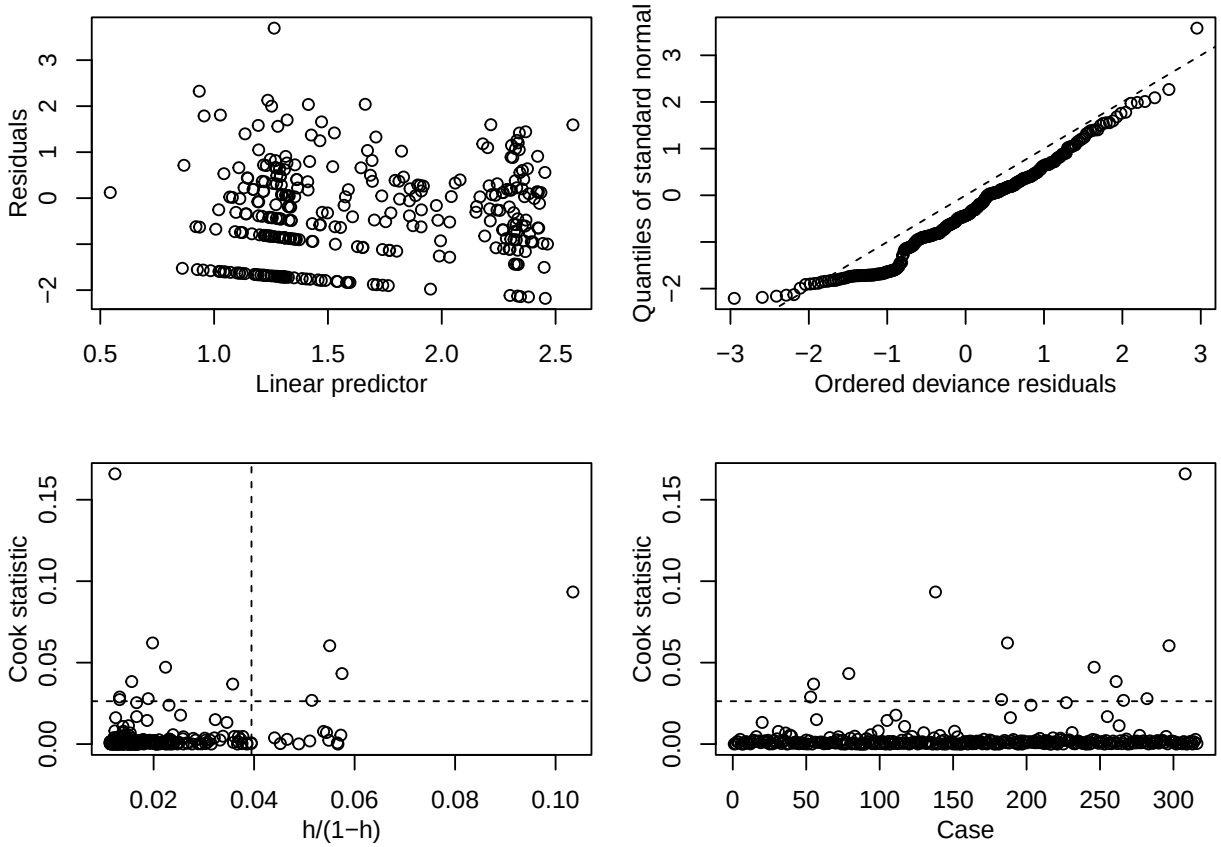
Table 16: Incidence Rate Ratios with 95% C.I.

	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.

Diagnostic Plots



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 affect attendance behavior for these students.