# Data Analysis with Quasi-Binomial and Negative-Binomial Regression Techniques

Jiaying Teng

2023-21-03

# Contents

# GLOVE DATA

**ABSTRACT** The purpose of this report is to investigate if attendance of an educational program regarding importance of glove usage improved actual glove usage during heart valve surgeries. Furthermore, this report also examines whether years of experience has any significant impact on overall glove usage. This report is done on a data set containing 92 observations with 4 variables. That of which being Period(time period when observed), Observed(number of times observed), Gloves(number of times gloves were worn when observed), and Experience(years of experience). The data collection process involves repeated observations over four time periods on 23 nurse's glove usage in a cardiology department at a specific hospital. More specifically, each nurse is observed numerous times throughout each time period on their particular glove use.

This report utilized quasi-binomial models to correct for over dispersion of regular binomial GLM models resulting from the data above. The results indicate that both the educational program and years of experience had significant impact on glove usage. Older, more experienced nurses were less likely to use gloves in comparison to their younger cohorts. Proportion of glove usage improved drastically compared to pre-program levels throughout every period. However, retention on the increase of proportional glove use decreased in tandem with levels of experience. To be more precise, more experienced nurses tended to suffer retention of increased glove usage compared to young nurses after attending the educational program.

More specifically, this report found that with no experience and no attendance of the educational program, the proportion of expected glove usage is 0.568. After taking the educational program, log odds of proportion of expected glove usage increased by 3.63 observed one months after intervention, 2.96 two months after intervention, and 1.966 five months after intervention. Furthermore, per every year of experience, a nurse decreases the log odds of proportion of glove use by -0.142.

**INTRODUCTION** Table 1 below outlines the first five observations of the data. In total, the given data set contains 4 variables and 92 observations. The data is collected on 23 nurses in a cardiology department from a specific hospital without their knowledge. This accounted for all nurses in that particular department. There are 29 missing rows of observations for which a nurse was not observed during a particular period. All nurses attended the educational program on the importance of glove usage. The variables are labeled and referred to as below:

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

Table 1: First 5 observations of data

| Period | Observed | Gloves | Experience |
|---|---|---|---|
| 1 | 2 | 1 | 15 |
| 2 | 7 | 6 | 15 |
| 3 | 1 | 1 | 15 |
| 4 | NA | NA | 15 |
| 1 | 2 | 1 | 2 |

The variable `Period` is treated as a factor denoting each of the four distinct time periods for observation. Furthermore, missing NA values in the data only occurs in tangent for both the variables `Observed` and `Gloves`. This is because missing data suggest that a nurse was not observed at all during that time period. There is no missing data for the variable denoting `Experience`.

For our model, proportion of glove use is treated as the response. The variables `Period` and `Experience` are treated as predictors. Thus, rows with missing observations indicate missing predictor values of proportion and thus are ignored and unaccounted for.

Since this an experiment on repeated measures of participants, there exist a high possibility of over dispersion a regular binomial GLM model. This is due to correlated sampling that causes the dispersion parameter to not equate to 1. Hence, this report will induce the quasi-binomial method instead. The quasi-binomial method accounts for the increase of variance to avoid possible over dispersion in our model.

To look at the full formula for quasi-binomial, please refer to **Quasi-Binomial** under **Glove Data** in the `Appendix` below.

Before any modeling is done however, it is appropriate to look at exploratory graphs of our data. Mainly the exploratory analysis will focus on missing data and any trends between different periods or years of experience.

**PRELIMINARY ANALYSIS:** To better connect visualizations in our data, 3 additional variables are created. The table below denotes the addition of three new variables for our exploratory analysis only. Note, these variables will not be included in our final model as they retain redundant information from the original data. The new variables are labeled and referred to as below:

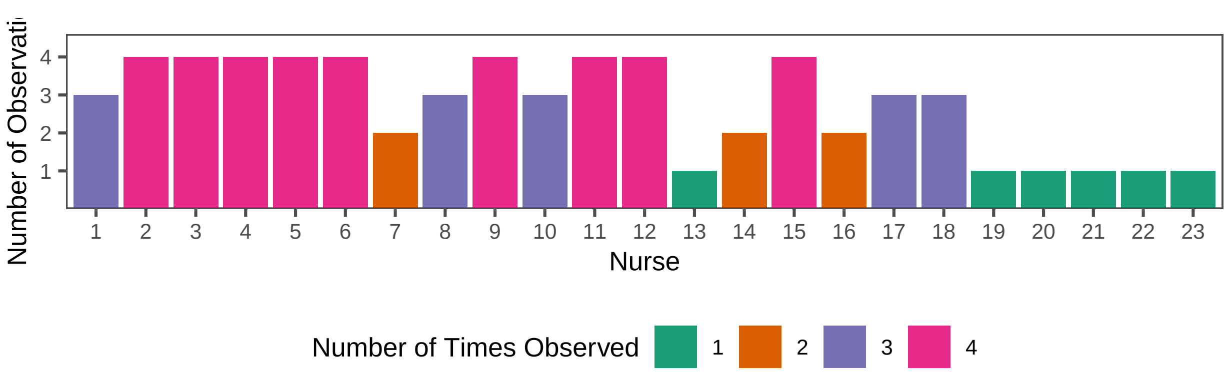
1. `Nurse`: Nurse that is observed (23 Nurses in total)
2. `Skill`: (Novice = (0-5 years), Intermediate = (6-9 years), and Seasoned = (>=10 years) levels of experience)
3. `Proportion`: `Gloves/Observed`

Table 2: First 5 observations of transformed data

| Period | Observed | Gloves | Experience | Skill | Nurse |
|---|---|---|---|---|---|
| 1 | 2 | 1 | 15 | Seasoned | 1 |
| 2 | 7 | 6 | 15 | Seasoned | 1 |
| 3 | 1 | 1 | 15 | Seasoned | 1 |
| 4 | NA | NA | 15 | Seasoned | 1 |
| 1 | 2 | 1 | 2 | Novice | 2 |

This report begin its preliminary study with graphical summaries on the missing observations. Below are two figures that explore the missing data by `Period` and by Nurse. The first of which looks at if the spread of missing data is even across all nurses, a bar graph is constructed below detailing how many observations each nurse have out of four.

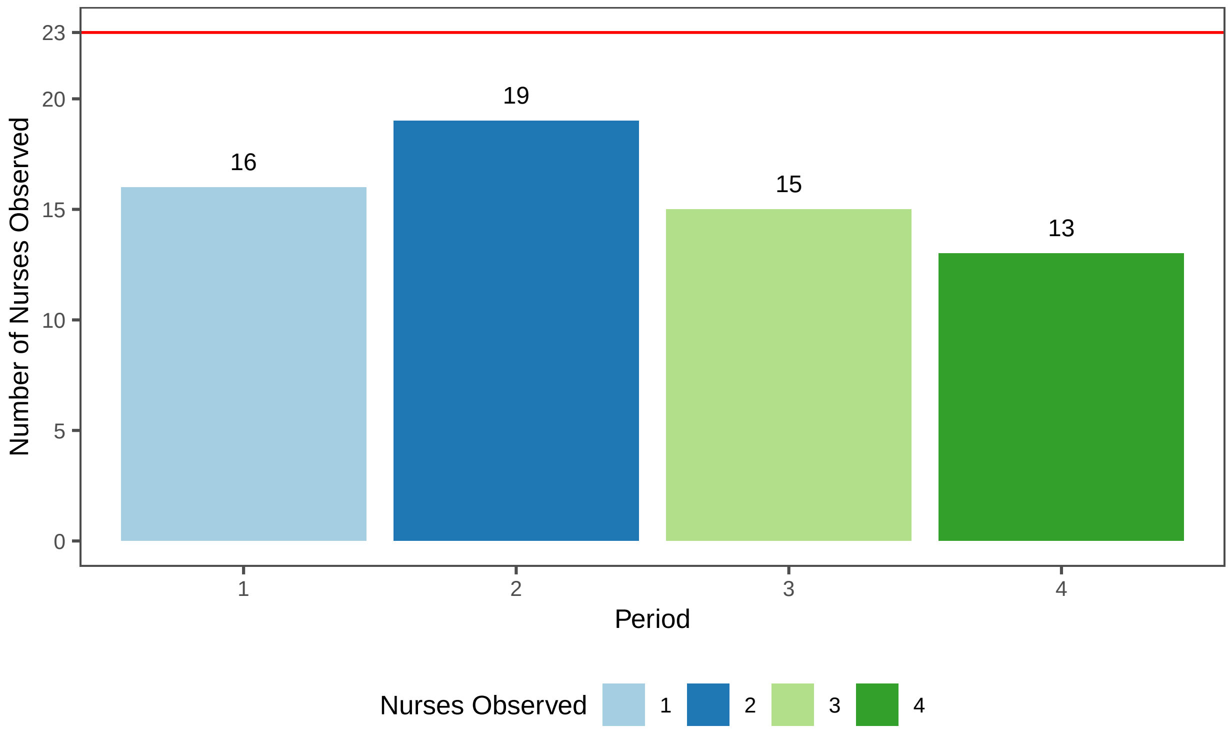
Figure 1: Number of Observation per Nurse



The above table counts the periods which a particular nurse is observed. The graph above relays information about possible missing data in our sample. The results indicate that there are five nurses with only one observed period and three nurses with only two observed periods. Thus, missing data does not appear to be uniform across all nurses. The graph suggests that missing data comes primarily from a few nurses with little to no observations. It is worth taking in consideration that the data collected may not be as diverse as it suggests. This will impact the reliability of conclusions since the data fails to capture diverse data about changes in glove usage throughout different periods.

To examine more closely how negatively it might impact our results, it is helpful to look at missing observations separated by distinct periods.

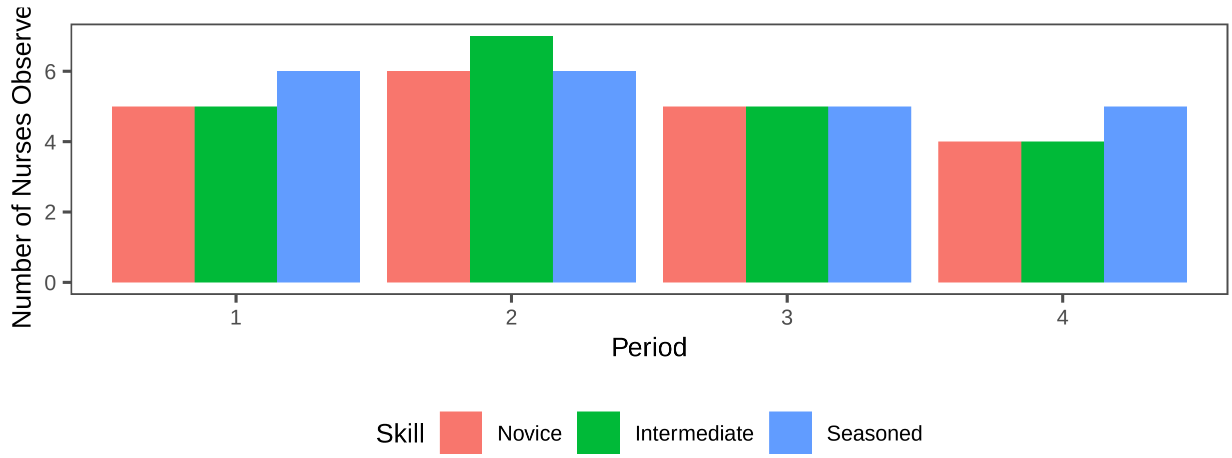
Figure 2: Number of nurse observation per period




From the figure above, it is clear that every period has missing data. It is good to note that although there is missing data, each period is missing proportional amount of data relative to each other. Although the lack of data would still effect the validity of end results, especially on this small of a data set, the results for changes throughout the period would prove to stay at least some what consistent. Period 4 is missing the most data with nearly half of the observations missing. Note that ideally, each period would have 23 observations, one for each nurse. Thus, this should be taken in account when interpreting the final model.

To visualize how negatively the missing data might affect our conclusions above `Experience` as a predictor, this report separates `Experience` into three categories. That of which being Novice (0-5 years), Intermediate (6-9 years), and Seasoned (>=10 years) levels of expertise. This separates the nurses into 8 novice nurses, 8 intermediate nurses, and 7 seasoned nurses. To ensure that the results are balanced between each group, the below table outlines the number of observations per each cohort group for distinct periods.

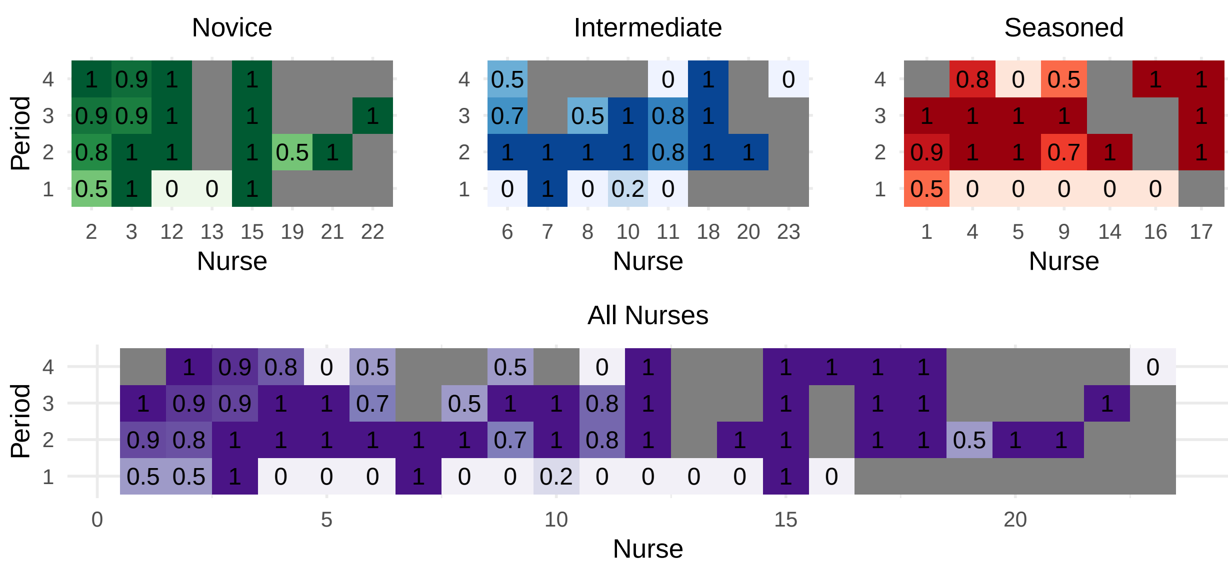
Figure 3: Number of nurse observations according to experience




As seen from above, there appears to be a equal distribution of observations for each period among the separated cohort groups by `Skill`. Novice and Intermediate levels of experience are most affected by loss of data overall. In period 4, both Novice and Intermediate cohort groups loses about 50% of their respective observations. However, the above graph conveys that the cut off points for separating the cohort groups seem appropriate for our purposes. The distinct cohort groups contain relatively the same amount of observations per each period and thus will convey a good balance of trend through out the periods for each level of `Skill`.

Now to connect the above graphics with the main question of whether or not attending the educational program actually had any effect on glove usage, a heat map is produced below to outline the proportion of glove usage for each nurse separated by cohort groups and overall.

Figure 4: Heat maps for proportional glove usage according to Experience




The figure above illustrates four heat maps that highlights proportion of glove usage for each nurse. The proportion is calculated as `Gloves` divided by number of times `Observed` for each `Period`. Having proportion be close to 0 suggests that the nurse has a low proportion of wearing gloves when observed whilst being close to 1 suggests that the nurse almost always wear gloves every time observed. The grey illustrates missing data for a particular nurse in a specific period. Looking at the heat map conveying All `Nurses`, the visualization supports our findings from above that missing data comes primarily from a cluster of nurses rather than being evenly spread.

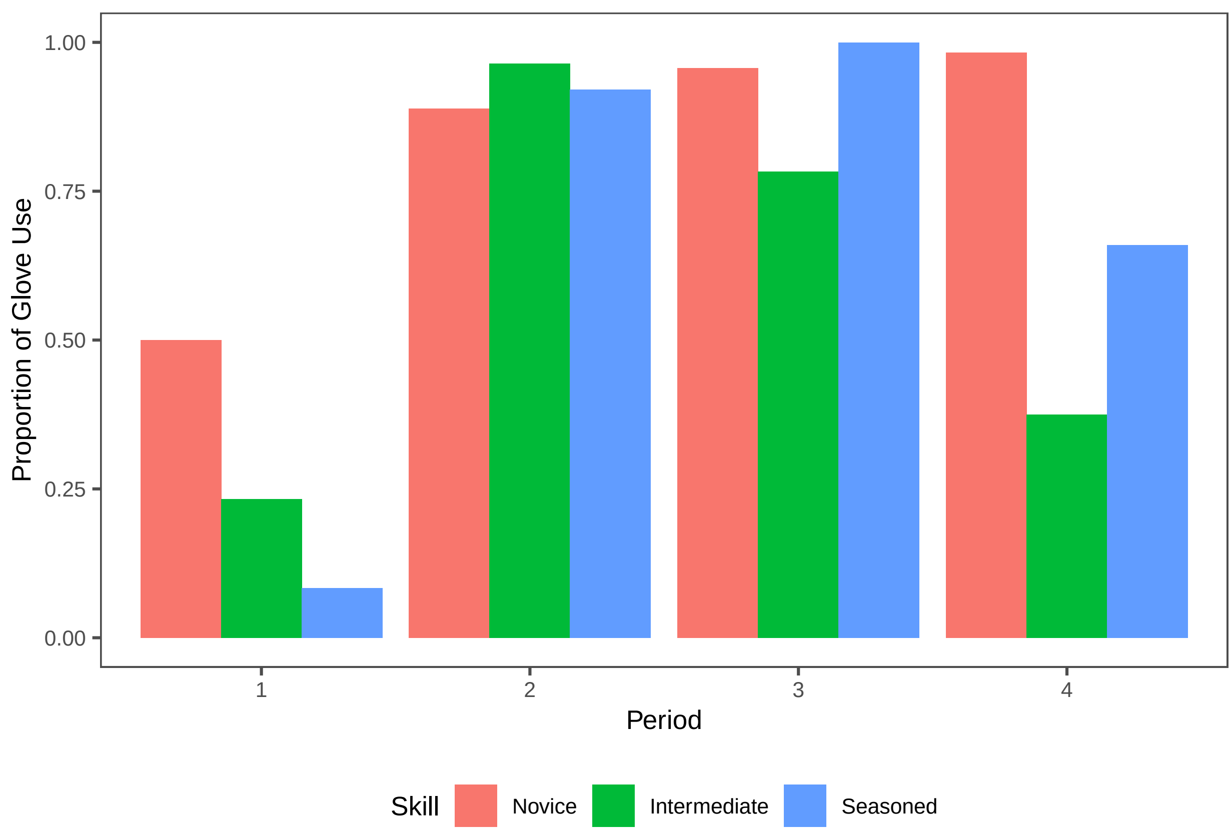
The graphs above support the claim that Experience and attendance of the educational program are significant factors in glove usage. Looking at primarily proportional glove usage before any intervention (Period 1) in the top three heat maps on the first row, seasoned nurses has low proportional glove usage compared to Novice and Intermediate nurses. Glove usage in intermediate groups also seem significantly lower than that of novice nurses. Thus, `Experience` may very likely be a significant predictor in proportional of glove usage.

Additionally, attending the educational program may have very likely increased glove usage. Every cohort group saw an increase in glove usage following attendance of the seminar in Period 2, one month after the program. This increase would remain compared to levels before attendance (Period 1) throughout period 3 and period 4. However, the heat maps convey that despite retention of glove use proportion being high in Novice nurses, Intermediate and seasoned nurses drop in retention in Period 3 and 4 (2 months and 5 months after intervention).

However, it is clear that across all cohort groups, proportional glove usage is lower across all levels before attending the seminar. The charts also clearly illustrates that an increase in Glove usage definitely did occur 1 months (Period 2) and 2 months (Period 3) after attending the educational seminar. And furthermore, the effects seem to be most impactful long term on Novice nurses where it saw the least decline in Glove Usage 5 months after intervention (Period 4).

To illustrate this point more clearly, below is a bar graph that depicts the proportion of glove usage across the 4 periods among the different cohort groups.

Figure 5: Proportion of Glove Usage per Period according to Experience




The bar graph above depicts a drastic increase in glove usage one month after attending the educational program. Interesting trends to note are that Novice nurses continued to retain high proportional usage of

Gloves while Intermediate and Seasoned nurses tend to decrease in glove usage in Period 3 and 4 (2 and 5 months after intervention). Interestingly, Intermediate nurses were the quickest to revert back to pre-intervention levels followed by seasoned nurses. Meanwhile, novice nurses. Although recall that this data set contains many missing observations that makes these confirmations unreliable without further analysis. Thus, although the above graphs support the claim that `Experience` is a significant predictor on glove usage, statistical analysis still needs to be performed to confirm actual significance.

To confirm this in a more statistical setting, this report will utilize Quasi-binomial methods to model the variables of `Period` and `Experience` as predictors to our response, which will be the proportion of glove usage. To obtain the response, this report will count the number of successes of glove over total number of times observed. To clarify, proportion will be represented by the variable, `Gloves`, divided by the variable, `Observed`.

**MODELING** This report begin the modeling process by first verifying the needed usage of Quasi-binomial models in preference to regular binomial GLMs. This is done by confirming through a Pearson's chi-square statistic on the residual deviance, divided by the degrees of freedom. When this test exhibits the need to reject the null hypothesis, it indicates that the assumption for binomial variability may not be valid and thus the model exhibits over dispersion as the dispersion parameter is not 1.

Below is a summary table of the binomial model, denoted as `model 1`, labeled by the follow equation:

Table 3: Summary Output of 'Model 1'

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 0.2734 | 0.4517 | 0.6051 | 0.5451 |
| **Period2** | 3.633 | 0.5927 | 6.13 | 8.772e-10 |
| **Period3** | 2.96 | 0.5661 | 5.228 | 1.714e-07 |
| **Period4** | 1.996 | 0.4982 | 4.007 | 6.139e-05 |
| **Experience** | -0.1424 | 0.03595 | -3.96 | 7.491e-05 |

(Dispersion parameter for binomial family taken to be 1 )

| | |
|---|---|
| Null deviance: | 179.40 on 62 degrees of freedom |
| Residual deviance: | 85.16 on 58 degrees of freedom |

The Chi-square test from the above results with 85.16 on 58 degrees of freedom resulted in a p-value of 0.01161295, indicating that the model does have issues with fit and is possibly missing sources of data variation in a binomial GLM. This is a case for evidence of over dispersion in Binomial data with repeated sampling and thus with this confirmation, this report will utilize Quasi-binomial models instead to account for added variation.

Below is a summary output of the quasi-binomial model with proportion of glove usage as the response and `Period` and `Experience` as predictors.

Table 4: Summary Output of 'Model 2'

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 0.2734 | 0.5511 | 0.4961 | 0.6217 |
| **Period2** | 3.633 | 0.723 | 5.025 | 5.124e-06 |
| **Period3** | 2.96 | 0.6906 | 4.286 | 6.957e-05 |
| **Period4** | 1.996 | 0.6077 | 3.285 | 0.001732 |
| **Experience** | -0.1424 | 0.04385 | -3.246 | 0.001944 |

(Dispersion parameter for quasibinomial family taken to be 1.488031 )

| | |
|---|---|
| Null deviance: | 179.40 on 62 degrees of freedom |
| Residual deviance: | 85.16 on 58 degrees of freedom |

As there only exist two predictors maximum, the model above, which will be denoted as `model 2` signifies that all predictors are significant in determining proportion of glove use. Since all estimated coefficients appears to be highly insignificant, there is no need to do variable selection and the best basic model without any interaction terms is the quasi-binomial model with both `Experience` and `Period` as significant predictors.

`model 2` above follows the below equation:

$$logit(\frac{Gloves}{Observed}) = 0.2734 + (3.633)Period_1 + (2.96)Period_2 + (1.966)Period_3 + (-0.1424)Experience \quad (1)$$

**CHECKING FOR INTERACTIONS:** Since there are only two predictors, the only interaction effect that is needed to be checked is between Period and Experience. Below is a summary table for the model including this interaction term.

Table 5: Summary Output of 'Model 3'

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 1.035 | 0.8664 | 1.195 | 0.2372 |
| **Period2** | 1.747 | 1.356 | 1.288 | 0.203 |
| **Period3** | 1.111 | 1.178 | 0.9436 | 0.3495 |
| **Period4** | 2.087 | 1.247 | 1.674 | 0.09988 |
| **Experience** | -0.2411 | 0.104 | -2.32 | 0.0241 |
| **Period2:Experience** | 0.2128 | 0.1471 | 1.446 | 0.1538 |
| **Period3:Experience** | 0.2369 | 0.1426 | 1.662 | 0.1023 |
| **Period4:Experience** | -0.008109 | 0.1397 | -0.05806 | 0.9539 |

The above summary suggests that there is no interaction effect between the variable `Experience` and the factor variable `Period`. From the p-values, every interaction effect is insignificant and thus is not taken into account. Thus the report deduces that there are no interaction effects between the predictor variables. Furthermore, since `Period` is a factor variable, dropping one interaction term suggests that we drop all interactions between Period and Experience. Thus, the final best model remains to be `Model 2` with fitted predictors of Experience and Period with no interaction terms from above.

**FINAL MODEL:** Thus, the best model is represented by model 2 with the following quasi-binomial regression formula:

$$logit(\frac{Gloves}{Observed}) = 0.2734 + (3.633)Period_1 + (2.96)Period_2 + (1.966)Period_3 + (-0.1424)Experience$$

The results of the above model suggests that with no attendance of the educational program and zero years of experience, proportion of Glove usage will be $\frac{e^{0.2734}}{1+e^{0.2734}} \approx 0.568$.

Recall above that Period is a factor variable, thus only one of the three Period variables above will be considered when calculating for the log odds of proportional glove usage. Thus, holding the variable Experience constant, being in Period 1, one month after attendance of the educational seminar increases the log odds of proportion of glove usage by 3.633. Being in period 2, three months after attendance of educational seminar, increases the log odds of proportion of glove usage by 2.96. being in period 3, five months after attendance of educational seminar, increases the log odds of proportion of glove usage by 1.996. Finally, holding the Period factor variable constant, per every increase in year of experience for a Nurse, the log odds for proportion of glove usage decreases by -0.1424.

Thus, the increase in glove usage is most effective one month after the educational program. As time passes, the educational program remains effective although not as impactful to early months. Furthermore, increases experience decreases chances of a Nurse using gloves.

**LIMITATIONS:** Limitations in our data mainly comes from the combination of small data sample and large amount of missing observations. The results above would be much more reliable if there was no missing or very few missing observations for the nurses sampled in the cardiology department of this particular hospital. Furthermore, it helps to illustrate that there are only two predictor variables. Though a general is identified that statistically infers the significance of Period and Experience as predictor variables, more collection on different variability will aid in confirming these results. Some suggestions include collecting data on a more frequent time scale or information about when observations on glove usage is collected.

**CONCLUSION:** In conclusion, both nurse experience and the attendance of the educational program affected the usage of Gloves for the nurses in this department. Attendance of the educational program positively affected glove usage while years of experience negatively impacted glove usage.

The summary results of our final model suggests that With no experience and no educational seminar, the proportion of glove usage is 0.568. After taking the seminar, observed nurses within one month had an increase in log odds of proportional usage of gloves of 3.63. Furthermore, this retained in Period 2, three months after taking the seminar, with observed increase in log odds of proportional usage of gloves of 2.96 compared to pre-seminar proportions. And finally, compared to pre-attendance of the educational program, log odds of glove usage remain increased by 1.996 5 months after attendance. Experience on the other hand had a negative inverse relationship with glove use. Per our report, the results suggest that per every year of increased experience, a nurse decreases the log odds of proportional glove use by -0.142.

Summarizing the above, Glove usage increased drastically within the first 3 months after attending the seminar. Retention for increased activity of glove usage dropped after the first month but remained high compared to prior glove usage before the educational program. Combining with our exploratory analysis in the report, this drop in retention coincided with levels of Experience for the nurse. More experienced nurses suffered harsher retention drops of increased glove usage compared to newer, less-experienced nurses.

This report suggests increased levels of data collection to ensure and improve upon accuracy of above results. Due to large amount of missing data, this report is not completely confident in its findings beyond just basic confirmations of possible relationship between the educational program and glove usage proportion. Further research can involve examination of glove usage over longer period of times with more frequent observations and diverse time periods. Furthermore, it would be more helpful to incorporate additional variables such as difficulty of operation for more explanatory variables that might affect actual glove usage.

# SCHOOL ABSENCES

**ABSTRACT:** The purpose of this report is to discern attendance behavior based on a data set of 316 randomly selected sixth graders from two elementary schools in a particular school district. More precisely, an in depth analysis will be done examining whether there exists statistical significance and relationship between absences and the variables of school, gender, math and language scores. In our research, absence, which is a count variable, will be the response with the remaining 4 variables serving as predictors.

The following research will perform negative binomial regression to account for over dispersion of our response parameter. More specifically, the variance of our response variable `absence` is much greater than that of our mean. Thus the basic assumption of poisson regression is violated since dispersion parameter would not equal to 1. The basic assumptions therefore for our model of interest is that the response variable `absence` follows a negative binomial distribution and that using a log link, the log of the mean of our response have a linear relationship to our predictors.

Overall, this report found that variables of school, gender, and language proved to be significant predictors. More specifically, students from school 1 were expected to be absent at a rate of 2.13 greater than of school 2. Female students are expected to be absent a rate 1.489 times more than their male counterparts. Finally, per one score increase of standardized language score, the rate of days absent decreases by a factor of .99154.

**INTRODUCTION** Table 1 below outlines the first five observations of the data. In total, the given data contains 5 variables and 316 observations with no missing data. More specifically, the data is collected from 316 randomly sampled sixth graders from two different elementary schools. The variables are labeled and referred to as below:

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

Table 8: First 5 observations of data

| 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 |

T he variable of interest ,`absence`, conveys count data. Thus, it suggest the usage of poisson regression. However, this report found a poisson model suffers from over dispersion from the data above. This over dispersion is caused by the difference in mean and variance from the response variable, `absence`. This vioates the basic assumption that `absence` follows a poisson distribution as its mean and variance are not equal. Thus, this will also violate the basic assumption that the dispersion parameter for poisson regression should be 1, thus making any modeling results with basic poisson regression skewed and possibly inaccurate. Thus, in this regard, this report will tackle over dispersion using a negative binomial model which incorporates an additional term to account for the excess variance.

**PRELIMINARY ANALYSIS:** The preliminary analysis will conduct a brief examination on the distribution of our data and any visual patterns that can be obtained from our observations.

Table 2 below is a summary output of statistical measures of the the data. This table will provide insight on if there exists any possible invalid observations in the data.

Table 2: Summary Output of Data

| | N | Mean | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| school | 316 | 1.50 | 0.50 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |
| math | 316 | 48.75 | 17.96 | 0.50 | 37.80 | 49.15 | 60.90 | 99.80 |
| language | 316 | 50.13 | 17.99 | 0.20 | 40.60 | 50.05 | 60.80 | 100.00 |
| absence | 316 | 5.81 | 7.45 | 0.00 | 1.00 | 3.00 | 8.00 | 45.00 |

One clear issue displayed in the summary table lies in the greater difference between the mean and the standard deviation of our response variable `absence`. Here, it is clear that the assumed dispersion parameter of 1 would be violated if we were to use a basic poisson regression. Over dispersion can happen in most GLM models but since Poisson assumes that the mean equates to the variance, this problem is exacerbated.

To examine more closely how the variance and mean relate to each other, it helps to first look at the shape distribution of our response variable depicted by the histogram below.

Figure 1 below shows the distribution of our response variable `absence`.

Figure 1: Histogram of Absence Data

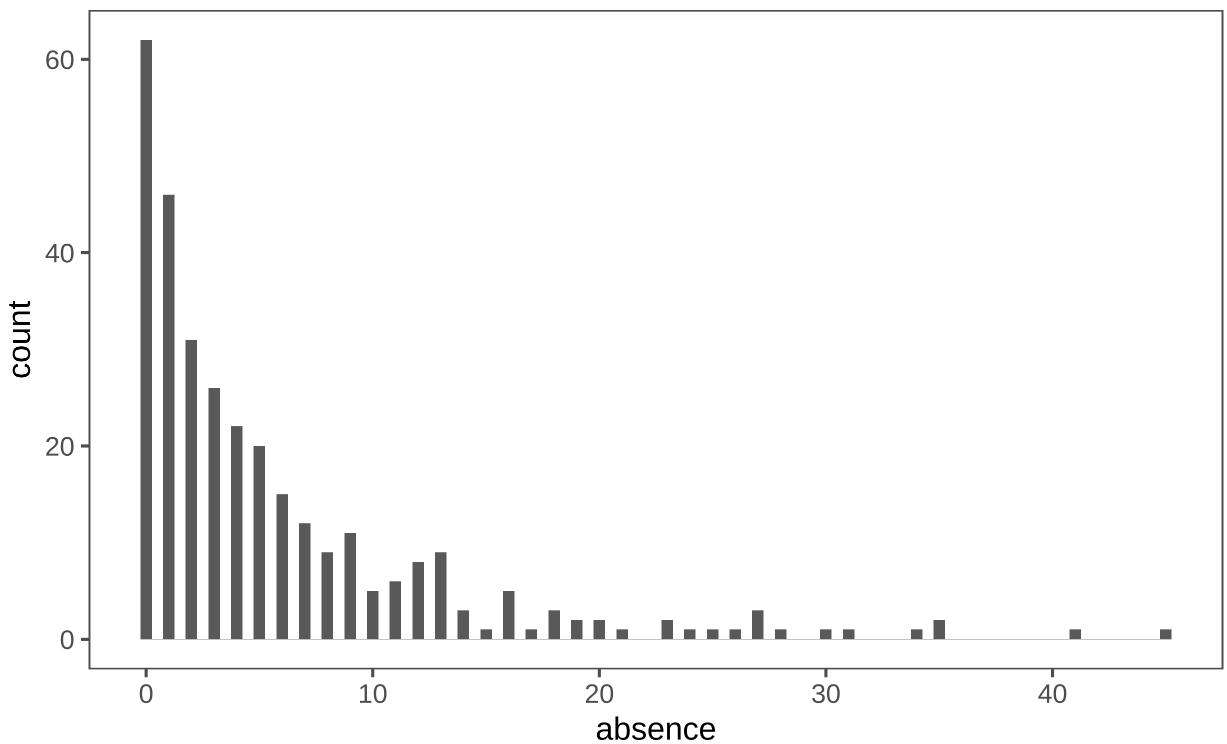
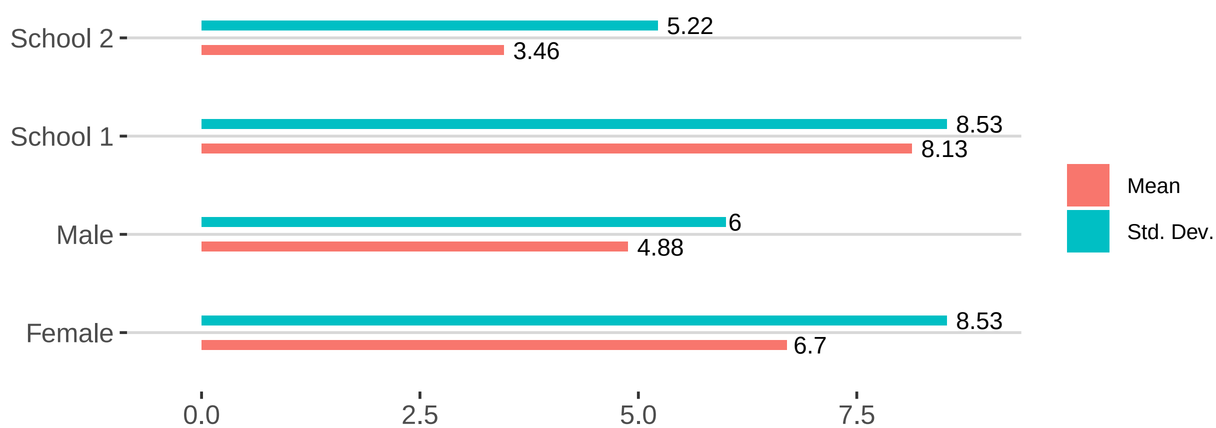



Figure 1 above shows a heavily skewed right distribution involving large concentration of small values with few counts of large values. This is a typical case of possible violation of the mean and variance equating to each other in poisson cases. From the shape of the data, it suggests very likely that the mean is much smaller than the variance, therefore implying that a regular poisson GLM fit will suffer from over dispersion.

To verify this suspicion beyond visual aid of the distribution's shape, figure 2 below also outlines the mean and standard deviation of each indicator cohort group numerically.

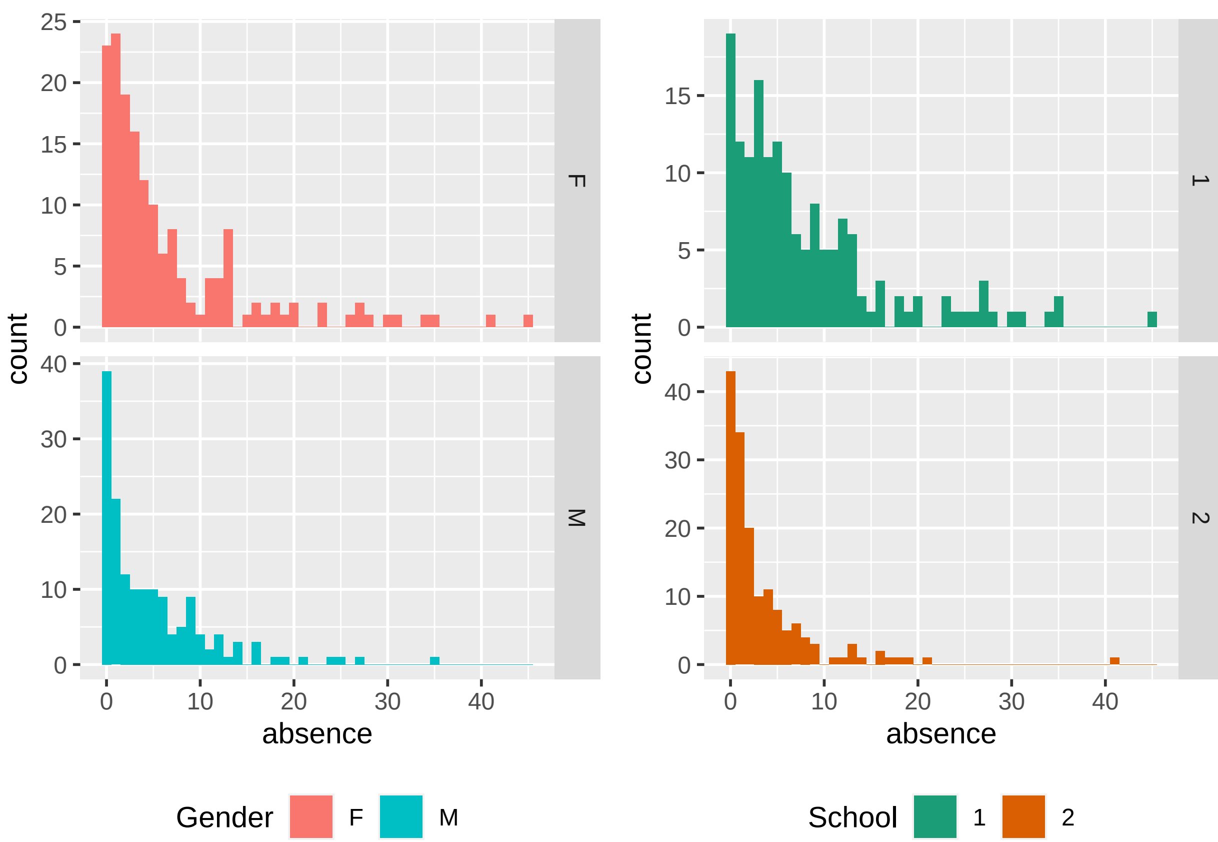
Figure 2: Mean and Standard Deviation of Indicator Cohort Groups




From the values above, we see that the standard deviation, regardless of the result of the indicator variable, is much higher than the mean. Thus, it would make sense to employ negative binomial regression which accounts for the gross over saturation of the variance over the mean for violation in assumption of our response following a poisson distribution.

To examine more closely possible relationships between different variables, it helps to examine distributional differences in absence between different cohort groups. In figure 3 below, the groups are divided into different cohorts to examine possible differences between the indicator variables.

Figure 3: Histogram of Absence Data by Indicator Variables



Above, Figure 3 depicts four histograms that conditions seperation based on the variables of `gender` and `school`. Here, the visual highlights that students who are female are more likely to be absent than their male counterparts. Furthermore, the graph also shows if divided into groups just by school, school 1 students are much more likely to be absent than their peers from school 2.

To verify if this relationship is actually significant, the report will construct negative-binomial regression models that outlines the statistical significant between the response and each of the predictor variables.

**PRELIMINARY MODEL:** Selection of preliminary model begins with confirmation of our suspicion on over dispersion with a fully fitted poisson regression. For the summary result for this model, please refer to the **summary output of `model 1`** under **Student Data** in the `Appendix`.

`Model 1` is a poisson regression model with residual deviance of 2007 on 311 degrees of freedom. This accounts for a ratio of 6.4 and therefore clearly violates the assumption that the dispersion parameter is 1, therefore confirming that poisson regression by itself suffers from over dispersion.

Thus, this report will instead utilize a negative binomial model to fit our response variable `absence` against all predictors. This will be done using the glm.nb() from the R package, MASS.

Table 2 below outlines the summary report of our fully fitted negative binomial model with `absence` as our response.

Table 2: Summary output of 'Model 2'

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 2.622 | 0.2234 | 11.74 | 8.029e-32 |
| **school2** | -0.7653 | 0.1446 | -5.292 | 1.21e-07 |
| **genderM** | -0.398 | 0.1339 | -2.973 | 0.002953 |
| **math** | 0.002106 | 0.005087 | 0.4139 | 0.6789 |
| **language** | -0.01005 | 0.005242 | -1.916 | 0.05531 |

The model, which we will denote as `model 2` follows the negative binomial equation below:

$$log(absence) = (2.622) + (-0.765)school + (-0.398)gender + (0.002)math + (-0.01)language \qquad (2)$$

From the summary output above, the fully fitted model suggests that the variable, `math`, is an insignificant predictor. This can be confirmed further with AIC or step regression. But below, this report confirmed the removal of this main effect using Pearson's Chi Square test. The table below includes the summary output of the anova summary output.

Table 3: Anova summary output comparing new model'

| Model | theta | Resid. df | 2 x log-lik. | Test | df | LR stat. | Pr(Chi) |
|---|---|---|---|---|---|---|---|
| -Math | 0.8701 | 312 | -1734 | | NA | NA | NA |
| Model 2 | 0.8705 | 311 | -1734 | 1 vs 2 | 1 | 0.1931 | 0.6603 |

Model 2, which contains the variable math resulted in a p-value of 0.6603 against the model without math, indicating that adding math as a predictor does not significantly affect the model. Thus we exclude math from our model.

Below will be a table outlining the summary output of the model with math removed as a predictor, denoted as `model 3`.

Table 4: Summary output of 'Model 3'

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 2.642 | 0.2143 | 12.33 | 6.152e-35 |
| **school2** | -0.7553 | 0.1432 | -5.275 | 1.327e-07 |
| **genderM** | -0.3981 | 0.1333 | -2.986 | 0.002829 |
| **language** | -0.008492 | 0.004014 | -2.115 | 0.0344 |

(Dispersion parameter for Negative Binomial(0.8701) family taken to be 1 )

| | |
|---|---|
| Null deviance: | 410.4 on 315 degrees of freedom |
| Residual deviance: | 357.4 on 312 degrees of freedom |

The model, which we will denote as `model 3` follows the negative binomial equation below:

$$log(absence) = 2.642 + school(-0.755) + gender(-0.398) + language(-0.00849) \tag{3}$$

From the above results, it is evident that the remaining variables are all significant. Furthermore, notice the residual deviance on degrees of freedom ratio is close to one, thus this negative binomial model suffers much less from over dispersion compared to `model 1`, the basic GLM poisson model.

Thus, the best basic model we obtain the variables of `school`, `gender`, and `language` as significant predictors.

**CHECKING INTERACTION TERMS:** To fully understand how the variables affect `absence` in the above data set, it is important to check for interaction terms in the model. Since there are few predictor variables, it is possible to find significant interactions using AIC through stepwise backwards selection.

To fully see the process of variable selection, please check **Variable Selection of `model 2`** under **Student Data** in the *Appendix* below.

In this report, `school:gender` as well as `gender:math` are found to have significant interaction effects. Below represent the summary output of the model with interaction terms.

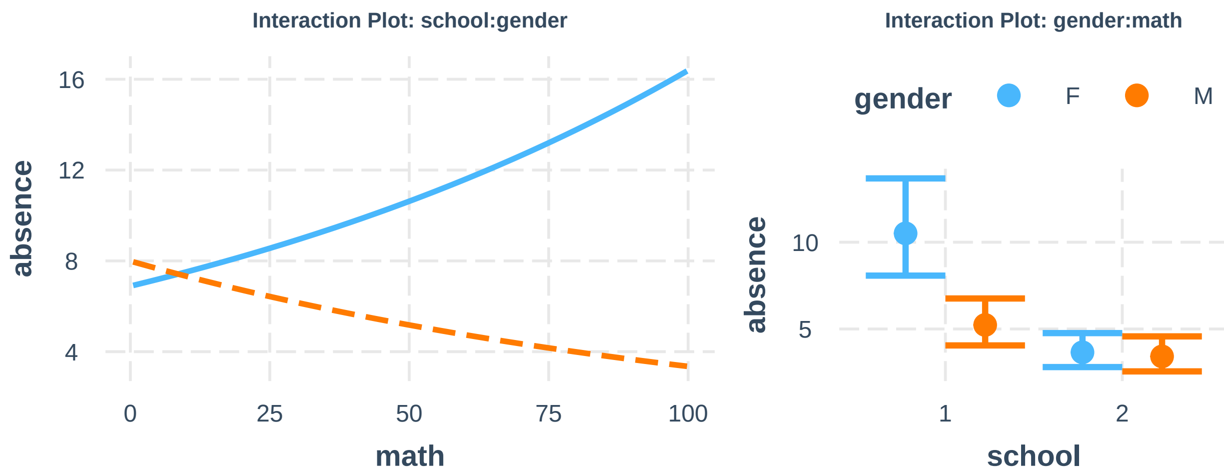
Table 5: Summary output of 'Model 4' and Anova Test (Model 3 vs Model 4)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 2.334 | 0.2831 | 8.246 | 1.634e-16 |
| **school2** | -1.057 | 0.1989 | -5.312 | 1.084e-07 |
| **genderM** | 0.1486 | 0.37 | 0.4016 | 0.688 |
| **math** | 0.00867 | 0.006159 | 1.408 | 0.1592 |
| **language** | -0.008067 | 0.005188 | -1.555 | 0.12 |
| **school2:genderM** | 0.6309 | 0.2812 | 2.243 | 0.02487 |
| **genderM:math** | -0.01736 | 0.007719 | -2.249 | 0.02452 |

| Model | theta | Resid. df | 2 x log-lik. | Test | df | LR stat. | Pr(Chi) |
|---|---|---|---|---|---|---|---|
| Model 3 | 0.8701 | 312 | -1734 | | NA | NA | NA |
| Model 4 | 0.9008 | 309 | -1726 | 1 vs 2 | 3 | 7.974 | 0.04654 |

From above, it is clear that the model with added complexity is significant from the p-value that resulted in the Pearson's Chi-square test. Thus, it is appropriate to construct appropriate plots to look at our two significant interaction effects to determine whether or not they are appropriate or logical to include in the final model.

Figure 4: Interaction plots




From the figures above, it is clear that both interaction terms are indeed significant. For the interaction effect between math and gender, it appears that as math score increase, the amount of absences go up for female student whilst inadvertently they are inverse for male students.

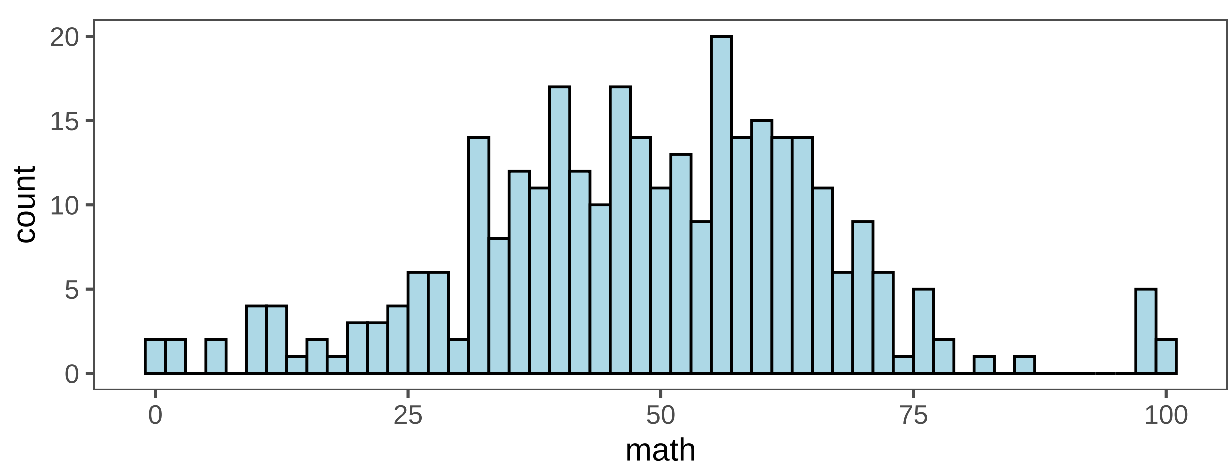
Furthermore, for gender and school, it appears that the female students in school 1 are much more absent in comparison to their male counter part. Furthermore, it appears that males and females do not differ at all in terms of absences in school 2.

Following these interaction plots, it becomes more clear why gender no longer is significant in our summary result above compared to our model without interactions.

However, notice in the summary result that the coefficient for the main effect of math is extremely small. Furthermore, notice that for the interaction effect between math and gender, the point of crossing lies near the very edge of the data.

Below is a histogram that outlines the distribution of math scores in our data.

Figure 5: Histogram of Math Scores




As seen from above, the histogram reveals that only a few data points reside near the point of intersection for our interaction plot between `math` and `gender` above. Thus, the result of this interaction being significant may not be accurate. Furthermore, the added complexity does not offer significantly better results compared to our basic model. Thus, for the purposes of clarity and better interpretability, this report aims to opt out this interaction effect from our final model. However, it warrants future research on possible consequences for the significance of such interaction effects.

Following the removal of interaction effect of `math:gender` as well as the main effect of `math` since it is insignificant according to p-value, we get the following model output.

Table 5: Summary output of 'Model 5'

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 2.729 | 0.2199 | 12.41 | 2.238e-35 |
| **school2** | -0.9428 | 0.1916 | -4.921 | 8.61e-07 |
| **genderM** | -0.5792 | 0.1794 | -3.228 | 0.001248 |
| **language** | -0.008369 | 0.004009 | -2.088 | 0.03684 |
| **school2:genderM** | 0.3929 | 0.264 | 1.488 | 0.1367 |

Thus, the interaction effect between school and gender also becomes insignificant and thus is removed from the model. However, both interaction effect should warrant further research on possible conditions for such relationships.

Removing both interaction effects above, we end up with `model 3` as our best final model.

**FINAL MODEL** Thus, the best model is represented by `model 3` with the following negative binomial regression formula:

$$log(absence) = 2.642 + school(-0.755) + gender(-0.398) + language(-0.00849)$$

Although the interaction term was significant in model 4, the added complexity did not warrant the loss of interpretability. Model 3 above can be easily interpreted and suffices for the purpose of this study.

The model above can be interpreted as follow: if all variables were set to zero except for the intercept, then a student is expected to have $e^{2.642} \approx 14$ absences. In this case, it would be a male student who went to school 2 and score all zeroes on his standard tests.

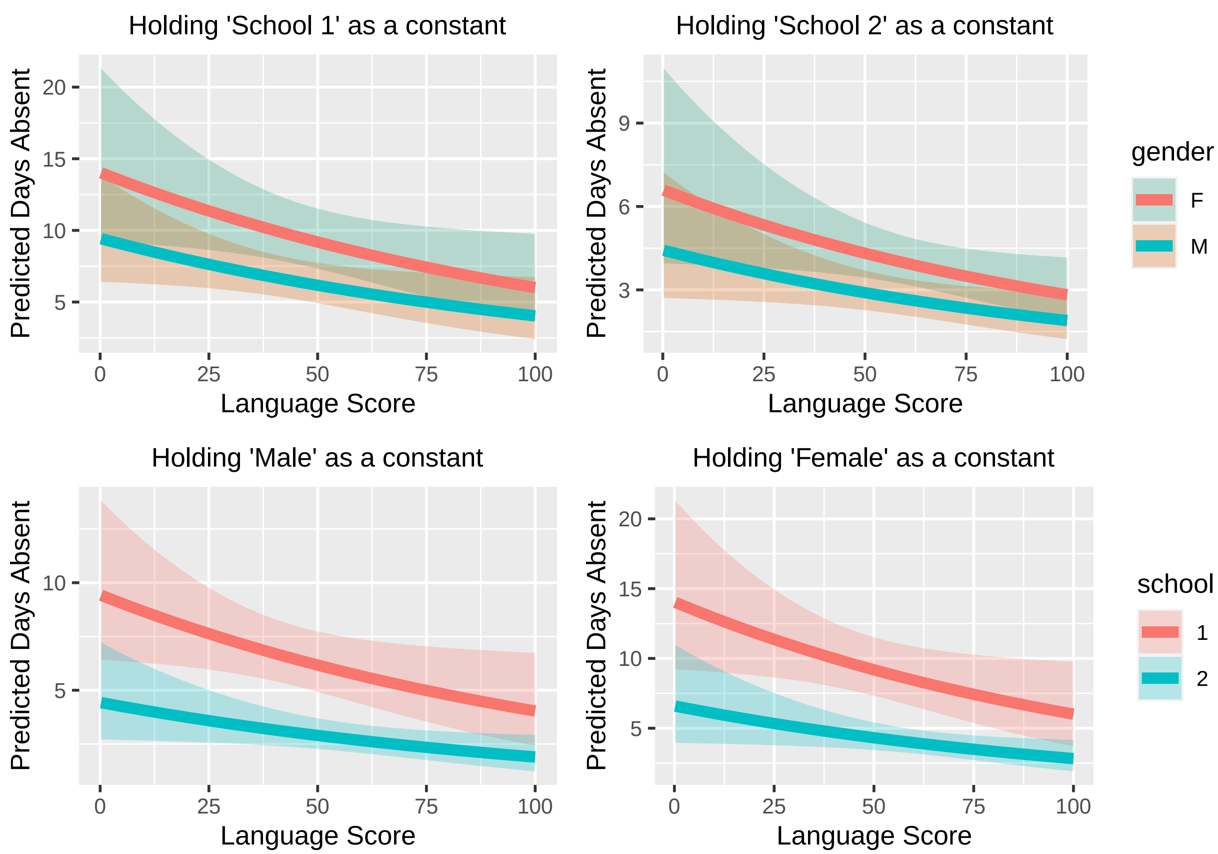
Similarly, if all other variables were held constant, a student from school 2 is expected to have a rate of absence 0.755 lower than that of school 2. In other words, $e^{-0.755} \approx 0.47$ suggests that a student from school 1 has a rate of absence 0.47 compared to a student from school 2. Or alternatively, a student from school 1 is expected to be absent at a rate $1/0.47 \approx 2.13$ greater for days absent.

For gender, if all other variables were held constant, then a male student would have a rate of absence $e^{-0.398} \approx 0.67$ compared to female student. In other words, a female students are expected to have a rate 1.489 times greater for days absent compared to male students.

Finally, for language score, considering all other variables were held constant, a one point increase in standardize language scores would decrease the rate for days absent by a factor of $e^{-0.00849} \approx .99154$.

**PREDICTION INTERVALS:** To more clearly illustrate how each predictor influence the number of days absent, below is figure six which depicts 95% confidence interval for each of our different cohorts. The below graphs holds other variables constant to highlight how absences change as the variable `language` increase.

Figure 6: Confidence Interval (95 Percent) based on fixed Constants




The graphs above show that if School 1 is held constant, female students are much more likely to be absent than male students. The same if School 2 is held constant but notice the y-axis of the graph and it is shown that female students do not have as great of a difference in absence compared to males in consideration to School 1. Furthermore, if the gender Male is held constant, then students from school 1 is absent much more than students in school 2. Having a student being Female does not change this claim and is in fact exacerbated as Female students from school 1 is much more absent than female students from school 2. Notice the confidence band being much larger on the sides of the graph compared to the center. This is due to a lack of data to confirm precision of the 95% confidence interval near the extremes.

**LIMITATIONS** Key limitations involve the small diversity in sample from our data. The key question involves finding significant statistical relationship between school, gender and standardized test for students but it is difficult to discern with only students from two schools. Furthermore, this issue is worsened by the fact that the two schools appear to have very different attendance habits that makes it difficult to evaluate actual significance of test scores on student absences. Since the results are so evidently clear that students from School 1 tends to be more absent than students from School 2, it makes it difficult to form a generalized relationship on a larger population on the effects of standardized scores on absences.

**CONCLUSION** In conclusion, this report has found that absences ties closely to the variables of school, gender, and standardized language scores. More specifically, school 1 students are expected to be absent at a rate of 2.13 greater than that of school 2 students. Furthermore, female students are expected to be absent at a rate 1.489 times greater than that of days absent for male students. Finally, for every one point increase in standardized language scores, the rate for days absent should decrease by a factor .99154.

These results suggest which specific school the student reside in results heavily in correlation with the number of days absent. Furthermore, gender itself also plays a significant role. However, during this research, it was discovered that there may exist a possible relationship and correlation when considering school and gender in tandem. To be precise, it appears that female students are much more likely to be absent compared to their male counter part in school 1 compared to students in school 2.

Although statistical significance was found, this research encourages data collection on more diverse schools to form a more general consensus on the effects of standardized testing on number of days absent. Further study is encouraged as the current findings support very narrowly only results on these two specific schools and even then, there appears to be possible correlation between significant hidden variables that does not exist in the current data. It is encouraged that more parameters are taken such as Grades or Distance away from School to identify whether or not other more important factors causes the variance in school 2 and school 1. Furthermore, it is encouraged to study more closely upon the difference between female and male students in school 1 to discern a reasonable conclusion to the difference between the two gender in attedance.

# APPENDIX

## GLOVE USAGE

### Quasi-binomal regression

Regular Binomial pdf:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Quasi-binomial pdf:

$$P(X = k) = \binom{n}{k} p(p + k\phi)^{k-1}(1 - p - k\phi)^n - k \;\;,\;\; where \; |\phi| \leq min[p/n, \; (1 - p)/n]$$

The $\phi$ parameter describes additional variance in the data that can not be explained by the binomial distrbution alone

## SCHOOL ABSENCES

### Summary Output of `Model 1`

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 2.578 | 0.07317 | 35.23 | 6.755e-272 |
| **school2** | -0.8063 | 0.05562 | -14.5 | 1.271e-47 |
| **genderM** | -0.4354 | 0.04819 | -9.036 | 1.619e-19 |
| **math** | 0.000906 | 0.001801 | 0.5031 | 0.6149 |
| **language** | -0.007261 | 0.001901 | -3.82 | 0.0001333 |

(Dispersion parameter for poisson family taken to be 1 )

| | |
|---|---|
| Null deviance: | 2410 on 315 degrees of freedom |
| Residual deviance: | 2007 on 311 degrees of freedom |

### Variable Seletion of `Model 2`

```
## Start:  AIC=1751.21
## absence ~ (school + gender + math + language)^3
##
##                           Df    AIC
## - school:math:language    1 1749.2
## - school:gender:language  1 1749.2
## - gender:math:language    1 1749.4
## - school:gender:math      1 1749.7
## <none>                      1751.2
##
## Step:  AIC=1749.21
## absence ~ school + gender + math + language + school:gender +
```

```
##     school:math + school:language + gender:math + gender:language +
##     math:language + school:gender:math + school:gender:language +
##     gender:math:language
##
##                          Df    AIC
## - school:gender:language  1 1747.2
## - gender:math:language    1 1747.4
## - school:gender:math      1 1747.7
## <none>                      1749.2
##
## Step:  AIC=1747.25
## absence ~ school + gender + math + language + school:gender +
##     school:math + school:language + gender:math + gender:language +
##     math:language + school:gender:math + gender:math:language
##
##                        Df    AIC
## - gender:math:language  1 1745.4
## - school:gender:math    1 1746.1
## <none>                    1747.2
## - school:language       1 1749.0
##
## Step:  AIC=1745.39
## absence ~ school + gender + math + language + school:gender +
##     school:math + school:language + gender:math + gender:language +
##     math:language + school:gender:math
##
##                      Df    AIC
## - gender:language     1 1743.4
## - school:gender:math  1 1744.1
## - math:language       1 1744.9
## <none>                  1745.4
## - school:language     1 1747.1
##
## Step:  AIC=1743.41
## absence ~ school + gender + math + language + school:gender +
##     school:math + school:language + gender:math + math:language +
##     school:gender:math
##
##                      Df    AIC
## - school:gender:math  1 1742.1
## - math:language       1 1743.0
## <none>                  1743.4
## - school:language     1 1745.2
##
## Step:  AIC=1742.12
## absence ~ school + gender + math + language + school:gender +
##     school:math + school:language + gender:math + math:language
##
##                   Df    AIC
## - school:math      1 1740.7
## - math:language    1 1741.5
## <none>               1742.1
## - school:language  1 1743.6
## - gender:math      1 1744.0
```

```
## - school:gender    1 1744.4
##
## Step:  AIC=1740.74
## absence ~ school + gender + math + language + school:gender +
##     school:language + gender:math + math:language
##
##                   Df    AIC
## - math:language    1 1740.7
## <none>              1740.7
## - school:language  1 1741.6
## - gender:math      1 1742.9
## - school:gender    1 1743.4
##
## Step:  AIC=1740.66
## absence ~ school + gender + math + language + school:gender +
##     school:language + gender:math
##
##                   Df    AIC
## - school:language  1 1740.0
## <none>              1740.7
## - school:gender    1 1743.4
## - gender:math      1 1745.4
##
## Step:  AIC=1740.01
## absence ~ school + gender + math + language + school:gender +
##     gender:math
##
##                 Df    AIC
## <none>            1740.0
## - language       1 1740.3
## - school:gender  1 1743.0
## - gender:math    1 1743.7
```