PM LEVELS AND SCHOOL ABSENCES

IN MIDDLE SCHOOL CHILDREN

by

Jaclyn Marie Henderson

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Mathematical and Computer Sciences).

Golden, Colorado

Date _____

Signed: _____

Jaclyn Marie Henderson

Signed: _____

Dr. William Navidi

Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Dr. Willy Hereman

Professor and Head

Department of Applied Mathematics and Statistics

ACKNOWLEDGEMENTS

The purpose of our research was to determine the relationship between the middle school in Libby, Montana and the level of particulate air pollution. Libby, Montana, home to approximately 2600 residents, is located in northwestern Montana. The stoves located inside most of the homes in Libby are wood-burning stoves. To keep warm in the winter, the residents would burn a significant amount of wood which is not good for air quality. This can result in high levels of particulate matter air pollution. The particles in the pollution that have a diameter less than 10 microns can cause severe health effects. These microns, known as PM10, have been linked to increased absence rates in schools. During the years 2003-2005 data was collected of the average particulate matter and the number of school absences. In the years 2005-2008, over 1100 of the wood-burning stoves in the town were replaced with newer, cleaner-bur ning models. This resulted in a significant reduction in wintertime particulate levels. We fit generalized linear models assuming both a Poisson and binomial distribution for the outcome. We found that higher levels of particulate matter are associated with larger number of absences, at least on some days of the week and during some months. We also found that higher wind speed, larger wind gust, and higher relative humidity were associated with higher number of absences. Higher temperatures were associated with lower number of absences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Colorado School of Mines for supporting my education the past two years. I would also like to thank my advisor, Dr. William Navidi as well as my committee members Dr. Luis Tenorio and Dr. Amanda Hering. Thank you for your help and support throughout this entire process. I enjoyed learning from all three of you. I would next like to thank my parents for all of their support as well these last two years. Thank you for believing and encouraging me every day. I would not be here without the two of you. Thank you to all of my family, friends, and colleagues again your support means very much to me. I would also like to thank Dr. Laci Rogers Lyons and Professor Jamie Seitz. Thank you for the opportunity to work with you and to learn many of the skills I needed in graduate school. The information, professionally and personally, I learned from the two of you will forever be invaluable to me.

CHAPTER 1
INTRODUCTION

The purpose of our research is to try to determine the relationship between the middle school in Libby, Montana and the level of particulate air pollution. Libby, Montana, home to approximately 2600 residents, is located in northwestern Montana. The stoves located inside most of the homes in Libby are wood-burning stoves. To keep warm in the winter, the residents would burn a significant amount of wood which is not good for air quality. This can result in high levels of particulate matter air pollution. The particles in the pollution that have a diameter less than 10 microns can cause severe health affects. These microns, known as PM10, have been linked to increased absence rates in schools. During the years 2003-2005 data was collected of the average particulate matter and the number of school absences. In the years 2005-2008, over 1100 of the wood-burning stoves in the town were replaced with newer, cleaner-burning models. This resulted in a significant reduction in wintertime particulate levels, as presented in Table 1.

Table 1: Annual Average Particulate Matter Level for Winter Months (November - February)

Year	Average PML
2003-2004	27.85238
2004-2005	27.91250
2005-2006	26.57971
2006-2007	20.97297
2007-2008	20.59211
2008-2009	18.59167

PM10 has been associated with increased absence rates in schools ([1], [2], [4]). In this study, we will examine the relationship between school absence rates and PM levels in Libby. We have daily PM measurements for November 2005 through March 2009. We also have complete records of all absences at Libby Middle School, the only middle school in Libby, serving grades 5 through 8.

The first part of our research deals with analyzing the data year by year and then all the years combined. We are looking to investigate the relationship between the number of absences and the average particle matter level. We will determine whether the expected number of absences decreased when particulate matter levels are lower. We will also see if the number of absences declined over time and whether that is related to a decline in particulate matter level.

CHAPTER 2

DATA IN 2008-2009 SCHOOL YEAR

We began our research with the data for the 2008-2009 school year. There are 9,794 absences with 464 students for this school year of 180 days. Each of these absences has 25 variables associated with it. The variables are student ID, date, gender, grade, illness code, asthmatic, school day number, particle matter level, particle matter minimum, particle matter maximum, solar radiation, wind speed, wind direction, wind gust, temperature average, temperature maximum, temperature minimum, barometic pressure, precipitation, day of the week, and student number. Some of the variables that were most important to the research are date, day of the week, particle matter concentration, and absence length. The average particle matter level variable is the concentration of particle matter in the air for that particular day in ug/m^3. The absence length variable is the number of days a student was absent from school; starting with the first school day that was missed and ending when the student attended school again. Figure1 presents a boxplot of daily absences for the 2008-2009 school year. One can see that there are several extreme values. One in particular almost reaches 140 absences. On the days with unusually large number of absences, factors other than PM are likely to be the cause. Examples of such factors include snow days, days near the end of school or near holidays, or the first day of hunting season. We therefore removed these outliers by using the inter-quartile range (IQR) test. The IQR test takes the difference between the third and first quartiles and multiplies this difference by 1.5. The difference is subtracted from the first quartile and added to the third quartile. The data outside of this range are considered outliers. The low numbers that fall below Q1 - 1.5(IQR) are considered outliers as well as the high numbers that are above Q3 + 1.5(IQR). Once these outliers were removed, there were 168 days for the 2008-2009 school year.

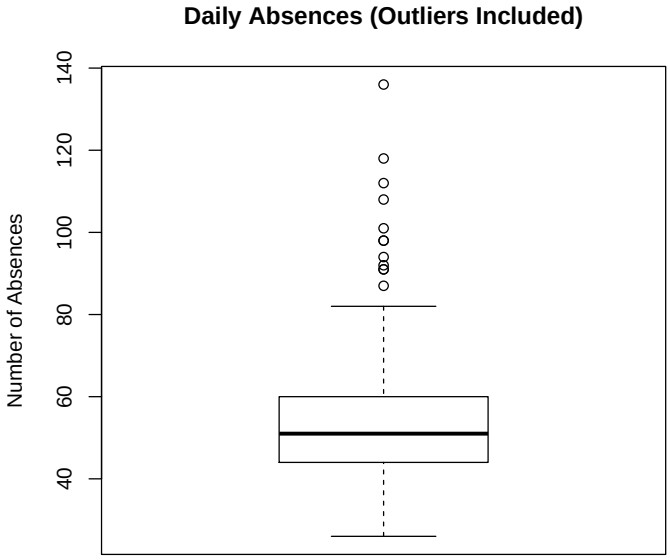

Figure 1: A boxplot of the daily number of absences for the 2008-2009 school year with the outliers included.

Figure 2 presents a boxplot of the daily absences with the outliers removed. The distribution is approximately

symmetric with two outliers. The two new outliers in Figure 2 are to be expected having adjusted for the previous outliers.

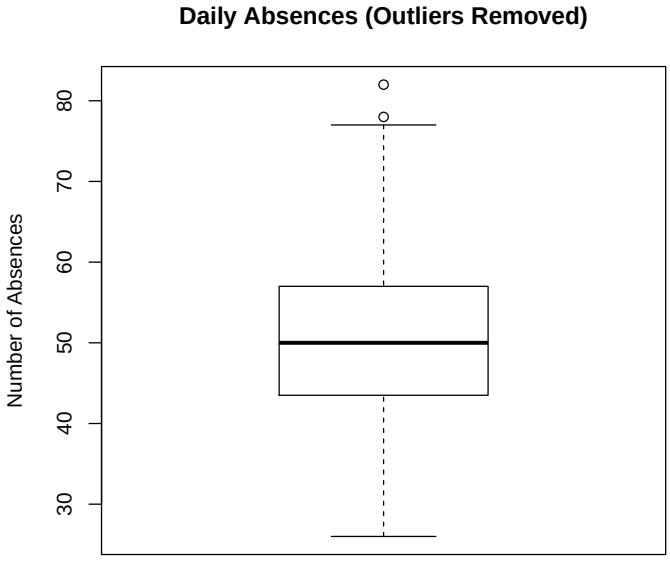

Figure 2: A boxplot of the daily number of absences for the 2008-2009 school year with the outliers excluded.

CHAPTER 3

TIME TRENDS IN ABSENCE RATES

To determine whether the number of absences varied over the school year, we plotted the number of absences against time. (The outliers were removed from this plot and the subsequent plots throughtout this paper.) Figure 3 presents the results.

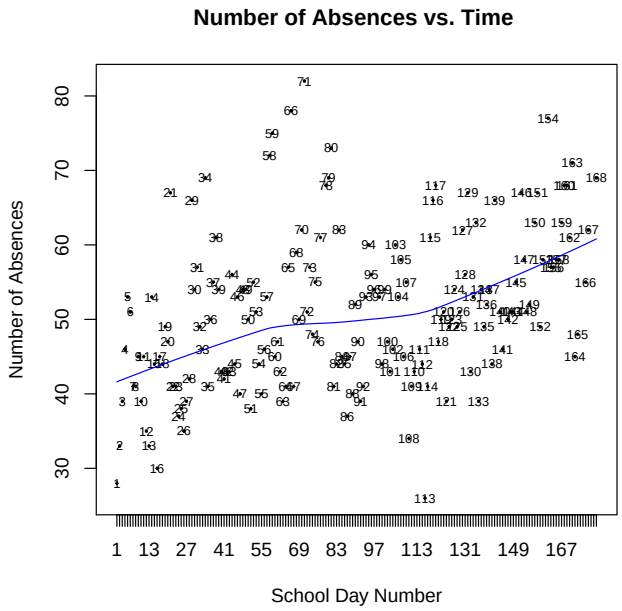

Figure 3: A scatterplot of each school day for the 2008-2009 school year versus the corresponding number of absences for that school day.

The blue line in Figure 3 is the lowess line. It is a line or curve fitted to the data using weighted least squares. It is computed using a locally-weighted scatterplot smoothing technique in which the data points that are centrally located receive more weight than points that are farther away which receive less weight. From the lowess line, there appears to be a clear increasing trend over the course of the year. To verify that the increasing trend seen in Figure 3 is statistically significant, we created a table that contains the school day number, the expected number of absences on each missed school day, the expected value (sample mean = 54.1105), and the chi-square value for each day. Table 2 is an abbreviated version of the frequency table. We calculated the chi-square value by first taking the absolute value of the difference between the actual number of missed school days and the expected value. We squared this number and divided it by the expected value. This calculation was computed for all 168 school days after removing outliers. The formula for the Chi-Square test is:

$$\chi^2 = \sum_{i=1}^{168} \frac{(O_i - E_i)^2}{E_i}$$

Table 2: Each school day is listed in chronological order with the number of absences, the expected value, and the chi-square statistic.

School Day	Number of Absences	Expected Value	Chi-Square Statistic
1	28	54.1105	12.59937
2	33	54.1105	8.23598
3	39	54.1105	4.21965
4	46	54.1105	1.21566
5	53	54.1105	0.02279
6	51	54.1105	0.17880
7	41	54.1105	3.17656
8	41	54.1105	3.17656
9	45	54.1105	1.53392
10	39	54.1105	4.21965
...	...	...	...
168	68	54.1105	3.56526

We performed a chi-square test to see if we would accept or reject the null hypothesis that the mean number of absences is the same on each day. The sum of column five, the chi-square values, is 367.8294. The chi-square value with a significane level of 5% and 167 degrees of freedom is 198.1542. Since our test statistic of 367.8294 exceeded the chi-square value of 198.1542, we can reject the null hypothesis and conclude that some days are missed more frequently than other days.

CHAPTER 4
CHI-SQUARE TESTS FOR SUBSETS

We expect that some of the variation in daily absences is due to varying days of the week or to varying months. To control for these sources of variation we repeated the chi-square tests, restricting the data to a certain day of the week or to a certain month. We split the data into smaller subsets and we repeated the chi-square test, using only absences on a given day of the week or during a given month. We expect that some of the variation between days is due to differences in days of the week or variations in the time of year. We tested to see if each Monday had the same mean number of absences. We repeated this process for each day of the week. We calculated each chi-square value in the same process as we did the total number of absences. We took the absolute value of the difference between the actual number of missed school days and the expected value. We squared this number and divided it by the expected value. This calculation was repeated for each missed day for a given day of the week. In Table 3, the chi-square value in column three is the sum of the chi-square values for each given day of the week. The p-value contained in the fourth column of Table 3 was calculated with the chi-square value and the corresponding degrees of freedom for that given weekday.

Table 3: Day of the week absences with the corresponding expected value, chi-square statistic, p-value, and degrees of freedom.

Day of the Week	Expected Value	Chi-Square Statistic	P-value	Degrees of Freedom
Monday	53.7647	57.7741	0.005	33
Tuesday	55.0270	110.6179	0.000	38
Wednesday	45.7222	1696.5530	0.000	36
Thursday	48.7188	48.1635	0.025	32
Friday	73.4828	219.4454	0.000	29

We also tested to see if each day within a given month had the same mean number of absences. This test was performed on months September through March. Table 4 shows the results of the rank test for each individual month.

Table 4: Monthly absences with the corresponding expected value, chi-square statistic, p-value, and degrees of freedom.

Month	Expected Value	Chi-Square Statistic	P-value	Degrees of Freedom
September	43.2381	37.2775	0.0000	20
October	47.7619	39.4417	0.0002	20
November	51.0588	32.9608	0.0054	16
December	53.3333	44.7625	0.0000	14
January	50.6842	28.8693	0.0007	18
February	49.7647	17.9858	0.0000	16
March	51.6364	5.5426	0.0031	21

There was considerable variation in the data even with taking into consideration month and day of the week. The range for the expected number of absences for weekdays was approximately twenty while the range for the expected number of absences for months was approximately ten. We will further examine the variation between week days and month with average PM level.

CHAPTER 5
RANK TEST

Figure 4 presents the number of absences versus the average particle matter level for the corresponding school day. There are a few outliers, which lead us to do a rank test on the data.

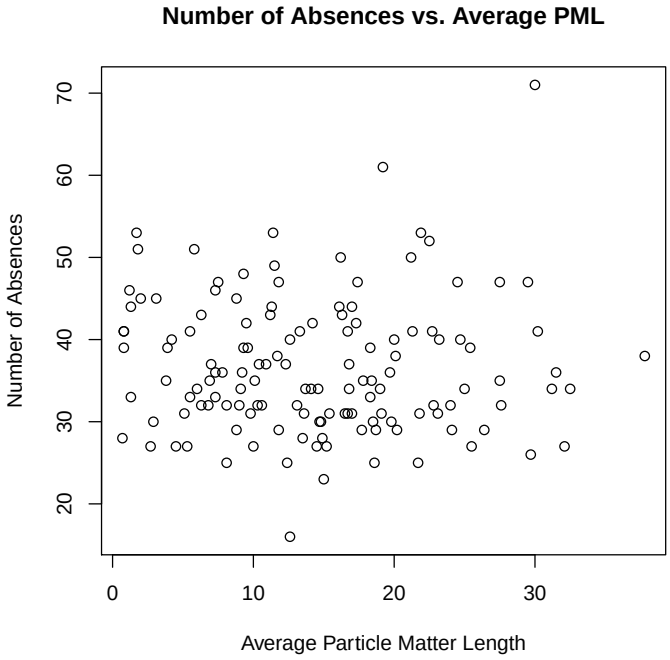

Figure 4: A scatterplot of the number of absences versus the average particle matter level for each school day for the 2008-2009 school year.

We decided to do a rank test to investigate the possibility of a relationship between average particulate pollution level and the expected number of absences. (Note for this test the outliers are still removed from the data because we expect that the days with unusually large numbers of absences are due to factors other than air pollution.) We computed the correlation between the rank of the average PML and the rank of the expected number of absences for the 2008-2009 school year. During this school year, pollution data was not recorded for the months of April, May, and June 2009. This further reduced the number of school days to 135. To do this we ordered the school days, 1 to 135, based on the average particle matter level. The smallest PML received a rank of 1 and the highest PML received a rank of 135. If there was a tie such as in school days 3, 7, and 8 when the average particle matter level was 0.8 for all three days, the ranks were averaged. The third largest PML was 0.8 ug/m^3 and subsequently these days received the same rank of 3.0. The school day with the next highest particle matter level received a rank of 5.0 and continued till all 135 days were ranked. The correlation between the rank of the average particle matter level and the rank of the expected number of absences is -0.07566395. The correlation is negative which suggests that days with more pollution have fewer absences. There is a weak linear relationship between the ranks of the two variables; however, the relationship is not statistically significant. Figure 5 presents the rank of the average particle matter level and the rank of the number

of absences.

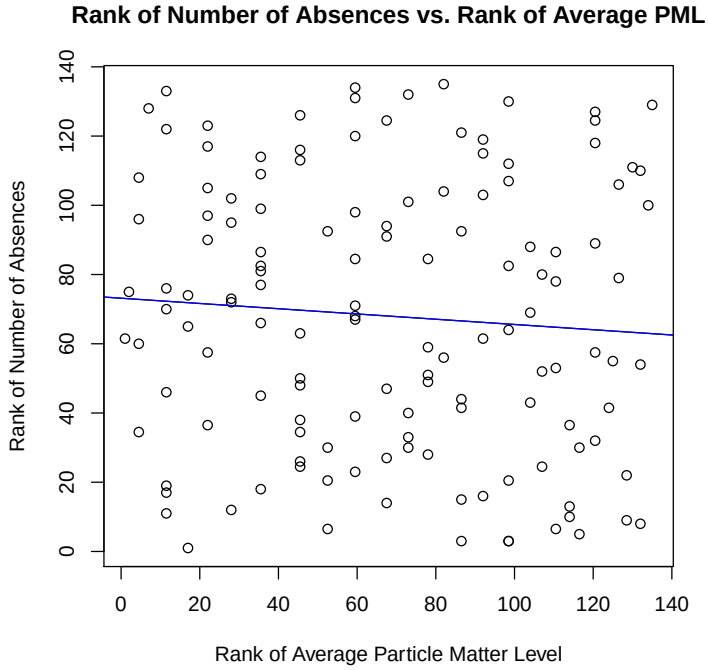

Figure 5: A scatterplot of the rank of the number of absences versus the rank of the average particle matter level for the 2008-2009 school year.

The line going through the graph is the least squares line. The β_0 term for the least squares line is 73.15107 with β_1 of -0.0759. From the least squares line, there is a slight negative linear trend of the ranks. We did a summary fit on the least squares line. The r-squared value is 0.005725 with a p-value of 0.3831. The p-value is greater than the alpha value of 0.05 and is not significant. We can conclude that there is no evidence of a relationship between the particulate pollution level and the expected number of absences.

Figure 6 presents the number of absences for Monday versus the average particle matter level for that particular day. From the figure there does appear to be a increasing linear trend, although we still have outliers. In particular a day with an average PML of around 5 ug/m^3 had over 65 absences. The existence of outliers, even after taking into consideration the day of the week, led us to do a rank test on each individual day of the week.

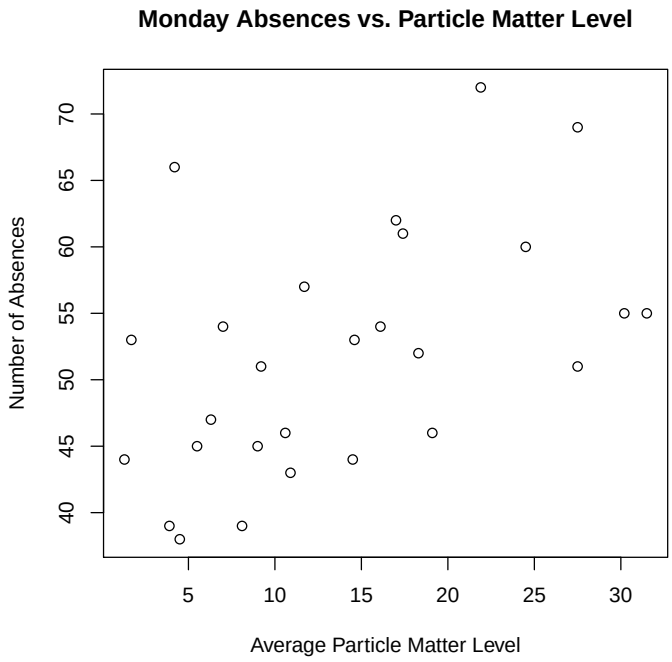

Figure 6: A scatterplot of the number of absences versus the average particle matter level for Monday for the 2008-2009 school year.

We performed the rank test on each given day of the week. The process for ranking the average particle matter level for each day of the week was the same process as when we ranked the average PML for each school day. Table 5 summarizes the results of the rank test performed on each weekday. The p-value for each weekday is compared against a 5% significance level. The weekdays that are significant are Monday and Thursday. We can conclude that there is evidence of a relationship between the particulate pollution level and the expected number of absences for Monday. We can also conclude there is evidence of a relationship between the particulate pollution level and the expected number of absences for Thursday. Table 5 contains each weekday with its correlation, r-squared value, and associated p-value before adjustments.

Table 5: A summary fit for the days of the week including: correlation, r-squared value, and p-value.

Day of the Week	Correlation	R-squared Value	P-value
Monday	0.5390	0.2905	0.0044
Tuesday	0.2577	0.0664	0.1468
Wednesday	0.2769	0.0767	0.2589
Thursday	0.4654	0.2166	0.0126
Friday	0.1308	0.0171	0.5242

We must take into consideration multiple tests were performed on the days of the week. Using the Bonferoni

adjustment, we will be able to tell the amount of impact the testing has had on the significance of Monday and Thursday. We performed five tests for each given day of the week. The original p-value for Monday is 0.0044.

Monday: P-value = 5(0.0044) = 0.022

With an adjusted p-value of 0.022, the day of the week Monday remains significant at the 5% level.

The original p-value for Thursday is 0.0126.

Thursday: P-value = 5(0.0126) = 0.0630

With an adjusted p-value of 0.0630, the day of the week Thursday is not significant at the 5% level.

Figure 7 presents the rank of the number of absences for Monday versus the rank of the average particle matter level. The line going through the scatterplot is the least squares line. β_0 is 11.560 and β_1 is 0.174. From the least squares line, there is a strong, positive linear trend of the ranks. We did a summary fit on the least squares line. The r-squared value is 0.2523 with a p-value of 0.0044.

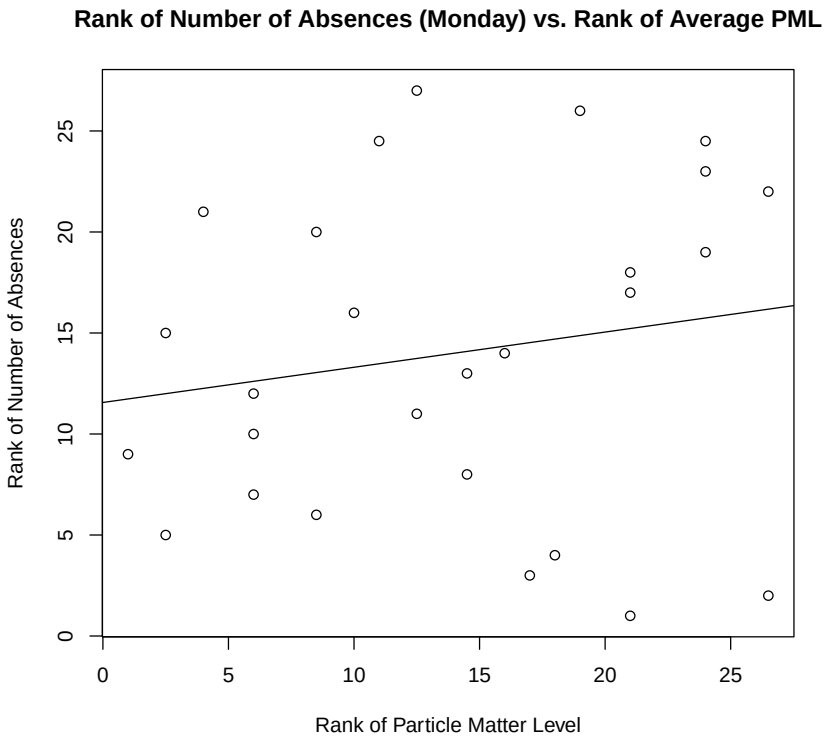

Figure 7: A scatterplot of the rank of the average particle matter level and the rank of the number of absences for Monday for the 2008-2009 school year.

Figure 8 presents the number of absences for October versus the average particle matter level for that particular day. From the figure there does appear to be a increasing linear trend, although we still have outliers. In particular, a day with an average PML of around 8 ug/m^3 had almost 70 absences. The existence of outliers, even after taking into consideration month, led us to do a rank test for the months.

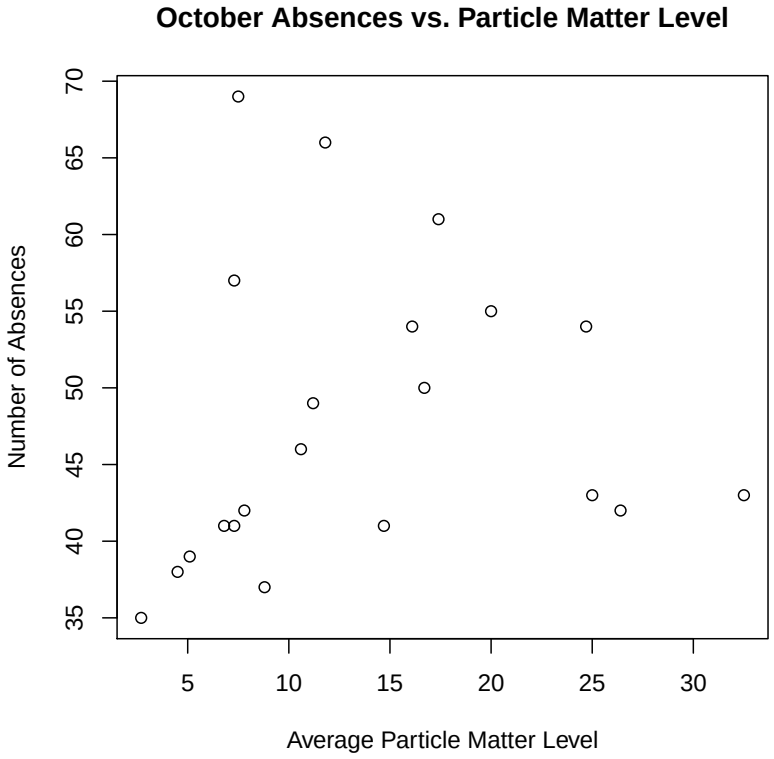

Figure 8: A scatterplot of the rank of the number of absences versus the rank of the average particle matter level for October for the 2008-2009 school year.

We performed the rank test on the individual months September through March. The process for ranking the particle matter level for each day of the month was ranked in the same process as the particle matter level for each school day was ranked. Table 5 summarizes the results of the rank test performed on each month. The p-value for each month is compared against a 5% significance level. The only month that is significant is October. We can conclude that there is evidence of a relationship between the particulate pollution level and the expected number of absences for October, although multiple tests were performed must be taken into account. The month of February is almost significant with a p-value of 0.0621. Using the Bonferoni adjustment we will be able to tell the impact the amount of testing has had on the significance of October and February. We performed seven tests on the months. The original p-value for October is 0.0466.

October: P-value = 7(0.0466) = 0.3262

With an adjusted p-value of 0.3262, October is no longer significant at the 5% level.

The original p-value for February was 0.0621.

February: P-value = 7(0.0621) = 0.4347

With an adjusted p-value of 0.4347, February is no longer significant at the 5% level. Table 6 consists of each month with its correlation, r-squared value, and associated p-value before adjustments.

Table 6: A summary fit for the months including: correlation, r-squared value, and p-value.

Month	Correlation	R-squared Value	P-value
September	0.3341	0.1116	0.1388
October	0.4388	0.1925	0.0466
November	-0.0025	0.0000	0.9925
December	0.0429	0.0018	0.8794
January	0.2500	0.0625	0.2878
February	0.4359	0.1900	0.0621
March	-0.3403	0.1158	0.1212

Figure 9 presents the rank of the expected number of absences for October versus the rank of the average particle matter level. The line going through the scatterplot is the least squares line. β_0 is 8.1589 and β_1 is 0.2583. From the graph, there is a strong, positive linear trend of the ranks. We did a summary fit on the least squares line. The r-squared value is 0.1501 with a p-value of 0.04659.

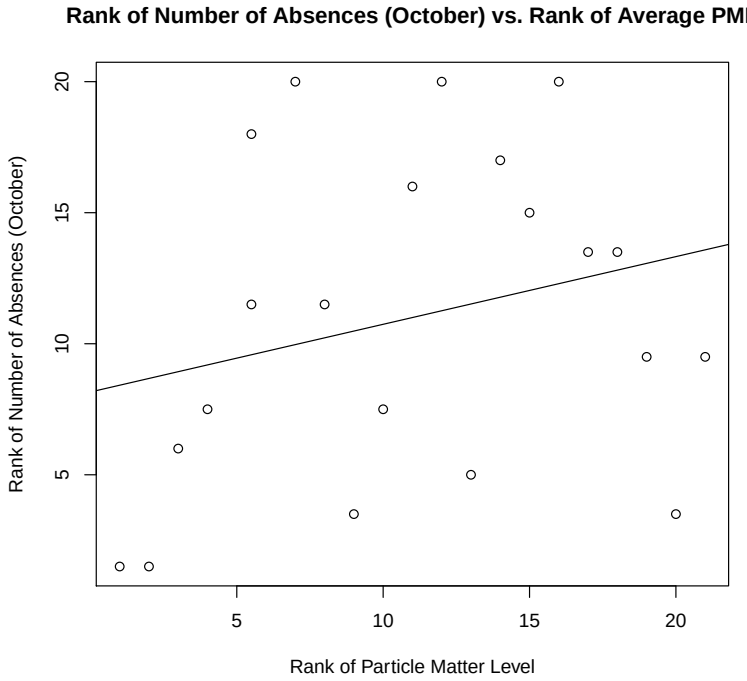

Figure 9: A scatterplot of the rank of the number of absences versus the rank of the average particle matter level for October for the 2008-2009 school year.

CHAPTER 6
COMBINING RESULTS FOR DAYS OF THE WEEK & FOR MONTHS

We performed a chi-square test on the data to see if a particular day of the week or month was significant. We performed a two-tail hypothesis test on each day of the week and on each month. The null hypothesis for a 2-tail test is $\rho = 0$. To find the test statistic for the two-tail hypothesis test for the weekdays, we took the p-values that we previously found from the rank test and took the log of 1 over the p-value. We summed these five values and multiplied by two. The test statistic is 26.9122. The formula we used is given below:

$$\chi^2 = 2\sum_{i=1}^{5} log\frac{1}{P_i}$$

We compared this test statistic to a chi-square value of 18.307 at 5% significance and ten degrees of freedom. Since the test statisitic is less than the chi-square value, we accept the null hypothesis that the rank correlation values are equal to zero. The p-value of this test is 0.0027. Table 7 presents the day of the week p-values for a two-tail hypothesis test.

Table 7: Statistics for a two-tail hypothesis test for each day of the week and the corresponding p-value.

Day of the Week	P-value	2*Log $(\frac{1}{P-value})$
Monday	0.00445	4.70328
Tuesday	0.21703	1.32696
Wednesday	0.000834	6.15767
Thursday	0.000231	7.27278
Friday	0.52417	0.56106
Test Statistic		20.02175

We performed the chi-square test on the months just as we had done for the days of the week. The null hypothesis for a 2-tail test is $\rho = 0$. To find the test statistic for the two-tail hypothesis test for the months, we took the p-values that we previously found from the rank test and took the log of 1 over the p-value. We summed these seven values and multiplied by two to get a test statistic of 22.62348. The formula we used is given below:

$$\chi^2 = 2\sum_{i=1}^{7} log\frac{1}{P_i}$$

We compared this test statistic to a chi-square value of 23.6848 with an alpha of 0.05 and fourteen degrees of freedom. Since the test statisitic is less than the chi-square value, we fail to reject the null hypothesis that the rank correlation values are equal to zero. The p-value for this test is 0.0666887. Table 8 presents the monthly p-values for a two-tail hypothesis test.

Table 8: Statistics for a two-tail hypothesis test for each month and the corresponding p-value.

Month	P-value	2*Log $(\frac{1}{P-value})$
September	0.1388	3.94944
October	0.0466	6.13231
November	0.9925	0.01506
December	0.8794	0.25703
January	0.2878	2.49099
February	0.0621	5.55802
March	0.1212	4.22063
Test Statistic		22.62348

In the tests restricted to a single day of the week or a single month, some of the days and months showed statistically significant results while others did not. Since multiple tests were performed, we cannot interpret the p-values in the same way as for single tests. In particular, it is more likely that one or more p-values may be small due to chance, resulting in a Type I error. We therefore combine all the results into a single test. The null hypothesis for the single test is that the correlation between the ranks is zero for each day of the week (or for each month). The test statistic is based on the fact that, when a null hypothesis is true the p-value has a uniform distribution on (0,1). Now assume we are conducting k tests, and let $P_1, ..., P_k$ be the p-values for each of these tests. Under the null hypothesis that the correlation between the ranks is zero for each day of the week (or for each month), $P_i \sim U(0,1)$. Therefore -log$P_i \sim$ Exp(1), and -2 log$P_i \sim$ Exp($\frac{1}{2}$), which is the same as χ^2_2. The sum, $\sum_{i=1}^{k}$ -2 logP_i, therefore has a χ^2_{2k} distribution.

CHAPTER 7
GLM FOR SIGNIFICANT DAYS & MONTHS

We did further testing on the days and months that were significant. For each day and month that was significant, we created a generalized linear model with each missed day and the corresponding average particle matter level. These models were computed with the outliers removed. The days of the week that are significant are Monday and Thursday. Let Y be the number of absences and let X be the particle matter level on a given day. We assumed that Y has a Poisson distribution with mean $e^{\beta_0+\beta_1 x}$. Since the number of students at risk is almost the same every day, we did not include their value in the model. The Monday model is:

$$\mathbf{E(Y)} = e^{3.813314+0.009548*x}$$

β_0 is 3.813314 and β_1, the term for PM, is 0.009548. For each increase of one in the average particle matter, the number of Monday absences increases by a factor of $e^{0.009548}$. Table 9 presents the statistics for the Monday model. β_1 is significant with a p-value of 0.00145.

Table 9: Statistics for the Monday model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, and the p-value for the coefficient.

Coefficient	Estimate	Standard Error	95% CI	P-value
Intercept	3.81331	0.051284	(3.7127, 3.9139)	0.000
PM	0.00954	0.002999	(0.0037, 0.0154)	0.001

Figure 10 presents the number of absences versus the average particle matter level with the model plotted. From the graph, the data has an upward trend which suggests that as the average PML increases then the number of absences on Mondays increases too. Also the days with the most absences are the ones where the average PML is neither high nor low.

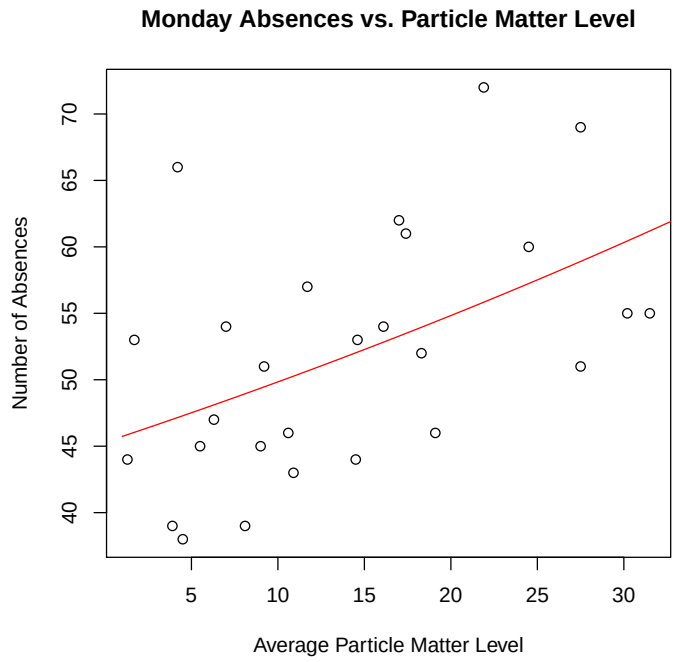

Figure 10: A scatterplot of the number of absences for Monday versus the average particle matter level for the 2008-2009 school year.

We also found Thursday to be a significant day of the week. The model for Thursday is:

$$\mathbf{E(Y)} = e^{3.766020+0.004747*x}$$

β_0 is 3.766020 and β_1, the term for PM, is 0.004747. For each increase of one in the average particle matter level, the number of Thursday absences increases by a factor of $e^{0.004747}$. Table 10 presents the statistics for the Thursday model. β_1 is not significant with a p-value of 0.21.

Table 10: Statistics for the Thursday model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, and the p-value for the coefficient.

Coefficient	Estimate	Standard Error	95% CI	P-value
Intercept	3.76602	0.061918	(3.6447, 3.8874)	0.000
PM	0.004747	0.003789	(-0.0027, 0.0122)	0.21

Figure 11 presents the number of absences versus the average particle matter with the model plotted. From the graph, the data has an upward trend which suggests that as the average PML increases then the number of absences on Thursday increases too.

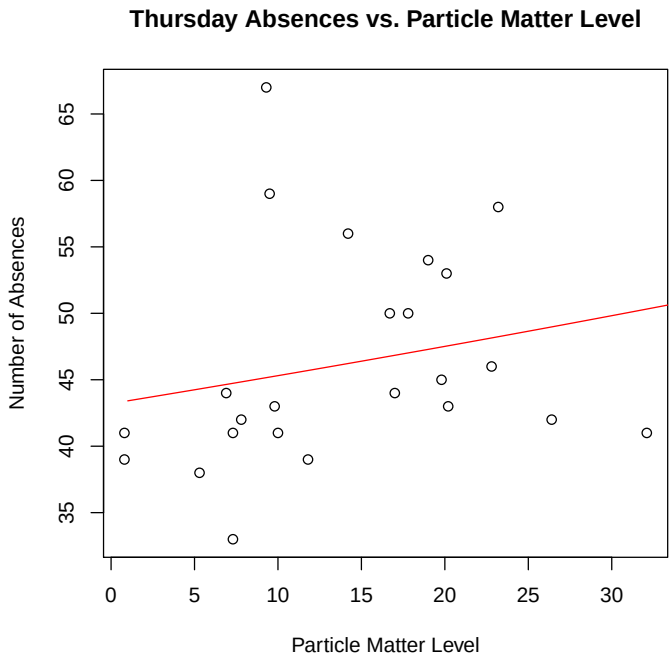

Figure 11: A scatterplot of the number of absences versus the average particle matter level for Thursday for the 2008-2009 school year.

We also did further testing on the months that are significant, October and February. We created the models for the months using the same process we did for the days of the week, with each missed day of October and February and the corresponding average particle matter level. These models were also computed with the outliers removed. The October model is:

$$\mathbf{E(Y)} = e^{3.820149+0.003369*x}$$

β_0 is 3.820149 and β_1, the term for PM, is 0.003369. For each increase of one in the average particle matter, the number of October absences increases by a factor $e^{0.003369}$. Table 11 presents the statistics for the October model. β_1 is not significant with a p-value of 0.386.

Table 11: Statistics for the October model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, and the p-value for the coefficient.

Coefficient	Estimate	Standard Error	95% CI	P-value
Intercept	3.82015	0.062208	(3.6981, 3.9421)	0.000
PM	0.003369	0.003888	(-0.0043, 0.0110)	0.386

Figure 12 presents the number of absences versus the average particle matter with the model plotted. From the

graph, the data seems to have an upward trend which suggests that as average particle matter levels increase then the number of absences during the month of October increases too.

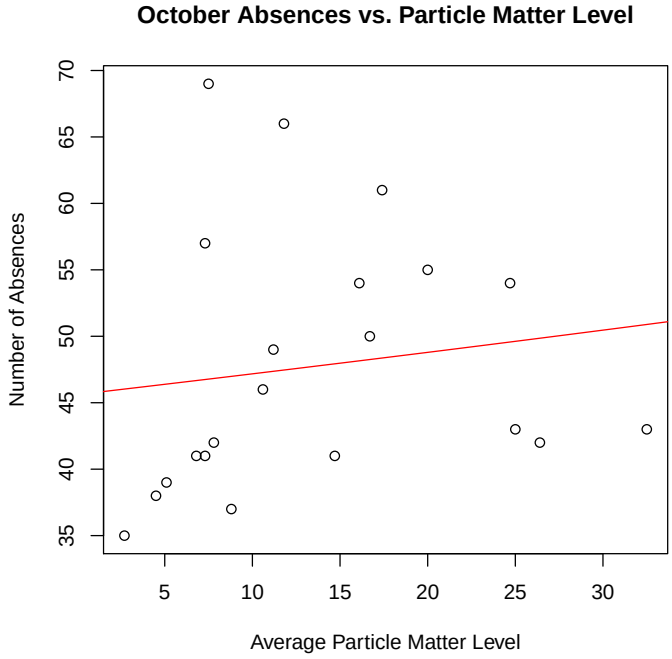

Figure 12: A scatterplot of the number of absences for October versus the average particle matter level for the 2008-2009 school year.

We also found the month of February to be significant. The February model is:

$$\mathbf{E(Y)} = e^{3.623808+0.013757*x}$$

β_0 is 3.623808 and β_1, the term for PM, is 0.013757. For each increase of one in the average particle matter, the number of February absences increases by a factor of $e^{0.013757}$. Table 12 presents the statistics for the February model. β_1 is significant with a p-value of 0.0265.

Table 12: Statistics for the February model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, and the p-value for the coefficient.

Coefficient	Estimate	Standard Error	95% CI	P-value
Intercept	3.623808	0.13353	(3.3621, 3.8855)	0.000
PM	0.013757	0.006198	(0.0016, 0.0259)	0.0265

Figure 13 presents the number of absences versus the average particle matter level with the model plotted. From the graph, the data seems to have an upward trend which seems to suggest that as PM levels increase then the number of

absences during the month of February increases too.

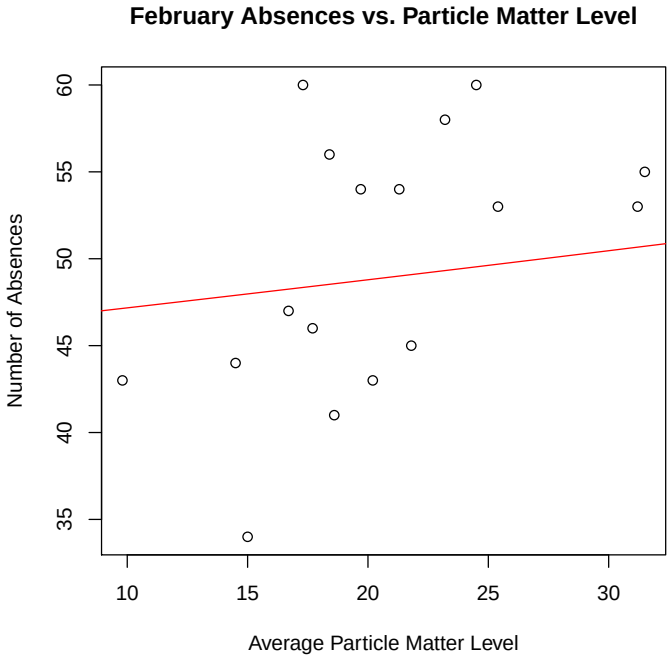

Figure 13: A scatterplot of the number of absences versus the average particle matter level for February for the 2008-2009 school year.

CHAPTER 8
ANALYSIS OF PARTICLE MATTER LEVEL ON ABSENCE LENGTH

We wanted to see if the particle matter level on the first day of absence had any effect on the length of the absence. We first looked at the entire dataset for the 2008-2009 school year. We took all the absences that had a length of one, a length of two, up to a length of twenty-five. For each length we averaged the particle matter levels on the first day of the absence. Table 13 presents the averages for the first three absence lengths.

Table 13: Average particle matter level for absence length.

Length	Average PML
1	15.21
2	25.66
3	15.08

We plotted the overall average particle matter level for an absence length of one to an absence length of twenty-five, this can be seen in Figure 14. The blue line in the figure is the lowess line. From the figure, the lowess line is relatively constant at approximately 15 ug/m^3 up till lengths of 15 or less. When a student is absent for 15 days or more, the average particle matter level decreases to below ten. We believe there are possible explanations besides average PML for why a student is absent for a period of 15 days or more from school. The resulting cause may not be due to air pollution but to a severe illness or a family emergency.

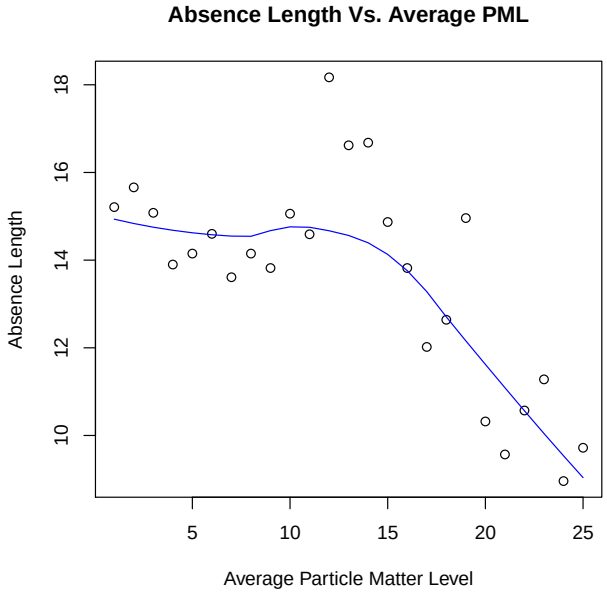

Figure 14: A scatterplot of the average particle matter level for absences of lengths 1-25 for the 2008-2009 school year.

We next considered the effect of average particle matter level on absence length for absences that began on each day of the week. We wanted to see if there was a weekly effect on absence length due to average PML. We repeated the same process that we had for the absence lengths for the entire school year. We took all the absences of length one up to a length ten for each day of the week. We then averaged the particle matters on the first day of absence for each length. Table 14 presents the day of the week with lengths one to three and the corresponding average particle matter level. We also wanted to see if the mean particle matter level was different for the different lengths of absence for each day of the week. To test if the difference between means was significant, we did an ANOVA test. In particular, we performed an F-test on the different lengths of absence for each day of the week. The null hypothesis was H_0: $\mu_1 = \mu_2 = \mu_3$, where μ_1 is the mean particle matter level for absences of length one, μ_2 is the mean particle matter level for absences of length two, and μ_3 is the mean particle matter level for absences of length three. The formula we used to compute the F-statistics is:

$$F = \frac{\dfrac{\sum_{i=1}^{3} n_i(\bar{x}_i - \bar{x})^2}{I-1}}{\dfrac{\sum_{i=1}^{3}(n_i-1)s_i^2}{N-1}}$$

where I is the number of samples taken, N is the total number of observations, n is the sample size, $\bar{x}_i$ is the sample mean, s_i is the standard deviation for each sample, and $\bar{x}$ is the population mean. The test statistics are also presented in Table 14. We compared each test statistic against a critical F-value with a 5% significance level. The general formula for the critical value is $F_{I-1,N-I}$, again where I is the number of samples taken and N is the total number of observations.

Table 14: Average particle matter level for the first day of an absence for each day of the week and the corresponding absence length. The F-statistic from the ANOVA test is also listed as well as the p-value.

Absence Length	Monday	Tuesday	Wednesday	Thursday	Friday
1	14.52	13.56	15.78	16.85	15.86
2	15.510	13.01	18.18	15.83	16.35
3	15.58	12.71	15.74	17.04	14.00
F-Statistic	1.48906	1.52207	5.21469	2.62077	2.97230
P-value	0.38705	0.38701	0.38710	0.38706	0.3870464

The day of the week that had a siginificant difference betweens means was Wednesday. The test statistic, 5.21469, was compared against a critical value of $F_{2,1142}$, or 3.00361. We can reject our null hypothesis and conclude that the mean particle matter value is different for absences of lengths one, two, and three for absences that begin on Wednesday. From Table 14, Wednesday, Thursday, and Friday have the highest particle matter level for each length. For absence lengths of one, two, and three, Wednesday, Thursday, and Friday have higher particle matter levels for

each length than Monday or Tuesday.

We also considered the effect of average particle matter level on absence length for each month because we wanted to see if there was a monthly or seasonal trend in absence length due to average PML. We repeated the same process we did for the absence lengths for the entire school year and days of the week. For each month we took all the absences up to a length of three. We averaged the particle matter levels for each length. Table 15 presents each month with lengths one to three and the corresponding particle matter level. There are differences in the mean PM levels among the lengths. We also wanted to see if the mean particle matter level was different for the different lengths of absence for each month. We performed an F-test on the different lengths of absence for each month. The null hypothesis was H_0: $\mu_1 = \mu_2 = \mu_3$, where μ_1 is the mean particle matter level for absences of length one, μ_2 is the mean particle matter level for absences of length two, and μ_3 is the mean particle matter level for absences of length three. The formula we used to compute the F-statistics is:

$$F = \frac{\frac{\sum_{i=1}^{7} n_i(\bar{x}_i - \bar{\bar{x}})^2}{I-1}}{\frac{\sum_{i=1}^{7}(n_i-1)s_i^2}{N-1}}$$

The variables in the model are the same as in the formula for the days of the week as stated previously. The test statistics are also presented in Table 15. We compared each test statistic against a critical F-value with a 5% significance level. The general formula for the critical value is $F_{I-1,N-I}$, again where I is the number of samples taken and N is the total number of observations.

Table 15: Average particle matter level for the first day of an absence for each month and the corresponding absence length. The F-statistic from the ANOVA test is also listed as well as the p-value.

Absence Length	September	October	November	December	January	February	March
1	4.048	13.36	17.61	12.73	22.99	21.06	12.0
2	4.074	14.13	17.96	12.74	25.08	19.99	11.75
3	3.581	16.41	16.37	13.41	24.73	21.41	10.61
F-Statistic	2.75310	5.00723	1.49795	1.46996	5.59127	1.51373	1.14971
P-value	0.38725	0.38725	0.38723	0.38727	0.38713	0.38712	0.38710

The months that had a significant difference betweens means were October and January. The test statistic for October, 5.00723, was compared against a critical value of $F_{2,801}$, or 3.00696, and the test statistic for January, 5.59127, was compared against a critical value of $F_{2,894}$, or 3.00579. We can reject our null hypothesis and conclude that the mean particle matter value is different for absences of lengths one, two, and three for absences that occur in

the months October or January. To determine whether average PML was associated with the length of absence, we performed an analysis of variance test (ANOVA). We tested the days of the week and months to see if the average particle matter level on the first day of absence was the same for different absence lengths. We wanted to see if the particle matter level was the same for absence lengths of one, two, and three days for each weekday and for each month. We also tested to see if the average particle matter level was the same for absence lengths of one, two, or three days for each weekday and for each month. To test the averages we performed ANOVA tests. We did the ANOVA testing because we wanted to see if there was a difference in the averages for weekdays and months for the two different groups of time. We are looking to further explain why there are increased absences during certain weekdays and months and whether this is due to increased average PML. We also wanted to see if there was a seasonal trend in the average particle matter level.

To create the models for the ANOVA testing we took the absences that are of one day in length and the corresponding particle matter level for that day, the absences that are of two days in length and the corresponding particle matter level for the first day of the absence, and the absences that are of three days in length and the correspoding particle matter level for the first day of the absence. The models were formed with the particle matter level as the response and the length of absence as the predictors.

Our null hypothesis for the 3-day models is H_0: $\mu_1 = \mu_2 = \mu_3$, where μ_1 is the mean for absences with a length of 1, μ_2 is the mean for absences with a length of 2, and μ_3 is the mean for absences with a length of 3. The 3-day model for September has a p-value of 0.4987 which makes this model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences in the month of September. Table 16 presents the summary statistics for the 3-day absence model for September and table 17 presents the ANOVA statistics for the model.

Table 16: Statistics for the September 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-Statistic	P-value
Intercept	4.04820	0.12725	31.814	0.000
2 days	0.02611	0.30737	0.085	0.932
3 days	-0.46716	0.40412	-1.156	0.248

Table 17: ANOVA statistics for the September 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	11.9	5.9425	0.6964	0.4987
Residuals	691	5896.4	8.5331	–	–

The 3-day model for October has a p-value of 0.0070 which makes this model significant.

The 3-day October model is:

$$\mathbf{E(Y) = 13.3606 + 0.7673}X_2 \mathbf{+ 3.0467}X_3$$

The expected value of Y in the model is the expected value of the average particle matter level for the first three lengths of absences for the month of October with absences of length two and absences of length three as predictors. X_2 is an indicator variable for absences of length two and X_3 is an indicator variable for absences of length three. For an absence of length two the average particle matter level increases by 0.7673 and for an absence of length three the average particle matter level increases by 3.0467. β_2 has a p-value of 0.2796 and does not appear to be significant. It has a 95% confidence interval of (-1.1104, 2.1573). β_3 has a p-value of 0.0021 with a 95% confidence interval of (1.1157, 4.9776). Since the model has a p-value of 0.0070, which makes it significant, we reject the null hypothesis and conclude the mean is the not the same for the first three lengths of absences in the month of October. Table 18 presents the summary statistics for the 3-day absence model for October and table 19 presents the ANOVA statistics for the model.

Table 18: Statistics for the October 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-Statistic	P-value
Intercept	13.361	0.317	42.155	0.000
2 days	0.767	0.709	1.082	0.280
3 days	3.047	0.985	3.093	0.002

Table 19: ANOVA statistics for the October 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	590	295.211	4.9895	0.007022
Residuals	801	47393	59.167	–	–

The 3-day model for November has a p-value of 0.2698 which makes the model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences in the month of November. Table 20 presents the summary statistics for the 3-day absence model for November and table 21 presents the ANOVA statistics for the model.

Table 20: Statistics for the November 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	17.6070	0.2657	66.275	0.000
2 days	0.3554	0.5914	0.601	0.548
3 days	-1.2372	0.8781	-1.409	0.159

Table 21: ANOVA statistics for the November 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	97.5	48.727	1.3125	0.2698
Residuals	709	26321.3	37.125	–	–

The 3-day model for December has a p-value of 0.2206 which makes this model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences in the month of December. Table 22 presents the summary statistics for the 3-day absence model for December and table 23 presents the ANOVA statistics for the model.

Table 22: Statistics for the December 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	12.73034	0.11900	106.975	0.000
2 days	0.01073	0.27841	0.039	0.969
3 days	0.68366	0.39505	1.731	0.084

Table 23: ANOVA statistics for the December 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	16.5	8.2691	1.1643	0.3128
Residuals	660	4687.6	7.1025	–	–

The 3-day model for January has a p-value of 0.0041 which makes the model significant.

The 3-day model for January is:

$$\mathbf{E(Y) = 22.9925 + 2.0831}X_2 \mathbf{+ 1.7348}X_3$$

The expected value of Y in the model is the expected value of the average particle matter for the first three lengths of absences for the month of January with absences of length two and absences of length three as predictors. X_2 is an indicator variable for absences of length two and X_3 is an indicator variable for absences of length three. For an absence of length two the particle matter level increases by 2.0831 and for an absence of length three the particle matter level increases by 1.7348. X_2 has a p-value of 0.0025 with a 95% confidence interval of (0.7626, 3.4036). X_3 has a p-value of 0.1167 and does not appear to be significant. It has a 95% confidence interval of (-0.3927, 3.8624). Since the model has a p-value of 0.0041, which makes it significant, we reject the null hypothesis and conclude the mean is the not the same for the first three lengths of absences in the month of January. Table 24 presents the summary statistics for the 3-day absence model for January and table 25 presents the ANOVA statistics for the model.

Table 24: Statistics for the January 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	22.9925	0.2973	77.327	0.000
2 days	2.0831	0.6737	3.092	0.00205
3 days	1.7348	1.0855	1.598	0.11036

Table 25: ANOVA statistics for the January 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	664	332.12	5.5406	0.00406
Residuals	894	53589	59.94	–	–

The 3-day model for February has a p-value of 0.0284 which makes the model significant.

The 3-day February model is:

$$\mathbf{E(Y) = 21.0598 - 1.0738}X_2 \mathbf{+ 0.3465}X_3$$

The expected value of Y in the model is the expected value of the average particle matter level for the first three lengths of absences for the month of February with absences of length two and absences of length three as predictors. X_2 is an indicator variable for absences of length two and X_3 is an indicator variable for absences of length three. For an absence of length two the average particle matter level decreases by 1.0738 and for an absence of length three the average particle matter level increases by 0.3465. β_2 has a p-value of 0.0117 with a 95% confidence interval of (-1.9066, -0.2410). β_3 has a p-value of 0.5938 and is not significant. Since the model has a p-value of 0.0284 we reject the null hypothesis and conclude the mean is the not the same for the first three lengths of absences in the month of February. From the model, as the expected number of absences of length two increase, the particle matter level decreases. Table 26 presents the summary statistics for the 3-day absence model for February and table 27 presents the ANOVA statistics for the model.

Table 26: Statistics for the February 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	21.0598	0.1906	110.503	0.000
2 days	-1.0738	0.4249	-2.527	0.0117
3 days	0.3465	0.6494	0.534	0.5938

Table 27: ANOVA statistics for the February 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	176.3	88.172	3.5752	0.0284
Residuals	911	22467.1	24.662	–	–

The 3-day model for March has a p-value of 0.021 which makes the model significant.

The 3-day model for March is:

$$\mathbf{E(Y) = 11.9982 - 0.2472}X_2 \mathbf{- 1.3917}X_3$$

The expected value of Y in the model is the expected value of the average particle matter level for the first three lengths of absences for the month of March with absences of length two and absences of length three as predictors. X_2 is an indicator variable for absences of length two and X_3 is an indicator variable for absences of length three. For an absence of length two the average particle matter level decreases by 0.2472 and for an absence of length three the average particle matter level decreases by 1.3917. β_2 has a p-value of 0.4636 with a 95% confidence interval of (-0.6607, 0.4135). β_3 has a p-value of 0.0058 with a 95% confidence interval of (-2.3782,-0.4052). Since the model has a p-value of 0.021, we reject the null hypothesis and conclude the mean is the not the same for the first three lengths of absences in the month of March. From the model, as the expected number of absences of length two increase, the particle matter level decreases. Table 28 presents the summary statistics for the 3-day absence model for March and table 29 presents the ANOVA statistics for the model.

Table 28: Statistics for the March 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-Statistic	P-value
Intercept	11.9982	0.1572	76.306	0.000
2 days	-0.2472	0.3371	-0.733	0.4636
3 days	-1.3917	0.5033	-2.765	0.0058

Table 29: ANOVA statistics for the March 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	136.5	68.270	3.8782	0.021
Residuals	984	17321.7	17.603	–	–

We performed a chi-square test on the combined months. Using the p-values that we found for each month, we took the log of 1 over the p-value. We added the seven values and multiplied the sum by two to get a test statistic of 18.29063. We compared this test statistic to a chi-square value of 23.6848 with a significance level of 5% and fourteen degrees of freedom. The p-value for this test is 0.19386. Month is not significant when taking into consideration its effect on average particle matter level and length of absence. The formula we used is given below:

$$\chi^2 = 2 \sum_{i=Sept}^{March} log(\frac{1}{p_i})$$

Again we must take into consideration the fact that multiple tests were performed. Using the Bonferoni adjustment, we will be able to tell the amount of impact the testing has had on the significance of the months. We performed seven tests on the months. Table 30 presents the p-value for each month and the adjusted p-value using the Bonferoni adjustment. After adjustments, the months October and January are the months that remain significant at the 5% significance level.

Table 30: The p-value and the adjusted p-value for each month.

Month	P-value	Adjusted P-value
September	0.4987	3.4909
October	0.0070	0.0492
November	0.2698	1.8886
December	0.3128	2.1896
January	0.0016	0.0284
February	0.0284	0.1988
March	0.0210	0.1470

We repeated this process for the ANOVA tests on days of the week but instead of taking the first three lengths of absences for each month, we did it separately for each day of the week. Our null hypothesis for the 3-day models is H_0: $\mu_1 = \mu_2 = \mu_3$. Where μ_1 is the mean for absences with a length of one, μ_2 is the mean for absences with a length of two, and μ_3 is the mean for absences with a length of three. The 3-day model for Monday has a p-value of 0.2198 which makes this model not significant. We accept the null hypothesis that the mean may be the same for the first three lengths of absences. Table 31 presents the summary statistics for the 3-day absence model for Monday and table 32 presents the ANOVA statistics for the model.

Table 31: Statistics for the Monday 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	14.5160	0.3105	46.746	0.000
2 days	0.9918	0.6467	1.534	0.125
3 days	1.0660	1.0248	1.040	0.298

Table 32: ANOVA statistics for the Monday 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	236	118.022	1.4908	0.2256
Residuals	1147	90807	79.169	–	–

The 3-day model for Tuesday has a p-value of 0.3966 which makes this model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences. Table 33 presents the summary statistics for the 3-day absence model for Tuesday and table 34 presents the ANOVA statistics for the model.

Table 33: Statistics for the Tuesday 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	13.5650	0.2213	61.292	0.000
2 days	-0.5590	0.5455	-1.025	0.306
3 days	-0.8540	0.8564	-0.997	0.319

Table 34: ANOVA statistics for the Tuesday 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-statistic	P-value
Group	2	96	47.894	0.9587	0.3837
Residuals	1291	64497	49.959	–	–

The 3-day model for Wednesday has a p-value of 0.00555 which makes the model significant.

The 3-day model for Wednesday is:

$$\mathbf{E(Y) = 15.7754 + 2.4056}X_2 \mathbf{- 0.0317}X_3$$

The expected value of Y in the model is the expected value of the average particle matter level for the first three lengths of absences for Wednesday with absences of length two and absences of length three as predictors. X_2 is an indicator variable for absences of length two and X_3 is an indicator variable for absences of length three. For an absence of length two the average particle matter level increases by 2.4056 and for an absence of length three the average particle matter level decreases 0.0317. β_2 has a p-value of 0.0016 with a 95% confidence interval of (0.91724, 3.8947). β_3 has a p-value of 0.9694 and does not appear significant. It has a 95% confidence interval of (-1.6428, 1.6110). Since the model has a p-value of 0.00555, we reject the null hypothesis and conclude the mean is the not the same for the first three lengths of absences for Wednesday. Table 35 presents the summary statistics for the 3-day absence model for Wednesday and table 36 presents the ANOVA statistics for the model.

Table 35: Statistics for the Wednesday 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	15.77541	0.34473	45.762	0.000
2 days	2.40598	0.75956	3.168	0.00159
3 days	-0.03171	0.83814	-0.038	0.96983

Table 36: ANOVA statistics for the Wednesday 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	823	411.44	5.222	0.00555
Residuals	967	76190	78.79	–	–

The 3-day model for Thursday has a p-value of 0.1937 which makes the model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences Thursday. Table 37 presents the summary statistics for the 3-day absence model for Thursday and table 38 presents the ANOVA statistics for the model.

Table 37: Statistics for the Thursday 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	16.8526	0.3021	55.780	0.000
2 days	-1.0258	0.5817	-1.763	0.0781
3 days	0.1884	1.1124	0.169	0.8656

Table 38: ANOVA statistics for the Thursday 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	230	114.95	1.6441	0.1937
Residuals	1107	77401	69.92	–	–

The 3-day model for Friday has a p-value of 0.06928 which makes the model not significant. We accept the null hypothesis and conclude the mean may be the same for the first three lengths of absences for Friday. Table 39 presents the summary statistics for the 3-day absence model for Friday and table 40 presents the ANOVA statistics for the model.

Table 39: Statistics for the Friday 3-day absence model including: the coefficient and its estimate, the standard error of the estimate, a 95% confidence interval for the estimate, the t-statistic for the coefficient, and the p-value for the coefficient.

Coefficients	Estimate	Standard Error	T-statistic	P-value
Intercept	15.861	0.235	67.458	0.000
2 days	0.490	0.671	0.730	0.466
3 days	-1.857	0.877	-2.117	0.035

Table 40: ANOVA statistics for the Friday 3-day absence model including: the degrees of freedom, the sum of squares, the mean sum of squares, the f-statistic, and the p-value. The degrees of freedom, sum of squares, and mean sum of squares are also given for the residuals.

Source	DF	SS	MS	F-Statistic	P-value
Group	2	279	139.493	2.676	0.06928
Residuals	1145	59691	52.132	–	–

We performed a chi-square test on the combined days of the week. Using the p-values that we found for each day of the week, we took the log of 1 over the p-value. We summed these five values and multiplied the sum by two to get a test statistic of 10.38128. We compared this test statistic to a chi-square value of 18.30704 with an alpha of 0.05 and ten degrees of freedom. The p-value for this test is 0.40770. Week day is not significant when taking into consideration its effect on average particle matter level and length of absence. The formula we used is given below:

$$\chi^2 = 2 \sum_{i=Mon}^{Fri} log(\frac{1}{p_i})$$

Again we must take into consideration multiple tests were performed on the days of the week. Using the Bonferoni adjustment, we will be able to tell the amount of impact the testing has had on the significance of the days. We performed seven tests on the days of the week. Table 41 presents the p-value for each day of the week and the adjusted p-value using the Bonferoni adjustment. With an adjusted p-value, Wednesday is the only day of the week that remains significant at the 5% significance level.

Table 41: The p-value and the adjusted p-value for each day of the week.

Day of the Week	P-value	Adjusted P-value
Monday	0.2256	1.128
Tuesday	0.3837	1.9185
Wednesday	0.00555	0.02775
Thursday	0.1937	0.9685
Friday	0.06928	0.3464

CHAPTER 9
GENERALIZED LINEAR MODELS

We fit the generalized linear models to determine which of several potential confounders were significant. The potential confounders are week day, wind speed, wind gust, average relative humidity, average temperature, month, and year. The first model was a generalized linear Poisson model. We started with the data in November of 2005, when the pollution data was recorded regularly, and used the data through March of 2009. In addition to average particle matter level, potential confounders included in this model were grade, day of the week, wind speed, wind gust, average relative humidity, average temperature, and month. Let Y be the expected number of absences on a given day. β_0 is the intercept, X_1 is an indicator variable for grade, X_2 is an indicator variable for the average particle matter level, X_3 is an indicator variable for week day, X_4 is an indicator variable for wind speed, X_5 is an indicator variable for wind gust, X_6 is an indicator variable for average relative humidity, X_7 is an indicator variable for average temperature, X_8 is an indicator variable for the month, and X_9 is an indicator variable for year. The generalized linear Poisson model is often parameterized as follows:

$$\text{E(Y)} = \text{n}(\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9))$$

where n is the known number at risk on a particular day. Because the number at risk is essentially constant, we parameterized the models as follows:

$$\text{E(Y)} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9)$$

where e^{β_0} represents a baseline that is constant across days.

The average particle matter level, parameter β_2, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0029753 ± 0.00033065 or (0.0026446, 0.0033059). The p-values for the β_4, β_5, β_6, β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 42 presents the results for the average particle matter level parameter β_2.

Table 42: The effect of average particle matter level for the Poisson model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0029753	0.0001687	(0.0026446, 0.0033059)	0.000

The estimate for β_2 is 0.0029753, the expected number of absences is increased by a factor of $e^{0.0029753}$ for

each increase of one ug/m^3 of average PML. The value for β_4, the parameter for wind speed, is 0.091199. The expected number of absences is increased by a factor of $e^{0.091199}$ for each increase of one mph. The value for β_5, the parameter for wind gust is 0.0028531. The expected number of absences is increased by a factor of $e^{0.0028531}$ for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0009593. The expected number of absences is increased by a factor of $e^{0.0009593}$ for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0021724. The expected number of absences is decreased by a factor of $e^{-0.0021724}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed, wind gust, and average relative humidity while temperature decreased the expected value of absence. To see how well the model fit the data, we looked at several residual plots. For this particular model we looked at the deviance residuals. With a Poisson model, the deviance residuals are calculated as follows:

$$r_i = sign(y_i - \hat{\lambda}_i)[2(y_i log(\frac{y_i}{\hat{\lambda}_i}) - y_i - \hat{\lambda}_i)]^{\frac{1}{2}}$$

where y_i is the number of students absent on day i and $\hat{\lambda}_i$ is the estimated number of absences for day i. Figure 15 presents two plots of the predicted values and residuals. On the left the plot is the deviance residuals versus the predicted values. On the right the plot is the deviance residuals versus the predicted link values. There is not a significant difference between the two plots; the other residual plots yielded similar results. The models fit the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the probability of an absence.

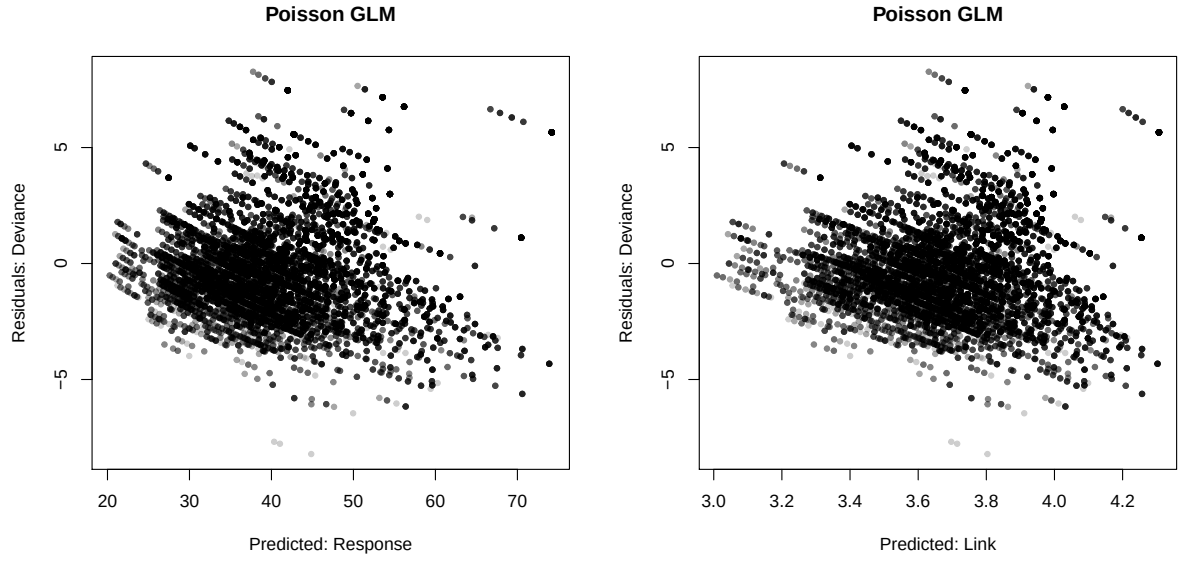

Figure 15: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

The second model was a generalized linear binomial model. The indicator variables are still assigned to the same

parameters as in the Poisson model accordingly. Below is the generalized linear binomial model:

$$\text{P(absence)} = \frac{\exp \sum_{i=1}^{p} \hat{\beta}_i X_i}{1 + \exp \sum_{i=1}^{p} \hat{\beta}_i X_i}$$

This is equivalent to:

$$\log \frac{P(absence)}{1 - P(absence)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9$$

In other words, the log odds of absence is linear in the parameters.

The average particle matter level, parameter β_2, was highly significant with a p-value of 0.000. A 95% confidence interval for parameter β_2 is 0.0030414 $\pm$ 0.00034476 or (0.0026966, 0.0033862). The p-values for parameters β_4, β_5, β_6, and β_7, corresponding to wind speed, wind gust, average relative humidity, and average temperature, were all significant with a p-values of 0.000. Table 43 presents the results for the average particle matter level parameter β_2.

Table 43: The effect of average particle matter level for the binomial model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0030414	0.0001759	(0.0026966, 0.0033862)	0.000

The estimate for β_2 is 0.0030414, the log odds of absence is increased by 0.0030414 for each increase of 1 ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.095450. The log odds of absence is increased by 0.095450 for each increase of one mph. Wind gust, parameter β_5, has an estimate of 0.0024186. The log odds of absences is increased by 0.0024186 for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0010010. The log odds of absence is increased by 0.0010010 for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0019279. The log odds of absence is decreased by -0.0019279 for each increase of one degree Farenheit. The confounders that increased the log odds of absence are wind speed, wind gust, and average relative humdity while temperature decreased the log odds of absence. To see how well the model fit the data, we looked at several residual plots. For this particular model we looked at the deviance residuals. With a binomial model, the deviance residuals are calculated as follows:

$$r_i = sign(y_i - n_i \hat{p}_i)[2[y_i log(\frac{y_i}{n_i \hat{p}_i}) + (n_i - y_i) log(\frac{n_i - y_i}{n_i(1 - \hat{p}_i)})]]^{\frac{1}{2}}$$

where y_i is the number of students absent on day i, $\hat{p}_i$ is the estimated probability of absence on day i, and n_i is the number of students at risk on day i.

Figure 16 presents two plots of the predicted values and residuals. On the left the plot is the predicted values versus the deviance residuals. On the right the plot is the predicted link values versus the deviance residuals. There is not a significant difference between the two plots; the other residual plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the log odds of an absence.

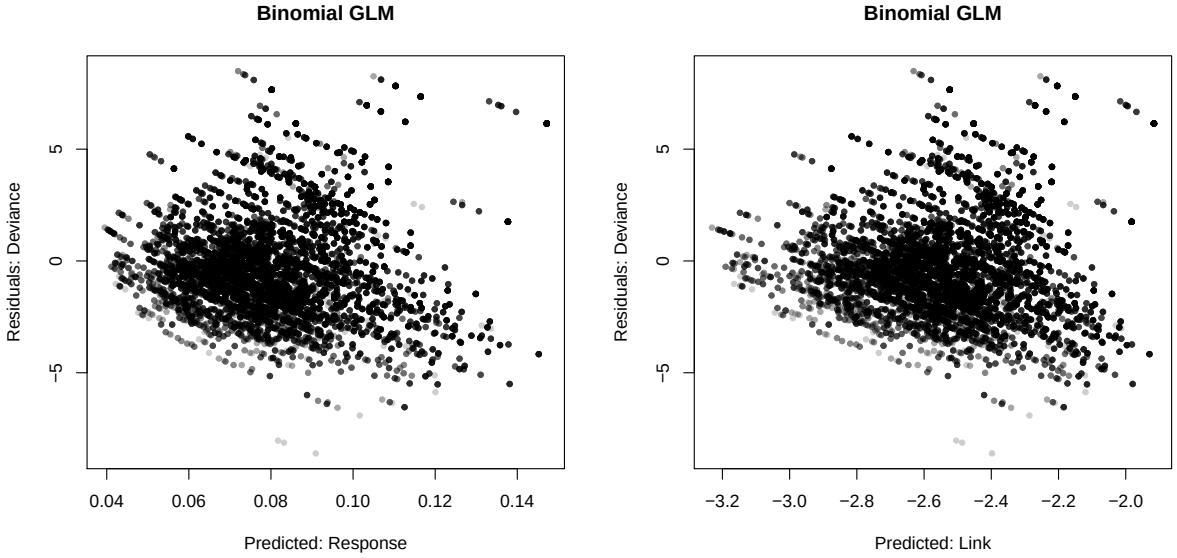

Figure 16: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

CHAPTER 10
GENERALIZED LINEAR MODELS FOR ANNUAL PM LEVELS

Table 44 presents the average particle matter level, the number of students, and the mean number of absences for each year. There has been a decline both in annual particle matter level and proportion of students absent since 2005-2006.

Table 44: Table of the average particle matter level along with the number of students, the mean number of absences, and the percent of students absent for each year.

Year	Year PML	Num. of Students	Mean Num. Absences	Percent of Students Absent
2005-2006	18.97510	482	38.51825	7.991
2006-2007	14.90359	541	36.80899	6.804
2007-2008	14.24304	509	35.22222	6.920
2008-2009	15.18453	464	32.86111	7.082

The next model we fit was a generalized linear Poisson model but the daily average particle matter level was replaced with the average particle matter level for each individual year. The indicator variables remained the same as they had in the previous Poisson glm accordingly, but now parameter β_2 is the year average particle matter level and parameter β_9 has been removed. β_9 was the parameter for year but since we am already accounting for year with the year average particle matter level then we do not need this parameter in the model. β_2, was highly significant with a p-value of 0.000. A 95% confidence interval for the X_2 parameter is 0.01263 ± 0.0011776 or (0.011452, 0.013808). The p-values for the β_4, β_5, β_6, and β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 45 shows the results for the year average particle matter level parameter β_2.

Table 45: The effect of the annual average particle matter level for the Poisson model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.01263	0.0006008	(0.011452, 0.013808)	0.000

The estimate for β_2 is 0.01263, the expected number of absences is increased by a factor of $e^{0.01263}$ for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.04802. The expected number of absences is increased by a factor of $e^{0.04802}$ for each increase of one mph. The value for β_5, the parameter for wind gust is 0.001669. The expected number of absences is increased by a factor of $e^{0.001669}$ for each increase of one mph. The value for β_6, the parameter for average relative humidity is -0.0003703. The expected number of absences is

decreased by a factor of $e^{-0.0003703}$ for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.002034. The expected number of absences is decreased by a factor of $e^{-0.002034}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed and wind gust while average relative humidity and temperature decreased the expected number of absences. To see how well the model fit the data, we looked at several residual plots. Figure 17 presents two plots of the fitted values and residuals. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the probability of an absence.

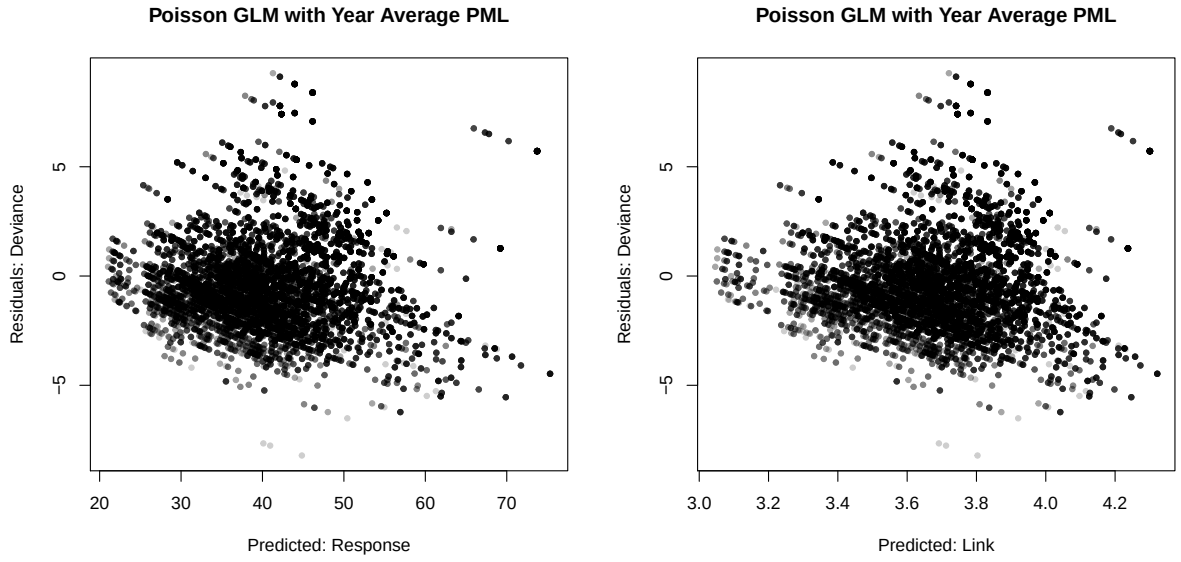

Figure 17: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

The next model was a generalized linear binomial model. The indicator variables are still assigned to the same parameters as in the previous binomial glm accordingly. The daily average particle matter level is replaced with the average particle matter level for each individual year. The parameter for year average particle matter level is β_2 and parameter β_9 has been removed. β_9 was the parameter for year but since we are already accounting for year with the year average particle matter level then we do not need this variable in the model. The parameter β_2 was highly significant with a p-value of 0.000. The p-values for the β_4, β_6, and β_7 parameters, corresponding to wind speed, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000 but β_5 was not significant with a p-value of 0.43940. I removed the β_5 parameter and re-tested the model. The parameter for year average particle matter level, β_2 , is significant with a p-value of 0.000. A 95% confidence interval for the β_2

parameter is 0.0316 ± 0.0012354 or (0.030365, 0.032835). Table 46 presents the results for the year average particle matter level parameter β_2.

Table 46: The effect of the annual average particle matter level for the binomial model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0316	0.0006303	(0.030365, 0.032835)	0.000

The estimate for β_2 is 0.0316, the log odds of absence is increased by 0.0316 for each increase of one ug/m^3 of PM. The parameters β_4, β_6, and β_7 remained significant with a p-value of 0.000. Wind speed, parameter X_4, increased the log odds of absences by 0.08306 for each increase of one mph. Average relative humidity, parameter β_5, increased the log odds of absence by 0.0004607 for each increase of one percent humidity. Average temperature, parameter β_6, decreased the log odds of absence by -0.003541 for each increase of one degree Farenheit. The confounders that increased the log odds of absence were wind speed and average relative humdity while average temperature decreased the log odds of absence. We looked at several residual plots to see how well the model fit the data. Figure 18 presents two plots of the fitted values and residuals. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the log odds of an absence.

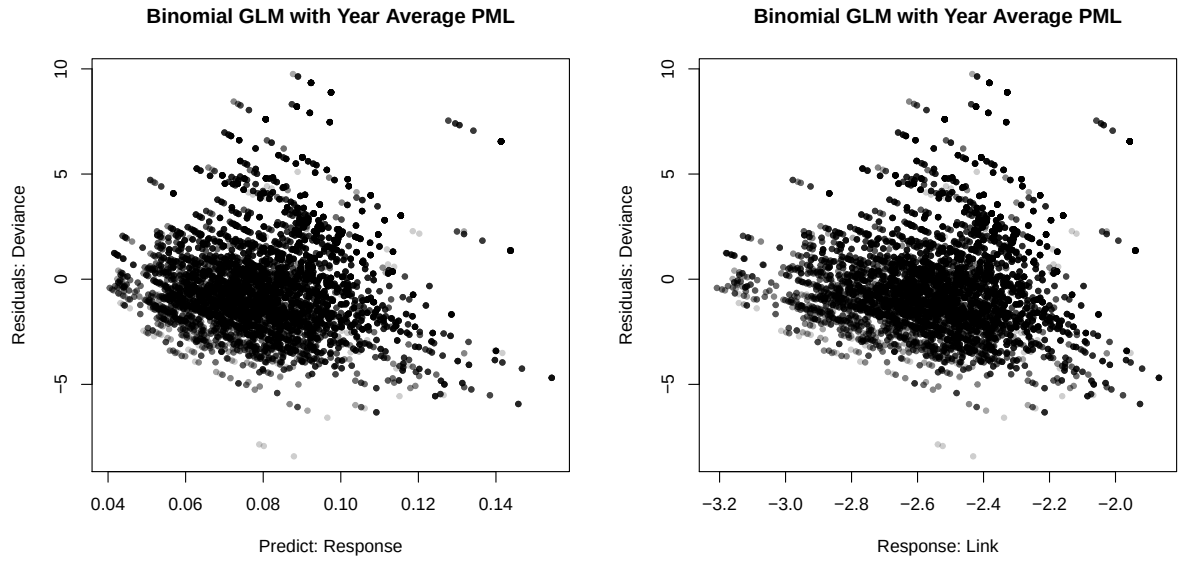

Figure 18: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

CHAPTER 11
GENERALIZED LINEAR MODELS FOR WINTER MONTHS

The next models we fitted were a subset of the previous models just fitted and tested. We decided to create a subset of the data for the winter months November through February. We wanted to fit a model for this part of the data by itself because the average particle matter level is higher during these months. In particular, wood stoves are used extensively for heating, so particle matter levels are much higher in the winter months. The generalized linear Poisson model for the winter months is the same formula as in the original generalized linear Poisson model. Again Y is the expected number of absences on a given day and β_2 is a parameter for the average particle matter level. β_2, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0025074 ± 0.00040925 or (0.0020982, 0.0029166). The p-values for the β_4, β_5, β_6, β_7 parameter, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 47 shows the results for the average particle matter level parameter β_2.

Table 47: The effect of average particle matter level for the Poisson model for the winter months November-February. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0025074	0.0002088	(0.0020982, 0.0029166)	0.000

The estimate for β_2 is 0.0025074, the expected number of absences is increased by a factor of $e^{0.0025074}$ for each increase of one ug/m^3 of PM. The estimate for β_4, the parameter for wind speed, is 0.15195. The expected number of absences is increased by a factor of $e^{0.15195}$ for each increase of one mph. The value for β_5, the parameter for wind gust is 0.012861. The expected number of absences is increased by a factor of $e^{0.012861}$ for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0067876. The expected number of absences is increased by a factor of $e^{0.0067876}$ for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0094626. The expected number of absences is decreased by a factor of $e^{-0.0094626}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed, wind gust, and average relative humidity while temperature decreased the expected number of absences. Figure 19 presents two residual plots. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. This model fits days with more than 50 absences better than the previous models. There is no downward trend of the data with days of absences greater than 50.

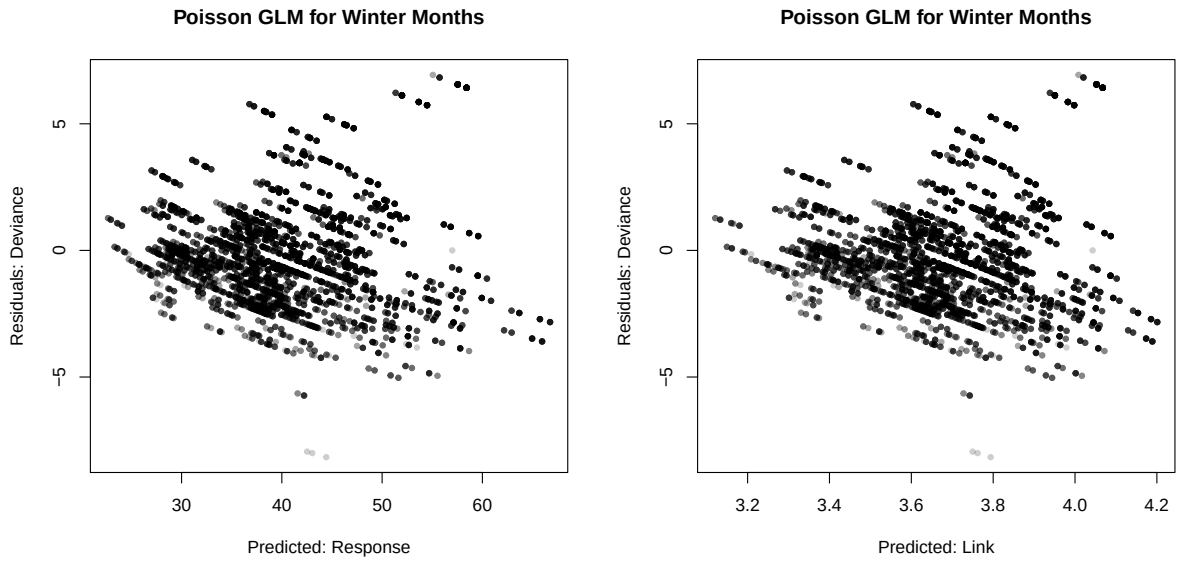

Figure 19: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

We also fitted a generalized linear binomial model to the winter months subset. This model is of the same form as the original generalized linear binomial model with the same indicator variables assigned to the parameters accordingly. β_2, the parameter for the average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0033758 ± 0.00042669 or (0.0029491, 0.0038025). The p-values for the β_4, β_5, β_6, β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 48 shows the results for the average particle matter level parameter β_2.

Table 48: The effect of average particle matter level for the binomial model for the winter months November-February. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0033758	0.0002177	(0.0029491, 0.0038025)	0.000

The estimate for β_2 for is 0.0033758, the log odds of absence is increased by 0.0033758 for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.14824. The logg odds of absence is increased by 0.14824 for each increase of one mph. The value for β_5, the parameter for wind gust is 0.013780. The log odds of absence is increased by 0.013780 for each increase of one mph. The value for X_6, parameter for average relative humidity is 0.0070176. The log odds of absence is increased by 0.0070176 for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0093993. The log odds of absence is decreased by

-0.0093993 for each increase of one degree Farenheit. The confounders that increased the log odds of absence are wind speed, wind gust, and average relative humidity while temperature decreased the log odds of absence. Figure 20 presents two residual plots. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. This model fits days with more than 50 absences better than the previous models. There is no downward trend of the data with days of absences greater than 50.

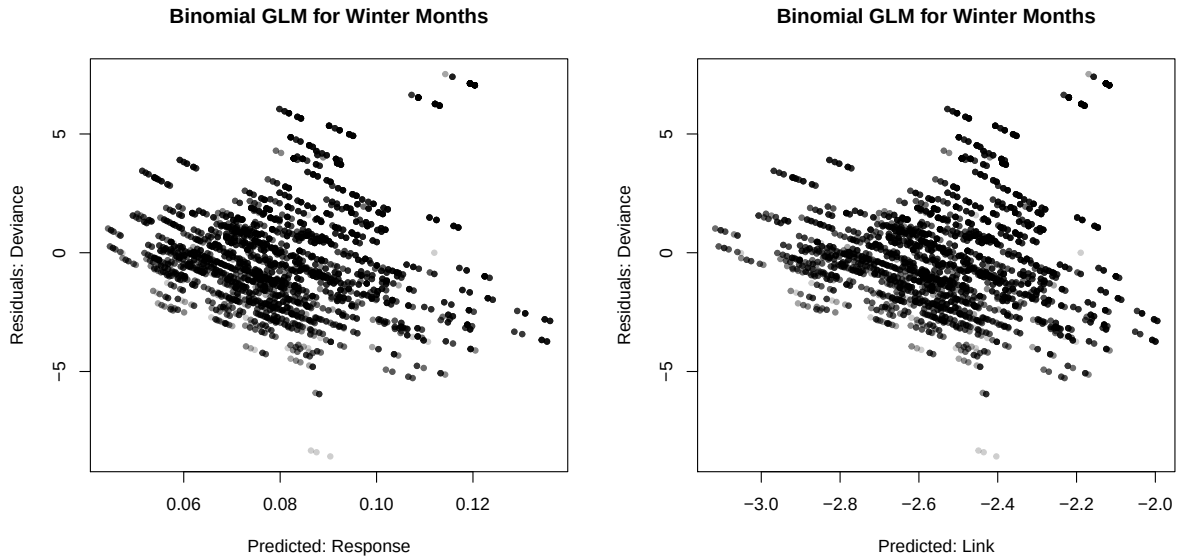

Figure 20: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

In addition to these winter models, we also fitted two winter models with the year average particle matter level in place of the daily average particle matter level. This model is of the same form as the generalized linear Poisson model but β_2 is now the year average particle matter level and the parameter for year, β_9 has been removed. Again we are already accounting for the year with the year average particle matter level. β_2, the parameter for the year average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.034254 ± 0.00175616 or (0.032498, 0.036010). The p-values for the β_4, β_5, β_6, β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 49 presents the results for the year average particle matter level parameter β_2.

Table 49: The effect of the annual average particle matter level for the Poisson model for the winter months November-February. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.034254	0.0008960	(0.032498, 0.036010)	0.000

The estimate for β_2 is 0.034254, the expected number of absences is increased by a factor of $e^{0.034254}$ for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.1076821. The expected number of absences is increased by a factor of $e^{0.1076821}$ for each increase of one mph. The value for β_5, the parameter for wind gust is 0.0108142. The expected number of absences is increased by a factor of $e^{0.0108142}$ for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0040438. The expected number of absences is increased by a factor of $e^{0.0040438}$ for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0083274. The expected number of absences is decreased by a factor of $e^{-0.0083274}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed, wind gust, and average relative humidity while temperature decreased the expected number of absences. Figure 21 presents two residual plots. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. This model fits days with more than 50 absences better than the previous winter models using the daily average particle matter level.

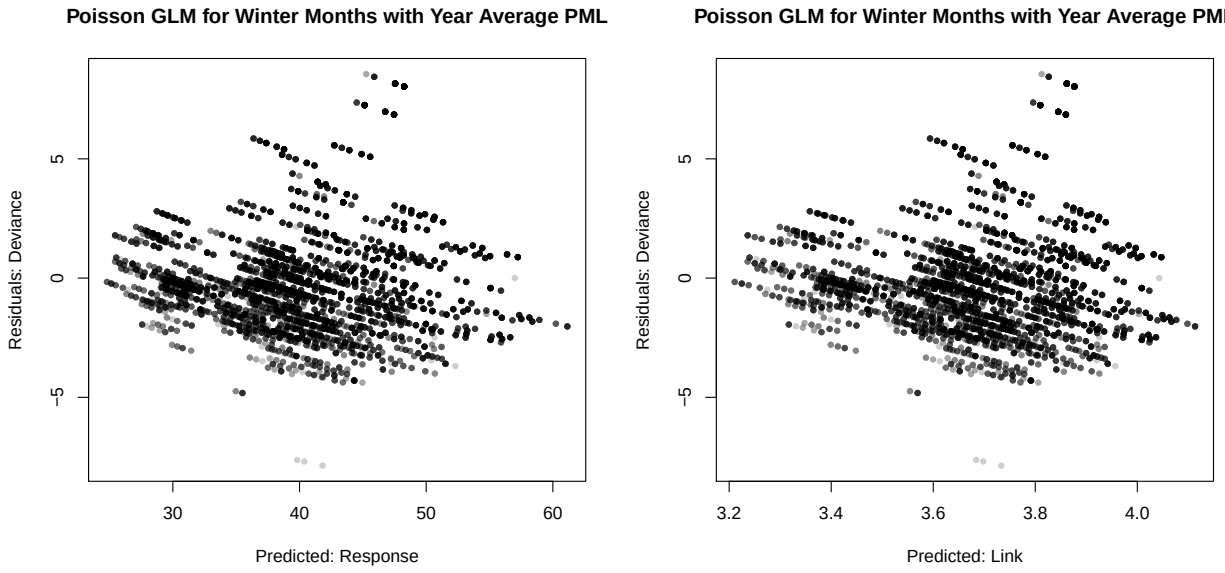

Figure 21: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

The generalized linear binomial model for winter months was also fitted with the year average particle matter level in place of the daily average particle matter level. This model is of the same form as the generalized linear binomial model but β_2 is now the year average particle matter level and the parameter for year, β_9 has been removed. β_2, the parameter for the year average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0532284 ± 0.0018350 or (0.051393, 0.055063). The p-values for the β_4, β_5, β_6, β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 50 presents the results for the year average particle matter level parameter β_2.

Table 50: The effect of the annual average particle matter level for the binomial model for the winter months November-February. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0532284	0.0009362	(0.051393, 0.055063)	0.000

The estimate for β_2 for is 0.0532284, the log odds of absence is increased by 0.0532284 for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.14647. The log odds of absence is increased by 0.14647 for each increase of one mph. The value for β_5, the parameter for wind gust is 0.0096688. The log odds of absence is increased by 0.0096688 for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0055806. The log odds of absence is increased by 0.0055806 for each increase of one percent humidity.

The value for β_7, the parameter for average temperature is -0.0105137. The log odds of absence is decreased by -0.0105137 for each increase of one degree Farenheit. The confounders that increased the log odds of absence are wind speed, wind gust, and average relative humidity while temperature decreased the log odds of absence. Figure 22 presents two residual plots. On the left the plot is the fitted values versus the deviance residuals. On the right the plot is the log of the fitted values versus the deviance residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. This model, as in the previous Poisson winter model using the year average particle matter level, fits the days with more than 50 absences better than the previous winter models using the daily average particle matter level.

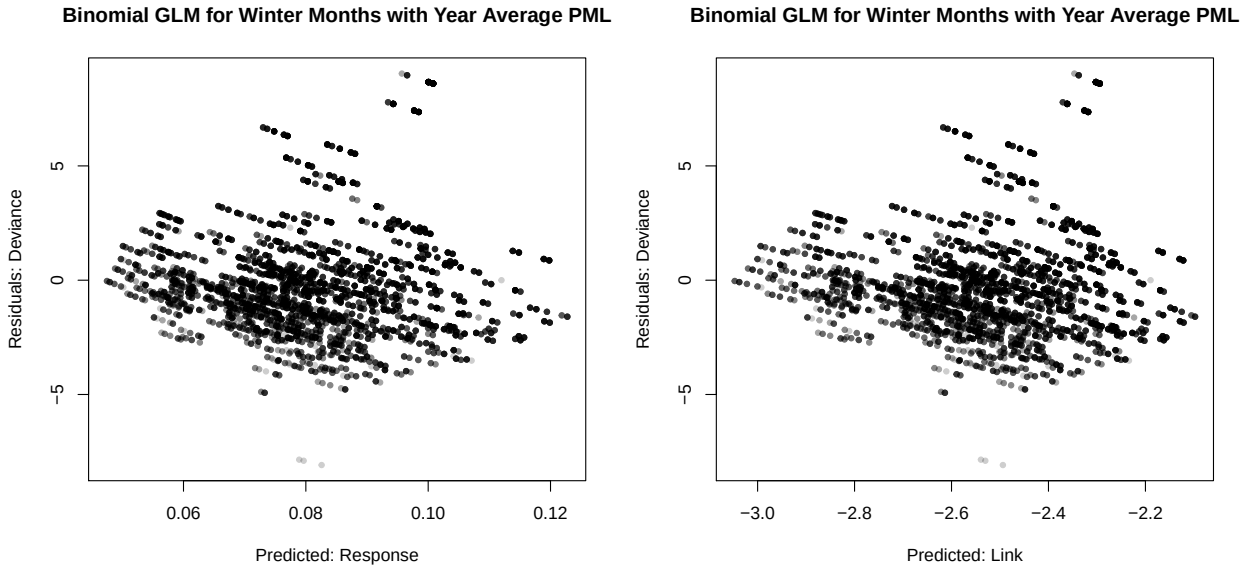

Figure 22: Two plots of the predicted values and residuals. The first plot is the deviance residuals versus the predicted values. The second plot is the deviance residuals versus the predicted link values.

CHAPTER 12
GENERALIZED LINEAR MODELS USING LAG 1 PM LEVEL

We also fitted two models with the previous day's particle matter level in place of the daily particle matter level. We fitted these models because we wanted to see if the previous day's particle matter level had an impact on attendance the following day. One of the models is of the same form as the generalized linear Poisson model but β_2 is now the previous day's average particle matter level. β_2, the parameter for the previous day's average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0043900 $\pm$ 0.00034986 or (0.0040401, 0.0047399). The p-values for the β_4, β_5, β_6, β_7 parameters, corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 51 presents the results for the previous day's average particle matter level parameter β_2.

Table 51: The effect of the lag 1 average particle matter level for the Poisson model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0043900	0.0001785	(0.0040401, 0.0047399)	0.000

The estimate for β_2 is 0.0043900, the expected number of absences is increased by a factor of $e^{0.0043900}$ for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.038040. The expected number of absences is increased by a factor of $e^{0.038040}$ for each increase of one mph. The value for β_5, the parameter for wind gust is 0.0098105. The expected number of absences is increased by a factor of $e^{0.0098105}$ for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0016947. The expected number of absences is increased by a factor of $e^{0.0016947}$ for each increase in one percent humidity. The value for β_7, the parameter for average temperature is -0.0038222. The expected number of absences is decreased by a factor of $e^{-0.0038222}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed, wind gust, and average relative humidity while temperature decreased the expected number of absences. As with the previous models, we tested the fit of the model to the data by looking at the residual plots. The following residual plots are with the Pearson residuals. For the Poisson models, the Pearson residual for day i is:

$$r_i = \frac{Y_i - \hat{\lambda}_i}{\hat{\lambda}_i^{\frac{1}{2}}}$$

The formula for the link response for the Pearson residual with a Poisson model is:

$$link : log(\hat{\lambda}_i)$$

where Y_i is the expected number of absences on day i and $\hat{\lambda}_i$ is the estimated mean number of absences on day i. Figure 23 presents two residual plots. On the left the plot is the predicted values versus the Pearson residuals. On the right

the plot is the log of the fitted values versus the Pearson residuals. There is not a significant difference between the two plots; the other residuals plots yielded similar results. The model fits the data well; most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the probability of an absence.

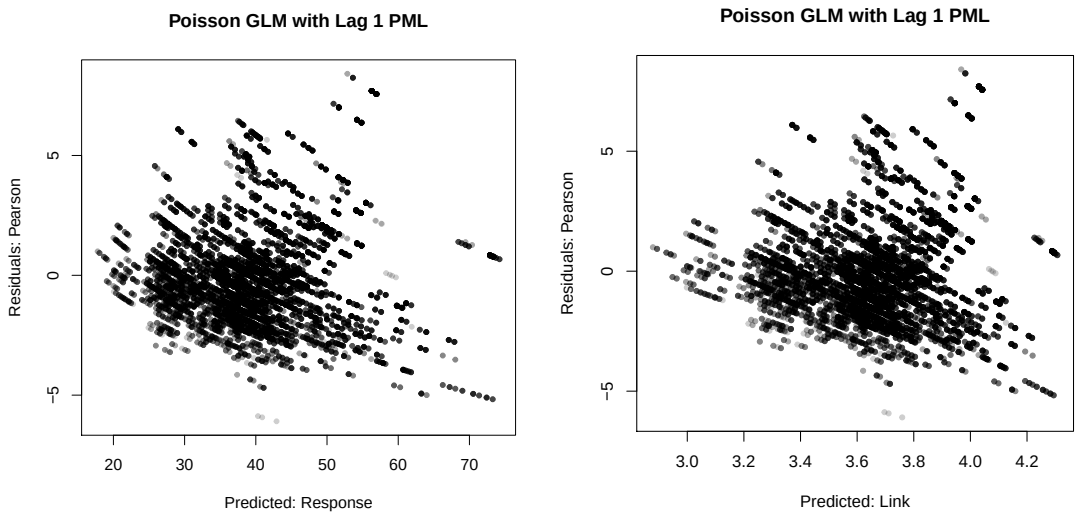

Figure 23: Two plots of the predicted values and residuals. The first plot is the Pearson residuals versus the predicted values. The second plot is the Pearson residuals versus the predicted link values.

The next model we fitted was a generalized linear binomial model using the previous day's average particle matter level. This model is of the same form as the generalized linear binomial model but β_2 is now the previous day's average particle matter level. β_2, the parameter for the previous day's average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for β_2 is 0.0050986 ± 0.00036417 or (0.0047344, 0.005462768). The p-values for β_4, β_5, β_6, and β_7 corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 52 presents the effects for the previous day's average particle matter level parameter β_2.

Table 52: The effect of the lag 1 average particle matter level for the binomial model. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0050986	0.0001858	(0.0047344, 0.0054628)	0.000

The estimate for β_2 is 0.0050986. The log odds of absence is increased by 0.0050986 for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.033634. The log odds of absence is increased by 0.033634 for each increase of one mph. The value for β_5, the variable for wind gust is 0.010800. The log odds of absence is increased by 0.010800 for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0018207. The log odds of absences is increased by 0.0018207 for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.0039110. The log odds of absence is decreased by -0.0039110 for each increase of one degree Farenheit. The confounders that increased the log odds of absence are wind speed, wind gust, and average relative humidity while temperature decreased the log odds of absence. Again we looked at several residual plots to see how well the model fit the data. The following residual plots are with the Pearson residuals. For the binomial models, the Pearson residual for day i is:

$$r_i = \frac{Y_i - n_i\hat{p}_i}{(n_i\hat{p}_i(1-\hat{p}_i))^{\frac{1}{2}}}$$

The formula for the link response for the Pearson residual with a binomial model is:

$$link : log(\frac{\hat{p}_i}{1-\hat{p}_i})$$

where Y_i is the expected number of absences on day i, n_i is the number of students at risk on day i, and p_i is the estimated probability of an absence on day i. The formula for p_i is:

$$\text{P(absence)} = \frac{\exp\sum_{i=0}^{p}\hat{\beta}_i X_i}{1+\exp\sum_{i=0}^{p}\hat{\beta}_i X_i}$$

Figure 24 presents two plots of the fitted values and residuals. On the left the plot is the predicted values versus the Pearson residuals. On the right the plot is the log of the predicted values versus the Pearson residuals. There is not a significant difference between the two plots. Most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the log odds of an absence.

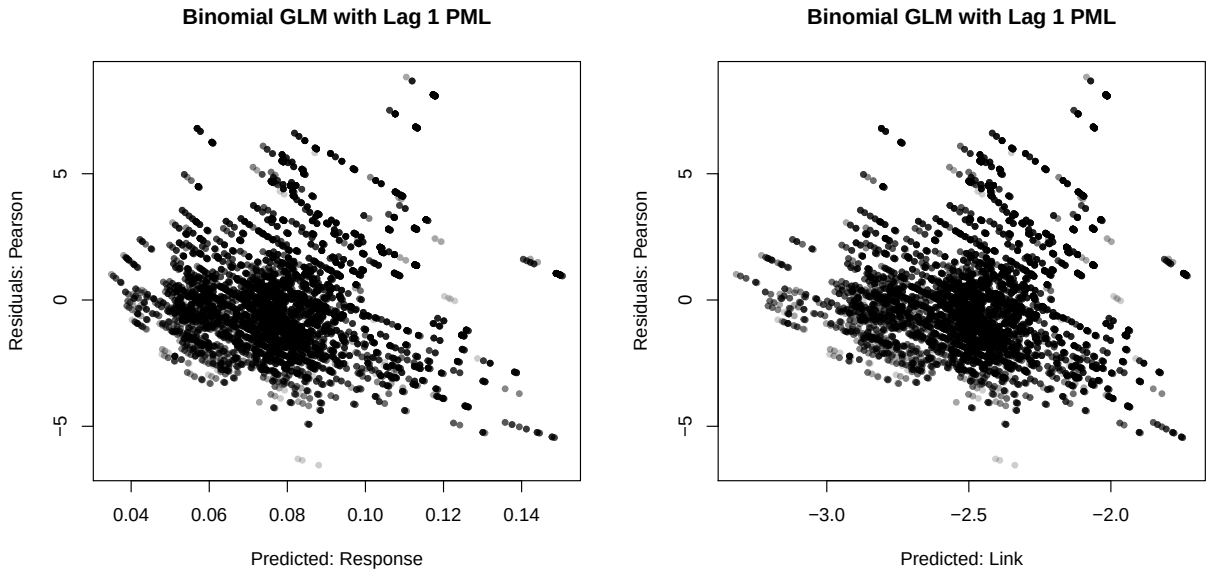

Figure 24: Two plots of the predicted values and residuals. The first plot is the Pearson residuals versus the predicted values. The second plot is the Pearson residuals versus the predicted link values.

The next set of models we fitted were generalized linear models during the winter months using the previous day's average particle matter level. The first model, a generalized linear Poisson model, is of the same form as the generalized linear Poisson model but β_2 is now the previous day's average particle matter level and the data has been restricted to the winter months November through February. β_2, the parameter for the previous day's average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for β_2 is 0.0040852 $\pm$ 0.000401996 or (0.0036832, 0.0044872). The p-values for parameters β_4, β_5, β_6, and β_7 corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 53 presents the results for the previous day's average particle matter level parameter β_2.

Table 53: The effect of the lag 1 average particle matter level for the Poisson model for the winter months November-February.. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0040852	0.0002051	(0.0036832, 0.0044872)	0.000

The estimate for β_2 for is 0.0040852, the expected number of absences is increased by a factor of $e^{0.0040852}$ for each increase of one ug/m^3 of PM. The value for β_4, the parameter for wind speed, is 0.13298. The expected number of absences is increased by a factor of $e^{0.13298}$ for each increase of one mph. The value for β_5, the variable for wind gust is 0.013497. The expected number of absences is increased by a factor of $e^{0.0134967}$ for each increase of one mph.

The value for $]beta_6$, the parameter for average relative humidity is 0.0076688. The expected number of absences is increased by a factor of $e^{0.0076688}$ for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.010825. The expected number of absences is decreased by a factor of $e^{-0.010825}$ for each increase of one degree Farenheit. The confounders that increased the expected number of absences are wind speed, wind gust, and average relative humidity while temperature decreased the expected number of absences. Again we looked at several residual plots to see how well the model fit the data. Figure 25 presents two plots of the fitted values and residuals. On the left the plot is the the predicted values versus the Pearson residuals. On the right the plot is the log of the predicted values versus the Pearson residuals. There is not a significant difference between the two plots. Most of the data is centered around 0 while varying between -6 and 6 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the probability of an absence.

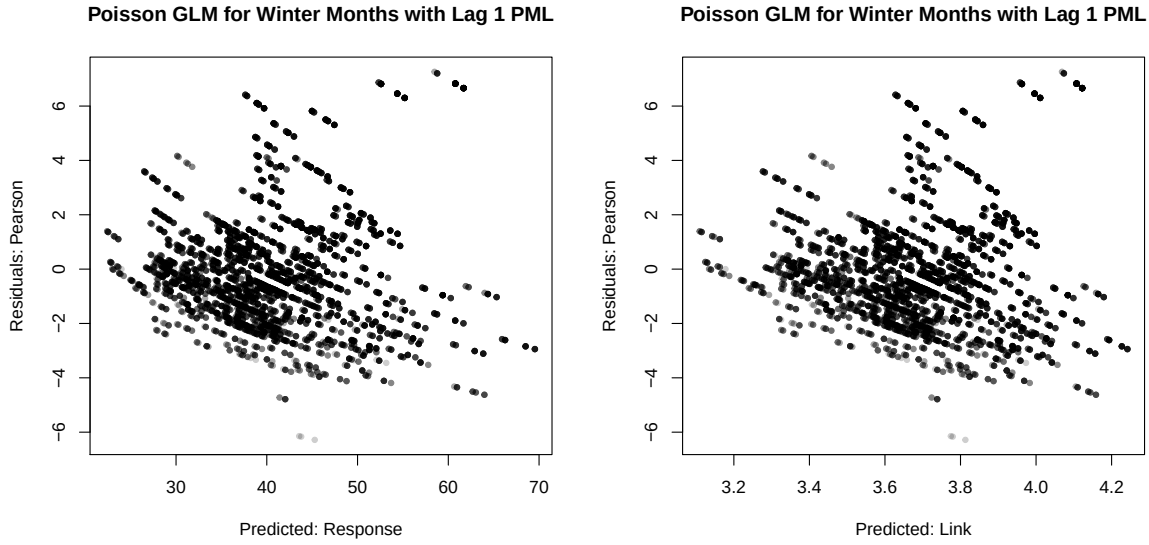

Figure 25: Two plots of the predicted values and residuals. The first plot is the Pearson residuals versus the predicted values. The second plot is the Pearson residuals versus the predicted link values.

The second model, a generalized linear binomial model, is of the same form as the generalized linear binomial model but β_2 is now the previous day's average particle matter level and the data has been restricted to the winter months November through February. β_2, the parameter for the previous day's average particle matter level, was highly significant with a p-value of 0.000. A 95% confidence interval for the β_2 parameter is 0.0048833 ± 0.00041924 or (0.0044641, 0.0053025). The p-values for the parameters β_4, β_5, β_6, and β_7 corresponding to wind speed, wind gust, average relative humidity, and average temperature accordingly, were all significant with a p-value of 0.000. Table 54 presents the results for the previous day's average particle matter level parameter β_2.

Table 54: The effect of the lag 1 average particle matter level for the binomial model for the winter months November-February. The estimate for β_2 is given along with the standard error for the estimate, a 95% confidence interval for the estimate, and the p-value for β_2.

Estimate	Standard Error	95% CI	P-value
0.0048833	0.0002139	(0.0044641, 0.0053025)	0.000

The estimate for β_2 for is 0.0048833. The log odds of absence is increased by 0.0048833 for each increase of one ug/m^3 of PM. The estimate for β_4, the parameter for wind speed, is 0.12401. The log odds of absence is increased by 0.12401 for each increase of one mph. The value for β_5, the parameter for wind gust is 0.014477. The log odds of absence is increased by 0.014477 for each increase of one mph. The value for β_6, the parameter for average relative humidity is 0.0080772. The log odds of absence is increased by 0.0080772 for each increase of one percent humidity. The value for β_7, the parameter for average temperature is -0.011057. The log odds of absence is decreased by -0.011057 for each increase of one degree Farenheit. The confounders that increased the log odds of absence are wind speed, wind gust, and average relative humidity while temperature decreased the log odds of absence. Figure 26 presents two plots of the fitted values and residuals. On the left the plot is the the predicted values versus the Pearson residuals. On the right the plot is the log of the predicted values versus the Pearson residuals. There is not a significant difference between the two plots. Most of the data is centered around 0 while varying between -5 and 5 with a few values lying outside of this range. On days with absences greater than 50, the model tends to over-estimate the log odds of an absence.

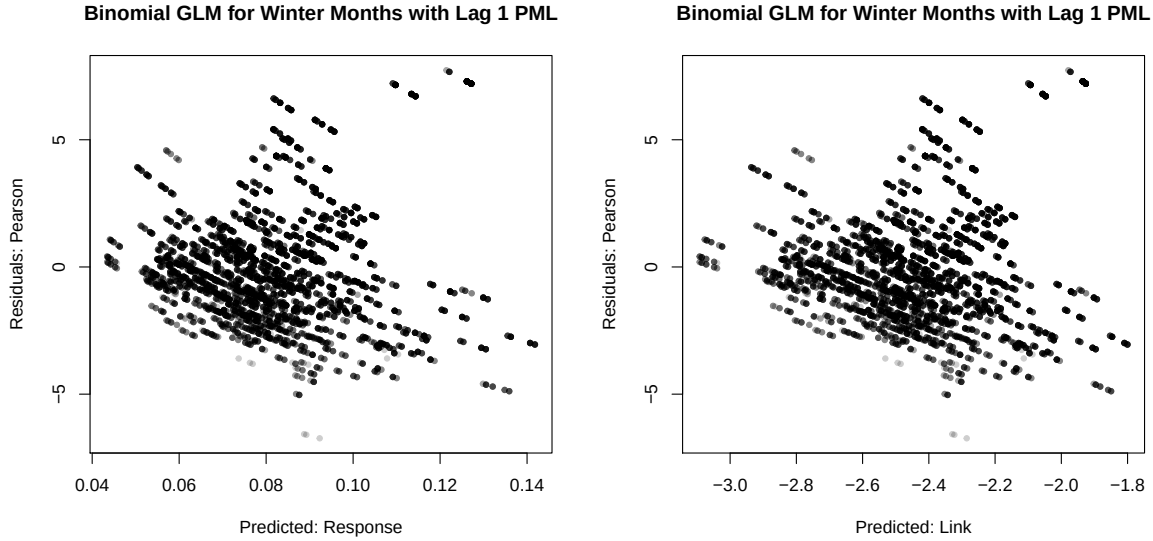

Figure 26: Two plots of the predicted values and residuals. The first plot is the Pearson residuals versus the predicted values. The second plot is the Pearson residuals versus the predicted link values.

CHAPTER 13
CONCLUSION

This thesis presents an analysis of average particle matter level and middle school absences in Libby, Montana. We first performed chi-square tests to see if absences varied over the course of the school year and whether absences varied over days of the week and months. We next performed rank tests to investigate the possibility of a relationship between average particle matter level and the number of absences. We also conducted rank tests within each day of the week and month. Using the results of the previous two tests, we performed a chi-test to see if a particular day of the week or month were significant. We created generalized linear models for the significant days of the week and months. We then looked to see if the average particle matter level on the first day of absence had any effect on the length of the absence. To test if the difference between means was significant, we performed an ANOVA test. We next wanted to see which of several potential confounders were significant. We had two types of models, Poisson and binomial, and several cases of each of these types. We created models where the data was just restricted to the winter months, November through February. We created models with the year average particle matter level. We also created models with the lag 1 particle matter level. Consistently throughout each of the generalized linear models increased wind speed, wind gust, and average relative humidity increased the expected number of absences and increased the log odds of absence while increased temperature did the opposite.

Our work has shown days with a high average particle matter level are associated with a higher number of absences. There are certain days of the week and months where a higher average particle matter level is associated with a higher number of absences. We can also conclude the average particle matter level on the first day of an absence does have an effect on the first day of an absence. We also identified several confounders that affect the number of absences.

REFERENCES CITED

[1] Currie, J, Hanushek, E, Kahn, E, Rivkin, S (2009) "Does Pollution Increase School Absences?" *The Review of Economics and Statistics*, 91:682-694.

[2] Makino, K (2000) "Association of School Absences with Air Pollution in Areas Around Arterial Roads" *Journal of Epidemiology*, 10:292-299.

[3] Park, J, Lee, B, Ha, E, Lee, J, Kim, H, Hong, Y (2002). "Association of Air Pollution With School Absenteeism Due to Illness" *Journal of the American Medical Association* 156:1235-1239.

[4] Wenten, M, Gauderman, W, Berhane, K, Lin, P, Peters J, and Gilliland, F (2009) "Functional Variants in the Catalase and Myeloperoxidase Genes, Ambient Air Pollution, and Respiratory-related School Absences: An Example of Expistasis in Gene-Environment Interactions" *American Journal of Epidemiology* 170:1494-1501.