Analysis of Racial, Socioeconomic, & Other Factors in Breast Cancer Incidence

Based on data from the Illinois Department of Public Health

Bailey W. Perry
University of Minnesota – Twin Cities

Introduction:

This study investigates the relationship between breast cancer incidence and several demographic, socio-economic and spatial factors near Chicago, IL. This topic was chosen because of the importance of medical access and the healthcare debates that are relevant to today's society; specifically, there is talk of the racial disparities that plague the system. The model selected from this study concurs with multiple reported relationships between breast cancer incidence and predictive variables, including incidence rate increasing with the racial predictor African American increasing. This analysis considers three identified models to better understand what predictors play important roles in incidence rates, what effect they have, and which model seems most valid.

Data:

The data for this study come from the Illinois Department of Health (IDH), the United States Census Bureau (USCB), and online mapping software. IDH reports cancer incidence cases by zip code for both in situ and invasive breast cancer cases, and this provided the foundation for this study. Instead of using the entirety of Illinois, a radius of 40 km using Glen Ellyn as the center point was conducted, and all zip codes within this range were included in the dataset. This was done to ensure feasibility of the project, but also includes major cities such as Evanston, Chicago, and Naperville.

As previously mentioned, the Illinois Department of Public Health was the key resource for obtaining incidence values by zip code. The dataset could be selected for many different time frames, but 2010-14 was the range of interest for this study. The data table produced had columns for year, zip code, cancer group, and total count. The values pulled from this included zip code and total count.

The next step for collection utilized the US Census Bureau FactFinder tool. Under the "Community Facts" tab, there is a search bar that allows for zip codes as input, and returns data applicable to that zip code. There are many tabs to the left, as well as a list of popular tables for the geographic region. The first demographic data variables extracted from this site were from the "American Community Survey (ACS) Demographic and Housing Estimates (2012-16)"; this included % white, % black, % Asian, and % over 65 for each zip code. Another table was used from the FactFinder tool under the "Income" tab, namely "Selected Economic Characteristics," which was developed from the ACS as well. The data withdrawn from this included: civilian population with health coverage insurance percentage, percentage of females over 16 in the workforce, and mean household income. Tobacco use data was also included, but was only available at the county level, so the percent of adults who are current smokers (2007-9) was included for the zip codes based on their county. The Illinois Department of Public Health provided this data, accessible online, as well.

Finally, mileage to the nearest hospital was also included in the dataset as a spatial parameter of interest. This was done using google maps directions, where the zip code is the starting point and “hospital” is typed as the destination; google automatically selects the nearest hospital (from the center of the zip code) with directions that includes the time and miles to get there.

Originally, the dataset contained 292 zip codes, but there were null values mixed into this. It was discovered that the Census Bureau will suppress data to maintain confidentiality. This arises when the zip code is particularly small and therefore the individuals could easily be identified. Additionally, it is also noted that if the data is not statistically valid or possibly erroneous, they do not publish it. Finally, missing data is also oftentimes imputed, so the nulls were occurring specifically due to the characteristics previously listed. Due to this information, it is believed that eliminating the nulls could be addressed as an assumption, and those values were removed as a means for continuing with the analysis. This resulted in 218 zip codes in the final dataset. A snapshot of the final excel spreadsheet is included in the Appendix.

Methods:

The variable being analyzed is breast cancer incidence rate within the identified range of zip codes in Illinois. The dataset was modelled three different ways, as mentioned above, and the validity of each model and outlier testing was implemented to confirm the practicality of using that specific method. Since many of the predictors relate to demographics or economics, there are some common approaches on analyzing that type of data that provided insight for potential transformations of the predictors.

The overarching statistical technique was regression, but the models that were examined were chosen based on the type of data being analyzed. Since incidence cases per zip code can be classified as a rate (incidence/population) which is what we are interested in, a linear model investigating rate was used to see the relationships of the predictors to changes in rate. Another approach used the fact that incidence case values are count data, which immediately points to using a Poisson regression (Turner, 2008). Although the data are measured in counts, the point of interest is rate, and thus using an offset within the code allowed that to be possible via Poisson as well, and ultimately this was used for the second model. Finally, it was also evident that a Binomial regression was relevant to this data since there are *c* number of cases, *n* for the total population, and therefore we observe *n-c* people who do not have breast cancer (LaMorte, 2016). This puts the problem in terms of "success" being breast cancer and "failure" being not having breast cancer, and therefore the probability of success depends on the predictors being used in the model. This is commonplace for many public health outcomes that are involved in regression.

The steps followed to derive these final three models were adapted from Weisberg (2014). The general steps (using R, 2015) are listed as follows:

1. Fit desired regression on all main effects and interactions
2. Review diagnostic plots for the original model
3. Use backwards elimination (using AIC) with k=4 to identify significant regressors
4. Fit new regression with only the significant regressors identified above
5. Review diagnostic plots for the model
6. If issues identified via step 5, consider outlier testing or transformations
7. Perform necessary transformations and refit *and/or* remove outliers and refit
8. Review diagnostics for the new model

The use of k=4 in the backwards elimination process was chosen because it is less liberal with term selection. If k=2 was used, which is the default, too many terms would be retained in the model, and we would lose selectivity as well as clarity in the model. Testing with k=4 corresponds roughly to a test with a critical value of 0.05. More specifically, a linear model was used for rate, whereas quasibinomial and quasipoisson GLMs were used to account for overdispersion in the data when fitting those GLMs. Overdispersion occurs when the actual variance exceeds the GLM variance, meaning that the predictors are not capturing some additional source of variability. This is seen quite commonly with both Binomial and Poisson models, and we need the quasi models because the data follows the original distributions proportionally, but not identically. In the Binomial case, overdispersion means that the variance of the response is greater than the variance expected from the distribution and can be written as: $\mathrm{var}(Y) > np(1\text{-}p)$ where Y~Bin(n, p). To adjust for this, it is recommended that a scale parameter is introduced to the variance function: $\mathrm{var}(Y)=np(1\text{-}p)*\boldsymbol{\sigma}^2$ where the estimate of $\boldsymbol{\sigma}^2=X^2/(n\text{-}r)$. The value for X^2 is the usual Pearson goodness-of-fit statistic, n is the number of sample cases, and r is the number of parameters, where (n-r)=degrees of freedom. In the Poisson case, it means that: $\mathrm{var}(Y) > \mu$, and it is recommended to scale with $\mathrm{var}(Y\ ) = \boldsymbol{\varphi}E(Y\ ) = \boldsymbol{\varphi}\mu$, where the estimate of $\boldsymbol{\varphi}= X^2/(n - r)$. Again, the same X^2 statistic, n, and r are involved. The standard errors are multiplied by the square-root of the scale (dispersion) parameter, but the coefficient estimates stay the same, so the test statistics and p-values are adjusted appropriately by this scaling of the standard errors. Overdispersion may also be caused by possible omitted covariates, but since more data cannot be collected now, the best method is to adjust for overdispersion as discussed above (and done by implementing the quasi functions in R). The R code containing the entirety of this analysis is included in the Appendix.

I. General Formulas

The following information includes additional description and formulas to further explain the techniques used for the dataset. To start, the linear model will be addressed. The following formula expresses the linear model of rate in general terms of the dataset:

$$E\left(\frac{cases}{population} \middle| X_1 \dots X_p\right) = \beta_0 + \beta_1(\%white) + \beta_2(\%black) + \beta_3(\%asian) + \cdots + \beta_{12}(\%white\!:\%black) + \beta_{13}(\%white\!:\%asian) + \cdots$$

Not all terms are explicitly written out, but the formula includes all main effects and two-factor interaction terms, until being reduced based on the outcome of backwards elimination. Note that the betas are dependent on population in the above rate model because population is used as the weights for that regression in R. Weighting in this model allows us to remove the assumption that all observations should be treated equally. In this case, weighting by population (which differs for each zip code) allows us to give each data point a proper amount of influence over the parameter estimates. More specifically, we need to weight the model so that small zip codes do not have the same influence as large zip codes. The next formula expresses the quasipoisson model in general terms of the dataset (where the betas are different from those listed above):

$$\ln\big(E(cases|X_1 \dots X_p)\big) = \beta_0 + \beta_1(\%white) + \beta_2(\%black) + \beta_3(\%asian) + \cdots + \beta_{12}(\%white\!:\%black) + \beta_{13}(\%white\!:\%asian) + \cdots + \log(pop)$$

The term log(pop) is included here because it was used as an offset for the quasipoisson model. This is done because absent any other effects, the mean is expected to be proportional to population, and that is believed to be true for this dataset. The underlying random variable is still cases, but when the offset is included, you can also look at the model equation as a rate of the events per unit population. In this scenario, the intercept plus the offset says the mean is proportional to population, and everything else in the model shows how count and population differ from this simplest possible relationship. Finally, the formula below expresses the quasibinomial model in general terms of the dataset (different betas yet again):

$$logit = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1(\%white) + \beta_2(\%black) + \beta_3(\%asian) + \cdots$$
$$+\beta_{12}(\%white{:}\%black) + \beta_{13}(\%white{:}\%asian) + \cdots$$

Something to note for the quasibinomial model is that population is the value for n when we consider our response following the binomial distribution: Y~Bin(n,p), where p = probability of incidence or the “successes”.

II. Diagnostic Techniques

The main things typically considered when checking model fit include investigating outliers, the residuals v fitted plot, the normal qqplot, and finally the actuals v fitted plot to check how well the model fits to the data. Checking outliers is essential to better understand their impact on the models/plots, as well as why they may be appearing. In this scenario, outlier testing was completed via R, and two cases appear to be outliers in all three models, numbers 210 and 55. When investigating these data points in the excel spreadsheet, there wasn’t any clear reason as to why they stood out. Most of their values for the various predictors were comparable to the other zip codes. After reviewing other external factors including maps, crime rates, manufacturing, and air quality indices, there still wasn’t anything that highlighted these cases (translating to zip codes 60148 and 60659). Since they had minimal effects on the regression outputs, but larger effects on the assumptions, it was assumed acceptable to drop these cases and move forward with the study. Regarding the diagnostic plots, since the data has a large sample size, we should expect to see approximate normality in the quantile-quantile plot, but this is not an assumption that needs to be met for the quasi models. The residuals v fitted and actuals v fitted are checks for model fit, and if there is no pattern in the residual plot, and a clear 45-degree diagonal in the actuals plot, then the model appears to be a good fit. These plots for each model are found below in the results section.

Although addressed previously, it is extremely important to understand that the models have flaws that may be explained by missing predictors. The inability to address true genetic characteristics and only using race as a generic proxy for that was a key issue for the models.

The diagnostics overall showed improvements as the models were finessed to the final three included in this analysis, and this was enough to investigate their meaning. These models could see improvement with further research and data in these other areas.

Results:

Throughout the iterations of the data analysis, it became clear that there were versions of each model that were most appropriate and valid. These models are discussed below in detail with interpretation of the final model. An important thing to note is that throughout the analysis, incidence cases were not standardized, and is solely in terms of number of cases. Another final note before investigating the models is that not all the predictors were independent. It is easier to understand the results with uncorrelated predictors, but it is not required. In our case, none of the predictors had exact linear relationships so although it is not ideal, the regression can still be performed, and this is an area that could be improved for future work.

III. Model Output

The quasibinomial model was the first model to go through the general steps outlined in the methods section, and thus will be discussed first. Some of the varying attempts at this model involved including or excluding certain predictors, as well as determining whether transformations may be necessary or not. Originally, the model resulting from backwards elimination contained terms that were not statistically significant. When checking how their removal affected the model, the null and residual deviances did not see much change and the dispersion parameter was minimally affected as well. Therefore, it was considered valid to remove those and create a simpler model. In the end, the quasibinomial model that provided the best diagnostics contained 18 terms, and was based on the dataset with the outliers removed. The scale parameter to adjust for overdispersion was σ^2=4.59. The resulting model coefficients and standard errors are found in the following table. Additionally, the significance codes for the next three tables are as follows from R: 0 ‘***”, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’, and 0.1 ‘ ’.

	Estimate	Standard Error	Significant?
Intercept	-0.11	1	***
White (%)	0.018	0.007	**
Black (%)	0.019	0.007	**
Asian (%)	0.011	0.0086	
Population	0.000004	0.0000024	.

Age (% over 65)	0.1	0.059	.
Income	0.000034	0.000012	**
Smoke (%)	0.0094	0.0055	.
Insured (%)	0.04	0.01	***
Work (% over 16 working)	-0.016	0.00046	***
Miles (to nearest hospital)	0.039	0.0014	
White:Asian	0.0002	0.000096	.
White:Pop	-0.000000088	0.000000034	**
White:Age	-0.014	0.00056	*
Black:Age	-0.014	0.0006	*
Asian:Age	-0.017	0.0006	**
Age:Work	0.0011	0.00027	***
Income:Insured	-0.00000034	0.00000012	**

Table 1. Quasibinomial Model Output

The quasipoisson model is the next for discussion. One of the key issues that arose during the variable selection process was the marginality principle (Nelder, 1977). The model that backwards elimination produced was in violation of this principle because it did not contain all the main effects that were involved in the interaction terms included in the output model. To address this, these main effects were re-added to the model, and it went through all the checks to see if it was still valid. Like the quasibinomial, the quasipoisson model resulting from backwards elimination also contained terms that were not statistically significant. Again, the null and residual deviances did not see much change and the dispersion parameter was minimally affected. In the end, the quasipoisson model that upheld the marginality principle, with insignificant interactions removed, maintained the best diagnostics, and was selected for analysis. The model contains 16 terms, and was again based on the dataset with the outliers removed. The scale parameter to adjust for overdispersion was φ=4.704. The resulting model is:

	Estimate	**Standard Error**	**Significant?**
Intercept	-0.1	1.009	***
White	0.019	0.066	**
Asian	0.011	0.0086	
Pop	0.0000046	0.0000024	.

Black	0.02	0.0065	**
Age	0.11	0.06	.
Income	0.000035	0.000012	**
Insured	0.038	0.0099	***
Work	-0.016	0.0046	***
White:Asian	0.0002	0.000096	*
White:Pop	-0.000000092	0.000000034	**
White:Age	-0.0014	0.00057	*
Black:Age	-0.0014	0.0006	*
Asian:Age	-0.0017	0.00064	**
Age:Work	0.0011	0.00028	***
Income:Insured	-0.00000035	0.00000013	**

Table 2. Quasipoisson Model Output

Finally, the linear model of rate is the last model analysis to consider. This model uses rate (cases/population) as the response variable, and the rest of the terms in the model were derived from backwards elimination on the full model of main effects and interactions. The final rate model contains 8 terms, and was again based on the dataset with the outliers removed. The resulting model is:

	Estimate	**Standard Error**	**Significant?**
Intercept	-0.0026	0.00016	.
Black	0.000015	0.000003	***
Age	-0.000054	0.000084	
Income	0.000000014	0.000000003	***
Smoke	0.00007	0.000028	*
Insured	0.000054	0.00002	**
Work	-0.000051	0.00002	*
Age:Work	0.0000049	0.0000014	***

Table 3. Linear Model of Rate Output

IV. Model Diagnostics

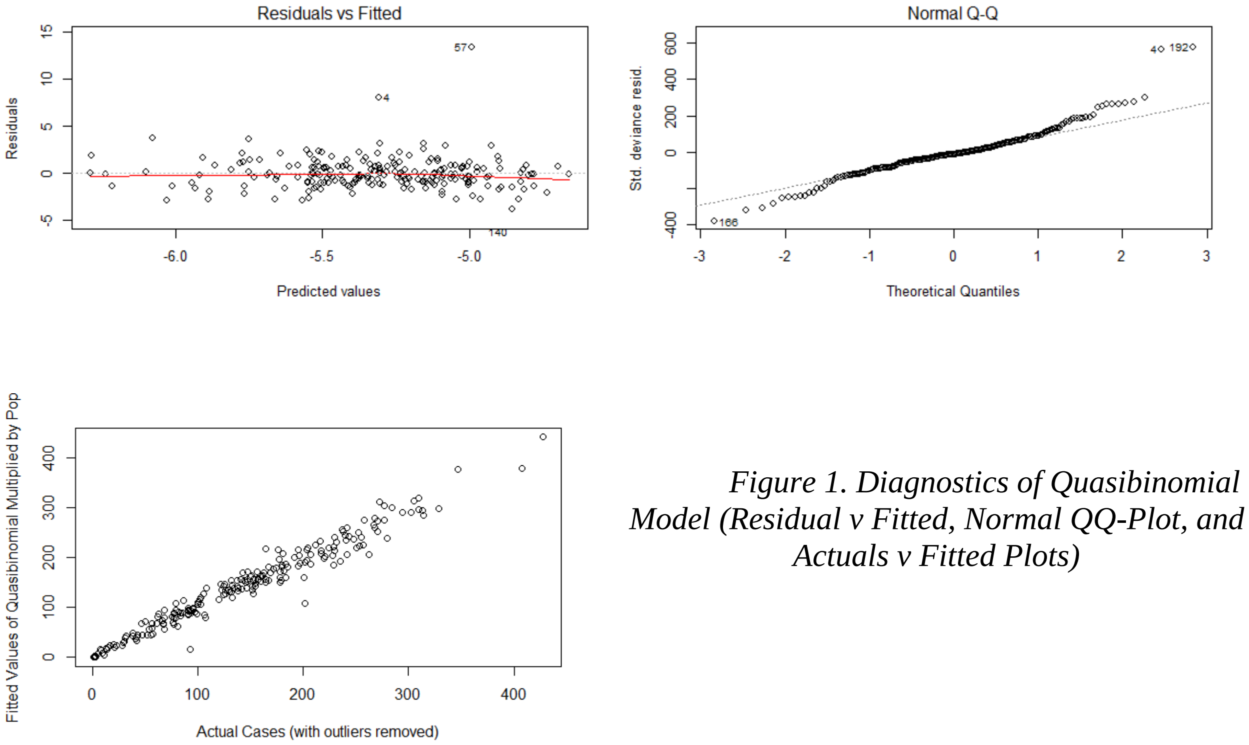

Figure 1. Diagnostics of Quasibinomial Model (Residual v Fitted, Normal QQ-Plot, and Actuals v Fitted Plots)

The above diagnostic plots for the quasibinomial model showcase all the characteristics that would be expected for a well-fitting model as described previously. The plots showcase the desired characteristics, except for case 57 and 4 standing out. When checked, these did arise as new outliers after cases 210 and 55 had been removed, but when these cases were removed, new outliers also arose. Therefore, it was determined that removing just 210 and 55 was the best method and would keep the models most comparable. Next is the quasipoisson on the following page.

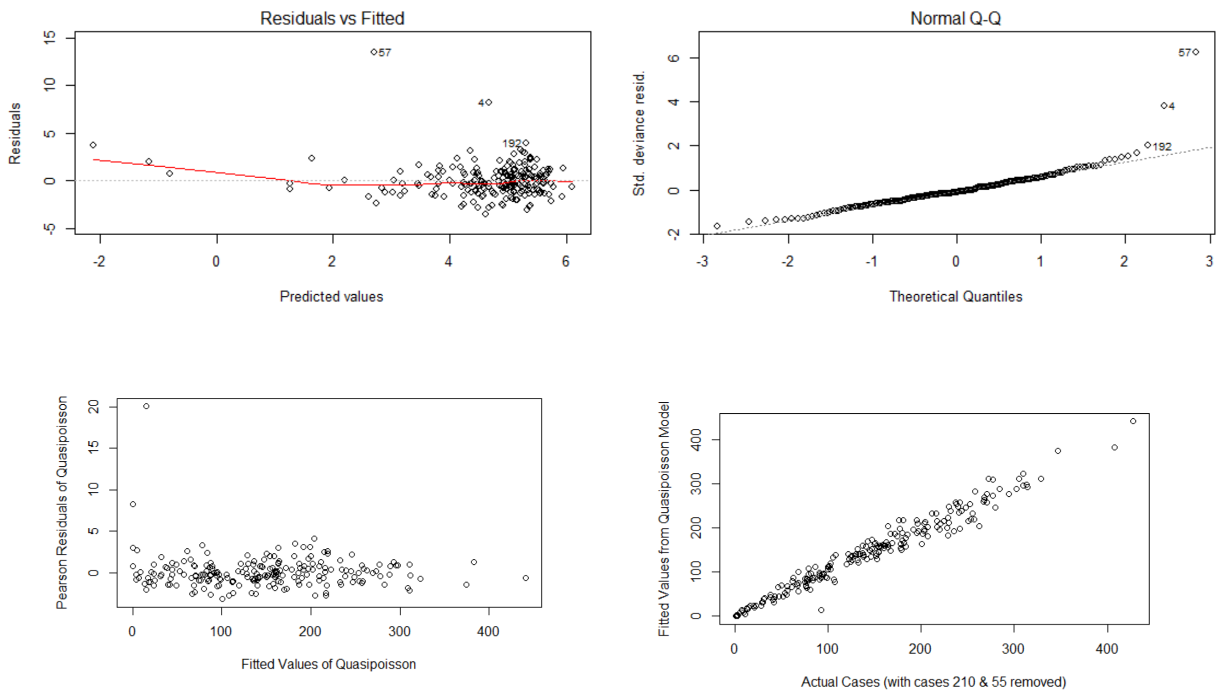

Figure 2. Diagnostics of Quasipoisson Model

(Residual v Fitted, Normal QQ-Plot, Pearson Residuals v Fitted, and Actuals v Fitted Plots)

The Residuals v Fitted plot (upper left) shows some of the points clustering on one side of the plot which tells us the model is underpredicting on the extreme low end, but this isn't necessarily an issue; We see fitted means between exp(2) and exp(6) with a few tiny means, which may be high leverage points, although we do not see this affecting the fitted values. To confirm or negate this issue, it is essential to check the Pearson residuals v fitted plot. This plot, on the bottom left, proves it was a non-issue since the points follow no clear pattern and form a band around zero, except for one big residual on the very low end where the model predicts nearly 0, but we have around 20. Investigating these residuals is key because this formula helps account for the variance function for the model, and thus are standardized. The remaining plots all match what is preferred except for two cases standing out.

Cases 57 and 4 stood out for the quasipoisson as well. Since these outliers did not arise with the linear model of rate, and new outliers were identified after removal again, they were left in the model and recognized as an imperfection. Finally, the linear model of rate is checked.

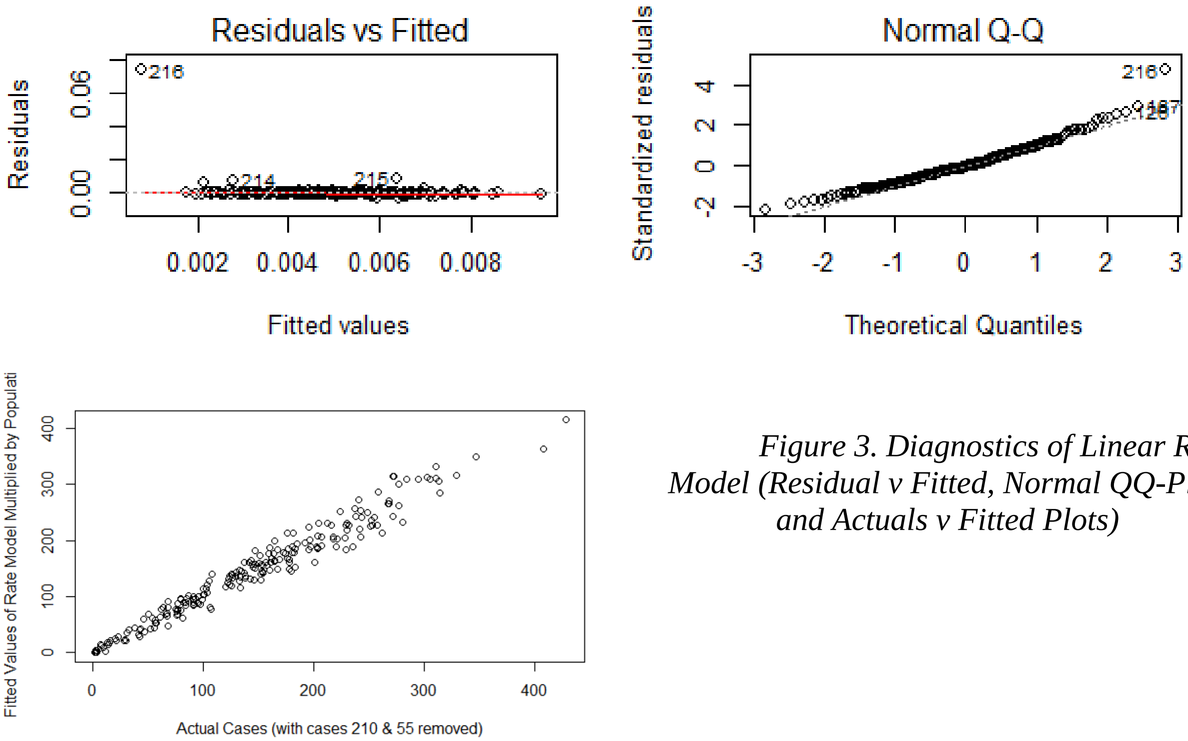

Figure 3. Diagnostics of Linear Rate Model (Residual v Fitted, Normal QQ-Plot, and Actuals v Fitted Plots)

The plots showcase the desired characteristics, except for case 216 standing out in the residuals v fitted plot. Removal of case 216, like the scenarios above, created new outliers and thus this case was left in the dataset as well.

I. Model Interpretation

For this section, two methods will be used to interpret the coefficients. First, since the predictors white, black, and asian do not sum to one, we are not forced to remove one of them from the dataset, and this implies there is an “other” category that encapsulates the final portion for race. Therefore, instead of shifting these racial predictors by one percent without consideration of the effect on the other three portions, it was determined that the percent shift should result in a proportional decrease across the others. Each of their means were calculated to start, and then a single predictor was shifted by one percent. Next, the mean of one of the other three was taken over the sum of the remainder and this determined the proportion of the one percent decrease that would take. This was done for all three remaining categories. For example: white increases by 1%, so sum black, asian, and other categories, 12.4+8.4+9.2=30% remaining total; then black proportion is 12.4/30=0.41, and therefore black decreased by 0.41. The following table was created to demonstrate the results of this process:

	White	**Black**	**Asian**	**Other**
Mean (%)	70	12.4	8.4	9.2
White ↑ 1%	71	11.99	8.12	8.89
Black ↑ 1%	69.2	13.4	8.31	9.09
Asian ↑ 1%	69.24	12.26	9.4	9.1

Table 4. Adjusted Race Percentages

The point of this was to determine values necessary for the interpretation process. To interpret the coefficient for a racial predictor, it was decided that the change in the linear predictor with a unit change in one predictor would be investigated. This was done instead of a shift of one standard deviation because that does not logically make sense for these predictors. Therefore, the coefficients were calculated using the coefficient for the predictor plus or minus the coefficient for any interaction terms involving the variable multiplied by the mean value of the other variable in the interaction. The one exception was interaction terms with race. Since the other racial categories had to decrease for an increase in the racial predictor being investigated, the mean needed to be multiplied by the new proportion that it represented, or in other words, the new percentage in Table 4 above needed to be implemented. To show this process, part of the calculation for the coefficient white for the quasibinomial model is: $white = 0.018 + 0.00017 * 8.12 - 0.000000088 * (31027.12)$ … In this equation, percent asian was in the interaction term and is highlighted in red, and the value in parenthesis is the mean of population. Following this process, the racial coefficients, accounting for interactions and a one-unit increase were calculated and are found in their respective tables of coefficients below.

The second step involved for interpretation was going through the remainder of the predictors and setting the interaction terms to their means. But, in this case, instead of looking at a unit change, investigating how the response changes when the predictor value is increased by one SD was determined to be the best method. A generic example of this process is: $age\ estimate = \beta_1(SD_{age}) + \beta_2(SD_{age}) * (\overline{white})$ … Where in this case, the second term is an interaction between the predictors age and white. The following tables indicate the results of these two processes for all the terms in each of the three models.

Predictor	**Coefficient**	**Odds Factor**
White*	-0.172	0.842
Black*	-0.17	0.844

Asian*	-0.20	0.819
Pop	-0.046	0.955
Age	-6.66	0.0013
Income	0.145	1.156
Smoke	0.023	1.023
Insured	0.035	1.036
Work	-0.072	0.931
Miles	0.409	1.505

*Table 5. Quasibinomial Coefficient Estimates (*unit change not SD)*

To interpret quasibinomial coefficients, it is important to recall that it is in logit form. Thus, the log odds are being considered, and it is a *factor* change. From above, for a one-percent increase in percentage white, the odds of having cancer decrease by a factor of $e^{-0.172} = 0.842.$ Following this method applies to other the terms in the model and these values are included in column three titled "Odds Factor".

As we saw before with the quasibinomial model, the quasipoisson model is also difficult to decipher visually. To interpret these coefficients, it is slightly simpler than the log odds, but it is also necessary to follow the two-step method above. This model only requires transforming the response back to cases by anti-logging the terms on the right-hand side of the formula. The following table summarizes the coefficient estimates for the quasipoisson model:

Predictor	**Coefficient**	**Exp. Incidence Rate Change**
White*	-0.12	0.887
Black*	0.00113	1.001
Asian*	0.0019	1.002
Pop	4.284	72.53
Age	0.283	1.327
Income	0.149	1.161
Smoke	N/A	N/A
Insured	-1.846	0.158
Work	-0.01	0.990
Miles	N/A	N/A

*Table 6. Quasipoisson Coefficient Estimates (*unit change not SD)*

The coefficient for population is slightly unique because log(pop) was included as an offset, so it played in to the final coefficient estimate on top of the regular main effect term and interactions. To interpret quasipoisson coefficients, the values are anti-logged and the result is the expected value of cases given the values of the predictors, found in column three above. From this, for a one-standard deviation increase in population, the expected number of cases of cancer increase by $e^{4.284} = 72.53$ and so on. Again, the racial predictors are a one-unit increase with all else held to their means.

Finally, the linear model of rate is considered. This model is unique compared to the other two in this analysis because there are no transformations of the response necessary to interpret it. Additionally, for this model, only two predictors are involved in an interaction term, age and work. Table 6 highlights these coefficient estimates:

Predictor	**Coefficient**
White*	N/A
Black*	0.000016
Asian*	N/A
Pop	N/A
Age	0.0014
Income	0.00064
Smoke	0.00015
Insured	0.00031
Work	-0.00041
Miles	N/A

*Table 7. Linear Model of Rate Coefficient Estimates (*unit change not SD)*

To interpret the predictor smoke: if percentage who smoke is increased by one percent, the rate increases by 0.00015. This table did not need a third column for understanding how the response is affected because no transformations were necessary. On top of this, it gives a clearer picture of the magnitude each predictor holds on affecting the overall rate. Recall that population weights this model since it is a linear model of rate.

Comparison:

To do further model comparison, it was important to look at the number of models that each of the main effects are found in. Following the marginality principle, the main effects are most important, and if they are involved in an interaction term, then it would be found as a main effect in the model.

	%white	%black	%asian	population	%over 65	%insured	income
# of models	2	3	2	1	3	3	3
	%smoke	mileage	%workforce				
# of models	2	1	3				

Table 8. Model Tally by Factor

From this table, certain factors were found to be common based on being present in all three model variations. One interesting thing to note is that the population main effect was only found in the quasibinomial model and only appears by implication as the sum of successes and failures, but it was utilized as an offset for the quasipoisson and was part of the response for the linear rate model so it was in the three models in different forms.

When comparing coefficients from the three tables above, it is essential to recall that the racial predictors were changed by one percentage (unit), while the other predictors were changed by one standard deviation. This means direct comparison across these groups is not necessarily fair, but the largest (absolute) magnitude predictors within these two groups are: asian and age for quasibinomial, white and pop for quasipoisson, and black and age for linear rate.

Something else to consider, is the general direction of association that each predictor had within the specific models. This gives insight into the effect that the term has individually (holding all else constant) on the response, without looking at magnitude (and excluding the intercept) as was done above. An upward arrow indicates an increase, and a downward arrow indicates a decrease in the respective response. Since some of these predictors may be related, such as percentage in the workforce being negatively related to income and percentage over 65, the interpretation and comparison of individual coefficients cannot be taken too far.

	white	black	asian	pop	over65	insured	income	smoke	miles	workforce
Quasibinomial	↓	↓	↓	↓	↓	↑	↑	↑	↑	↓
Quasipoisson	↓	↑	↑	↑**	↑	↓	↑	N/A	N/A	↓
Linear Rate	N/A	↑	N/A	N/A*	↑	↑	↑	↑	N/A	↓

Table 5. General Direction of Association by Term & Model
*(**with the log(pop) term included and *population as part of the response)*

Another important thing to consider is which model seems to perform the best. When referencing the Actuals v Fitted plots of the data, there are not major differences between the three models which means they all fit approximately the same. Additionally, the null and residual deviance are very similar for both quasi-models, so they are not necessarily the best checks or comparisons for goodness of fit, especially since it does not apply to the linear model of rate.

Some users may prefer the linear model of rate due to the minimal number of terms in the model, and a relatively high adjusted R-squared, 0.723. Other than comparing diagnostics it can be very difficult to compare statistical models that are not nested, and in this case the models also have different formats for the response variable. In the end it comes down to preference on which method of comparison is most important to the user or statistician making the decision. For this scenario, I would select the linear model of rate since it describes the relationship with cancer incidence cases very well as seen in the almost perfect Actuals v Fitted plot. It also does not require any transformation to understand the effects of each term on the response, and has the least amount of terms to consider in the model. Therefore, the model not only fits well, but it is also relatively user friendly.

Discussion/Summary:

To date, cancer is still a major research topic as cures and treatment are evading those in need. Ideally, laboratories are working to identify genomic based-solutions to solving this problem, but both time and money are burdens on this. On top of this, other people have hypothesized about the impact of race on cancer incidence cases and mortality, and there have been studies on whether African Americans are at higher risk than other races. There are still no decisive answers, and the data is constantly changing and growing, but some researchers have identified the disparity and are working to uncover the complex reasons as to why it exists (Desantis et al., 2015). Although this regression analysis does not give definitive answers, it helps explore the relationships these included predictors may have with cancer rates, and gives insight into areas for more exploration. The study revealed complex relationships within predictors as well as with the response, and these are summarized in the Results section. Overall, the three models did a good job of fitting to the data, but the linear model of rate is the most intuitive and user-friendly.

This model, along with the quasipoisson, suggested that the racial factor black had a positive association with expected cases, which supports the literature regarding these racial/health disparities. Yet, surprisingly, the model also had positive associations with income and the percentage of people insured. One possible explanation is that diagnosis occurs for those with insurance and/or the ability to pay for doctors' visits, but one could also assume that diagnosis occurs at some point; it just may be later when it is further developed or too late for those with less money or access to insurance. In either case, the model provokes interesting conclusions to consider moving forward.

Further investigation could include genetic factors (if accessible), and the use of genomic profiles to consider survival analysis. On top of this, it would be important to try to identify the other factors that could aid in rounding out the dataset to gain a better picture of the predictors involved in understanding this disease. Additional studies could take another more “proactive” approach by focusing on risk analysis, and exposures that lead to higher risk of development rather than the “reactive” approach of reviewing cases that have already occurred. Finally, it could be interesting to gain a larger geographical perspective and dive into more spatially related characteristics and comparing regions and their risk. Another method used by many statisticians is the process of cross-validation. In this case, that would involve splitting the data into training and testing sets before creating the model. This is done to see how the results of the model generalize to an independent dataset. In this project, splitting the data was not performed because cancer incidence rates are recorded at regular time intervals for these zip codes, and therefore cross-validation could be performed by using future data for the area.

The goal of this paper was to not only research and deepen personal understanding about disparity in public health, specifically through cancer, but to also gain a better understanding of the regression techniques and statistical methodology utilized to form the three models analyzed. The articles and journals with past and present research were motivating, and an inspiration to continue pursuing a career in health/medical statistics. There is much more to learn, but this analysis was a valuable lesson on the application of statistics and the importance of digging deeper into why things are happening around us.

References

American Cancer Society. (2016). Cancer Facts & Figures for African Americans. Retrieved February 13, 2018, from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-african-americans/cancer-facts-and-figures-for-african-americans-2016-2018.pdf

Desantis, C. E., Fedewa, S. A., Sauer, A. G., Kramer, J. L., Smith, R. A., & Jemal, A. (2015). Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA: A Cancer Journal for Clinicians,66*(1), 31-42. doi:10.3322/caac.21320

Ghafoor, A., Jemal, A., Ward, E., Cokkinides, V., Smith, R., & Thun, M. (2003). Trends in Breast Cancer by Race and Ethnicity. *CA: A Cancer Journal for Clinicians,53*(6), 342-355. doi:10.3322/canjclin.53.6.342

LaMorte, W. W. (2016, July 24). The Role of Probability. Retrieved March 18, 2018, from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability7.html

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. Retrieved March 15, 2018, from http://ww2.amstat.org/publications/jse/v9n3/stanton.html

Turner, H. (2008, April 22). Introduction to Generalized Linear Models. Retrieved March 6, 2018, from http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

University of California – Los Angeles (UCLA): Statistical Consulting Group. (n.d.). Deciphering Interactions in Logistic Regression. Retrieved March 20, 2018, from https://stats.idre.ucla.edu/stata/seminars/deciphering-interactions-in-logistic-regression/

Wallace, A. (2017, September 25). Ethnic minorities, elderly underrepresented in cancer clinical trials. Retrieved March 19, 2018, from https://www.upi.com/Ethnic-minorities-elderly-underrepresented-in-cancer-clinical-trials/2721506368866/

Wang, F., Mclafferty, S., Escamilla, V., & Luo, L. (2008). Late-Stage Breast Cancer Diagnosis and Health Care Access in Illinois∗ [Abstract]. *The Professional Geographer,60*(1), 54-69. doi:10.1080/00330120701724087

Weisberg, S. (2014). *Applied linear regression*. Hoboken, NJ: Wiley.