# Predicting Cancer Incidence of Counties in the United States

Michael Strange, Toshan Doodnauth, Joshua Geiger, and Christopher Friedman

Georgia Institute of Technology

ISYE 6414: Statistical Modeling and Regression Analysis

**Abstract**

The team aimed to develop a regression-based model to predict age-adjusted cancer incidence rates at the county level across the United States, utilizing macro-level social, economic, and lifestyle predictors. Such a model is intended to guide the allocation of resources, such as cancer screening programs and treatment services. The approaches explored included elastic net, logistic regression, Poisson regression, and multiple linear regression, with some models incorporating variable selection techniques. Key findings highlighted the significance of state and rural/urban classifications as predictors, consistently observed across all models. While the study served as a valuable practical application of regression principles as learned throughout the course, no single model was identified that simultaneously satisfied all assumptions, demonstrated statistically significant coefficients, provided a strong goodness of fit, and exhibited meaningful predictive power.

**Predicting Cancer Incidence of Counties in the United States**

The origin of predicting cancer incidence rates at a county level can be seen in its necessity for an effective and responsive public health surveillance operation. This is crucial for determining where to allocate resources, such as screening programs or cancer treatment services. The progression of this problem can be seen through advancements in data collection, statistical methods, and machine learning models. Initially, simple age-adjustment techniques were used, but over time, more sophisticated models have incorporated environmental, socioeconomic, and behavioral factors that contribute to cancer rates (Dong et al., 2022). For example, Zhang discusses how the physical environment may be involved, with exposures to specific factors such as cigarette smoking, alcohol use, physical exercise, dietary intake, water or air pollution, and radiation, all of which have been linked to affecting the rate of cancer incidence (Zhang, 2017). Furthermore, predictive models and geographic information systems are used to identify cancer trends more accurately at the county level, as seen with the American Cancer Society's use of a segmented-line linear regression model (Yu, 2013). In our dataset, we follow the standard practice of using an age-adjusted cancer incidence rate to account for the discrepancy between age groups (Withrow et al., 2019). Obtaining a more accurate prediction of cancer incidence by county could result in targeted public health campaigns, better resource allocation, and a reduction in cancer occurrence by directing more care towards at-risk populations.

Clegg et al (2009) discuss an approach to looking at cancer incidence through a lens of socioeconomic status (SES). Their approach involved combining nationwide data gathered over thirty years from the National Cancer Institute and the US Census to examine how often cancer in general - and different types of cancer - were diagnosed for different age groups stratified by different measures of socio-economic status. This study found that individuals from lower socioeconomic groups are more likely to be diagnosed with cancer and are more likely to be diagnosed with a higher stage cancer. This study uses the individual level to map cancer diagnosis and journey to individual socioeconomic

characteristics that may contribute to cancer. What this approach lacks is a wider view into other factors that may contribute to cancer incidence, such as environmental or dietary factors. In addition to Clegg et al, others have found that socioeconomic status may be related to cancer incidence and development of cancer (Clegg et al., 2009; Tabuchi, 2020). Geographic area (metropolitan vs. non-metropolitan areas) has been indicated to be related to certain types of cancer incidence (Henley et al., 2017; O'Neil, 2019). There have also been indicators that social inequalities can lead to differences in cancer deaths (Shi et al., 2005).

Moss et al. attempted to determine county level cancer-incidence for specific types of cancer by relating the incidence rate to ethnic compositions, cancer type (i.e. breast, cervical, colorectal), physician density, and median household income (Moss et al., 2021). While this study does use county level data for analysis, it does not account for all other cancer types, cancer-related factors such as air quality, poverty statistics (further than median household income), education levels, and prior health outcomes by county (Zhang, 2017). Another county level prediction method investigated by the National Institute of Health is the "Joinpoint Regression". This model does not address SES, environment, or dietary predictors(National Cancer Institute, n.d.). Others have attempted to use logistic regression to predict the probability of mortality rates using traditional socioeconomic predictors for specific 'sites' of cancer (Liu et al., 2021).

## Analysis

In the following sections, the different analytical methods pursued are discussed along with the results of each analysis. The response data used in these analyses, cancer incidence per 100,000 people in United States counties come from the National Cancer Institute (National Cancer Institute & Centers for Disease Control and Prevention, 2024). Predictors regarding socioeconomic status factors such as education, poverty status, and unemployment come from the US Department of Agriculture (Economic Research Service, US Department of Agriculture, 2025). Predictors regarding county-level characteristics

related to food access come from the Food Environment Atlas (US Department of Agriculture, 2024). County-level air quality index statistics come from the Environmental Protection Agency (Environmental Protection Agency, 2024b, 2024a). Predictors regarding health-related measures come from the PLACES dataset (Centers for Disease Control and Prevention, 2024).

## Logistic Regression

Logistic Regression was pursued as a potential solution to address the collective problem of predicting age adjusted cancer rate at the county level. Through this analysis one could suggest that a higher rate of age adjusted cancer could be present and predictable if access to healthy food is lower and poverty is higher. This hypothesis is being tested through the lens of logistic regression to determine if a statistically significant and good fit model can be demonstrated that can predict a county to be in the top 25 percent of age adjusted cancer rate using poverty and food variety predictors.

### *Datasets*

The 'incd.csv' dataset was used at the core dataset. The core data was merged with 'Poverty2023.xlsx' for the poverty predictors considered and 'FoodEnvironmentAtlas.xls' for the food quality predictors considered. The common key across all datasets was the Federal Information Processing Standards (FIPS) County Code. The age adjusted cancer rate data is from 2023.

The poverty predictors considered were poverty count of all ages, poverty percentage of all ages, and median household income. Each of these predictors were based on data from 2023.

The food quality predictors considered were from two sets of data. The first was from a restaurant dataset from 2012 and 2016. The predictors from this dataset were fast food restaurants in 2016 and percentage change in fast food restaurants from 2011 to 2016 with both normalized per 1,000 people. The dollars spent per capita in 2012 on fast food and on full-service restaurants were also considered.

The second food quality predictor dataset was regarding farming data. The predictors considered from this dataset were vegetable, orchard, and berry acre total and per 1,000 people. This data was from 2012. An additional predictor from 2018 in this data was the quantity of farmer's markets per 1,000 people.

***Data Quality***

After data quality exploration, the farming acreage datapoints were removed as imputation would have accounted for up to 26% of the rows. Poverty data had 1 row of missing data, so it was removed. The base dataset had 23 datapoints of missing trend values and trend type which were imputed with 0 and stable trends respectively.

Overall, 2617 out of the possible 3144 counties in the US were retained for logistic regression modeling.

One expected outlier for age adjusted cancer rate was identified and models with and without this datapoint were investigated.

A correlation chart is presented in Figure 1. On the diagonal is the histogram of the variable. The left side of the diagonal is the scatter plot between variables. On the right side of the diagonal is the Pearson correlation coefficient between variables. The response variable is the first row and column for the continuous scaled response. The last row and column represent the binary response. All other plots are predictor variables considered.

***Modeling***

Two paths of logistic regression modeling were considered. The first is with the age adjusted cancer rate per 100k people scaled between 0 and 1. The second was a true binary response by rank ordering the cancer rates with the value of 1 representing the highest 25% of cancer rates while the bottom less than 75% represented a value of 0.

***Statistical Significance***

The models using continuous probability responses resulted in no models with statistically significant predictors at a high level based on p-values. The predictors in models with binary responses were often statistically significant based on p-values. Further inspection of the predictive power proves why even though mathematically this was the case, practically it was not correct.

***Goodness of Fit***

The null deviances and Pearson residual analysis of the models suggests the models were not good fits of the data.

***Predictive Power***

The only potential success experienced is with the accuracy of the 10-fold cross validated models using the binary response. Upon further inspection this was a false success. Based on the confusion matrices examined, the models tended to always predict a value of 0. Using different cutoffs of percentage of the rank, the accuracy moved to match the percentage classed as 0. This, in fact, is not strong predictive power as the True Positives (TP) were always near 0. Accuracy was an overly optimistic metric to use in this case. Accuracy as a metric does not differentiate between TP and True Negative (TN) separately allowing TN to dominate in the accuracy calculation in this example. Consequently, neither approach yielded models with predictive power.

***Results***

The predictors selected were not sufficient to create a model with statistical significance, good fit, and high predictive power. To improve the potential success, additional predictors need to be researched and included. The response variable represented all types of cancer. Dividing the type of cancer could prove useful using smaller subsets of predictors. Another component not considered that should be researched is the effect of the year the predictor data represents versus the year represented by the response variable. In the data sources used, the years of the data were not aligned; years of the data were used based on the datasets identified for use. In fact, one may not want this to be aligned as

cancer rates today could be an effect of predictors in a previous year. For example, the response (cancer rate) is from 2023, while predictors came from many different years. Ultimately, the relative years of predictors and response should be researched and considered beyond the base data used in this initial analysis.

## SES and Healthcare Access Predicting Cancer Incidence with Poisson Regression

### *Methods*

Socioeconomic status and healthcare access and related statistics were analyzed to examine their relationship to cancer incidence. Poverty, unemployment, and education were analyzed as proxies for socioeconomic status. Healthcare access and related statistics were gathered from the PLACES dataset (Centers for Disease Control and Prevention, 2024) and analyzed as well. Preliminary analysis of cancer incidence found that the distribution is slightly left skewed and that there is a single outlier county, Union County, Florida. The mean of cancer incidence is 454 cases per 100,000 and the standard deviation of cancer incidence is 58 cases. Union County's cancer incidence is 1,248 cases per 100,000. In the rest of this section, analysis was performed with this county removed from analysis. See Figure 2 for a histogram of cancer incidence, showing the outlier county.

Many diagnostic figures were created to explore the data and perform preliminary analysis. The authors looked at how rural/urban class of a county, the recent trends of how cancer rate has changed, state, frequency of different educational level attainments, percent of population in poverty, unemployment, and air quality are related to cancer rate. In addition, frequency of disability risk factors, health outcome risk factors, healthy behavior risk factors, and health status risk factors were all evaluated for their relationship with cancer incidence.

Poisson and Quasi-Poisson models were built and analyzed for their goodness of fit and dispersion using both deviance residuals and Pearson residuals. Models were built using all the variables

that were promising after preliminary analysis described above. Two more sets of progressively simpler models with less predictors were generated using variables that seemed to be most important.

***Results***

Rural/urban classification of county showed to have little relationship with cancer incidence. A box plot comparing cancer rate to rural/urban classification is in Figure 3. Recent Trend of Cancer incidence appears to be related to cancer incidence: counties with rising cancer incidence look to have higher cancer rates than counties with decreasing cancer incidence. This makes sense and was expected by the authors. See Figure 4 for a boxplot visualization of cancer incidence trend. Cancer incidence was plotted for all states on a map (in Figure 5) and the authors noticed how there appears to be some regional component of cancer incidence: Counties in the west and southwest appear to have lower cancer incidence than counties in the northeast. A boxplot was generated (in Figure 6) of cancer incidence in each state and this figure confirmed that there is some effect that state has on cancer incidence.

Table 1 provides summary statistics and Pearson correlations to cancer incidence for quantitative variables. In addition, cancer incidence is plotted against different levels of education in counties in Figure 7, against percent of population in poverty in Figure 8, and against unemployment in Figure 9.

The data were broken into a training set comprising 80% of the data set and a testing set comprised of the rest of the data. A Poisson model was fit using rural/urban classification, recent cancer trend, unemployment rate, percent of population in poverty, the education variables, percent of population with any disability, cognitive disability, independent living disability, mobility disability, all teeth being lost, arthritis, COPD, obesity, currently smoking cigarettes, having short sleeps, physically inactive, mental distress, having an annual health checkup, colorectal cancer screening, mammography, high blood pressure medication, and percent of population lacking health insurance. This model failed

goodness of fit testing with a p value of approximately 0, and dispersion using deviance residuals of 362, and dispersion using Pearson residuals of 792. This model was re-fit as a quasi-Poisson model and had the same results.

A simpler model was fit using rural/urban classification, percent of population with all teeth being lost, arthritis, COPD, obesity, currently smoking cigarettes, having short sleeps, physically inactive, mental distress and unemployment rate. This model failed goodness of fit testing again with a p value of approximately 0 and dispersion using deviance residuals of 467, and dispersion using Pearson residuals of 1,180. This model was re-fit as a quasi-Poisson model and had the same results.

An even simpler model was fit using rural/urban classification, percent of population with obesity, currently smoking cigarettes, having short sleeps, physically inactive, and frequent mental distress. This model failed goodness of fit testing again with a p value of approximately 0 and dispersion using deviance residuals of 483, and dispersion using Pearson residuals of 1,197. This model was re-fit as a quasi-Poisson model and had the same results.

**SES and Healthcare Access Predicting Cancer Incidence using Multiple Linear Regression**

***Methods***

Socioeconomic status (SES) and healthcare access and related statistics were analyzed to examine their relationship to cancer incidence. In terms of socioeconomic status, poverty, unemployment, and education, urban influence, were analyzed.

The analysis plan was to use multiple linear regression to create a model that can be used to predict age-adjusted cancer rates using many predictors available in the data sources selected. The data sources include: age-adjusted cancer rates per 100K, level of education by percentage of adults, poverty by county, unemployment rate, and food environment information. The plan was to join several of the datasets together via county (FIPS) code and / or county name.

This approach used a variety of variable selection methods including: using industry knowledge to select variables that seemed appropriate, forward stepwise regression (using AIC and BIC), backwards stepwise regression (using AIC and BIC), and both forward/backward regression (using AIC and BIC).

***Results***

This analysis started with some initial data exploration on the age-adjusted cancer rate cases per 100,000 (hereafter cancer rate). There was one outlier county in Florida that was well outside the rest of the data as seen in Figure 2. Once removed the distribution of cancer rate was still a little skewed to the right side but generally looked a lot better as shown in Figure 10 where this county was removed.

Categorial variables were then examined to see if any were good differentiators for cancer rates and unfortunately there was not much separation even for predictors expected like rural/urban split as the journals we had researched seemed to indicate a difference in cancer rates for rural vs. urban areas (Henley et al., 2017). But there were also a lot of outliers within the boxplots so it was hard to get a good read as shown in Figure 11.

The US states predictor did show some nice separation, but it was a bit hard to get a clean read from the data but provided us an idea of at least one variable that might make its way into the model as shown in Figure 6. Numeric variables were then analyzed to see how they looked compared to cancer rates. The data did not look very good as there was a lot of clustering with some outliers and the r-squared for things looked dismal as seen in Figure 12. An initial model using 5 variables based on industry knowledge was then created. The model was not very good with an adjusted r-squared of 0.047.

When looking at the assumptions for the model the normality assumptions really stood out. The qqplot of residuals, seen in Figure 13, was very heavy tailed. The histogram did not look that bad but in total it felt like this was a concern.

A Box-Cox transformation was attempted to see if transformations would improve the model. This effort yielded some success but still resulted in heavy tails with a bit of a skew in the histogram. The transformation did not significantly improve the r-squared value and consequently was dropped from further models.

A few more attempts were made manually selecting different variables, but the results did not yield significant improvements worth further exploration.

The decision was then made to try using forward stepwise regression, backward stepwise regression and forward/backward stepwise regression using both AIC and BIC to see if they could provide better models than obtained previously.

All the different models had come out with adjusted r-squared between 0.4 and 0.48, indicating similar explanatory power. To optimize for model simplicity, forward stepwise regression using BIC was pursued further.

Using these processes, models were created that had much better adjusted r-squared but still had issues with failing goodness of fit tests. The diagnostic plots for goodness of fit testing are in Figure 14. The normality check was still not within the optimal range for normal, shown in Figure 15.

The final variables included were: US state, recent trend, percent of population Hispanic in 2010, SNAP benefits per capita from 2017, estimated percent of people of all ages in poverty 2023, rural/urban Indicator, percent of American Indian or Alaska Native, percent with low access to stores in 2015, unemployment rate from 2005, percent of population under age 18 in 2010.

With these variables, the model returned an adjusted r-squared of 0.4603. However, besides failing most of the goodness of fit tests, it was using data across different years, and the final model did not seem explainable or that it could be extrapolated to data in the future. For example, this model has a response variable from 2023 and predictors from 2017, 2015, 2010, and 2005. If this model were used

to predict a response in 2024, it's unclear that predictors from these years in particular would have the same predictive power that they do for 2023.

**Predicting Cancer Incidence with Elastic Net Regression**

***Methods and Results***

Elastic net regression with cross-validation has been found to be better than its other regularization counterparts, LASSO & ridge regression, as it combines the ridge shrinkage features in conjunction with the lasso-regression type thresholding, while gaining the benefit of generalization from cross-validation (Zou & Hastie, 2005). Additionally, Zhou & Hastie discuss the superiority of elastic net and multicollinearity by addressing the limitations LASSO and ridge regression both run into with respect to the calculated lambda values. Elastic net employs a hybrid approach with the use of lambda values. These two issues are limited by the use of elastic net regression because it includes the ridge regression penalty term to account for the limitations of LASSO regression (Freijeiro-González et al., 2022). To organize the specific elastic net methods used in this portion of the project, each of the following sub-headings will describe their respective method. Furthermore, the data used is described in the referenced figures for each of the subheadings.

*0.5-Alpha 0.01-Lambda Elastic Net Regression*

Beginning this elastic net analysis, a simple version of the elastic net regression model was used. An alpha of 0.5 was selected as it served as the midpoint between LASSO regression (alpha = 0) and ridge regression (alpha = 1). The lambda choice of 0.01 was arbitrary as a new model was later constructed that calculated the optimal lambda for the regression. This preliminary model generated results on the test set of 3119.596 for MSE, and a 0.0693 R-Squared. The coefficient path seen in Figure 18 details that all the coefficients used experience minimal shrinking due to the extremely small lambda value. These results are considered a baseline for which all the other optimizations will be compared.

*0.5-Alpha 10-Fold CV Elastic Net Regression*

The next step towards improving the elastic net method involved implementing a 10-fold cross-validation step to find the optimal lambda. Cross-validation in conjunction with an elastic net regression is used to account for overfitting and the model's generalization ability (Zou & Hastie, 2005). The determined lambda value of 0.89 was optimized by the model to select for the lowest MSE generated. By combining these two procedures, the model improved as the MSE decreased, and the R-Squared increased, as demonstrated in Table 2. The coefficient path in Figure 17, as well as Table 3 describe the values of the coefficients and their effect on the MSE.

*0.38-Alpha 10-Fold CV Elastic Net Regression (Alpha Optimization)*

The final step towards improving the elastic net method involved selecting the best performing alpha value. Calculating this alpha value was done by iterating through alphas of x/50, where x starts at 1 and increases by one until x=50. An elastic net regression with all of the features of the previous model were run for each alpha value, and the alpha value with the lowest MSE and highest R-Squared was chosen. The culmination of all of these elastic net optimizations resulted in the best performing elastic net model. This model did improve as the MSE decreased, and the R-Squared increased, giving values of 3110.888 and 0.0719, respectively. The coefficient path in Figure 16, as well as Table 3 describe the values of the coefficients and their effect on the MSE. Overall, the coefficients were slightly altered when the calculated alpha of 0.38 was used.

***Overall Elastic Net Analysis***

With respect to the entire elastic net analysis, it seems that the model did underfit some of the predictors. For example, one of the data sources involved the age adjusted cancer/melanoma rate among adults. Common knowledge would deem this as a very important factor for determining cancer rate (as those counties with a higher diagnosed cancer/melanoma rate would experience similar predicted cancer rates), however the elastic net regression set this coefficient to only have the fourth largest effect size for the best model, as seen in Table 3. Also, between the four educational

components of the model, there cannot be any conclusions drawn from why the model chose to shrink two of them to zero and leave the other two with a positive effect size and a negative effect size. Furthermore, cancer rate prediction is an extremely complex topic and selecting a few predictors to do this would be near impossible given its still unknown root causes (Zhang, 2017). Although the MSE and the R-Squared do not support a satisfactory prediction due to the complex nature of cancer prediction, the model did support conclusions made from other studies with respect to poverty and socioeconomic status. Finally, each of the tables and graphs have captions that further describe the results generated from this analysis.

## Explanation of Changes

### Logistic Regression

The approach outlined in the analysis plan stayed the same for the logistic regression analysis with one exception. An additional dataset ("FoodEnvironmentAtlas") was used to account for fast food density and access to farmer's markets as discussed in the logistic regression details section of this paper.

### SES and Healthcare Access Predicting Cancer Incidence Using Poisson Regression

Originally, this analysis was going to incorporate an analysis of air quality index (AQI) data, but the data set missed data for approximately two thirds of counties. Because of this, the AQI analysis was discarded. Additionally, a multiple linear regression analysis was proposed but this was replaced with a Poisson model to account for the fact that the response variable is a rate instead of a count of things.

### SES and Healthcare Access Predicting Cancer Incidence using Multiple Linear Regression

The original plan was to use the air quality and "PLACES Local Data for Better Health" (PLACES) data sources as a part of this analysis. After preliminary analysis, the air quality data was too sparse to be of use and the PLACES dataset was very complicated. An additional dataset that was not discussed originally (FoodEnvironmentAtlas) was included. The analysis plan did not specify the variable selection

methodologies that would be used. Forward and backward stepwise (and combination forward/backward) regression with both AIC and BIC were used. Different subsets of the variables, only using certain of the datasets, as well as all the variables together was not originally discussed in the analysis plan but did end up being used in the analysis. Although data imputation was mentioned in the analysis plan, it was not needed because the vast majority of counties were not missing data used in this analysis. In the few instances where FIPS code was missing the decision was made to just move forward with instances where the FIPS code was in both/all sources (and it was for over 99% of FIPS codes).

**Predicting Cancer Incidence with Elastic Net Regression**

The only change from the initial analysis plan was the exclusion of the air quality data. This dataset only accounted for 38.27% of all counties in the other datasets. Use of this dataset would dramatically decrease the sample size, so it was decided it would not be included. In addition, MSE was used over root mean square error due to having less sensitivity to outliers.

## Conclusions

As stated previously, the implications of obtaining a more accurate prediction of cancer incidence by county could result in targeted public health campaigns, better allocation of resources, and a reduction in cancer occurrence by directing more care towards at-risk populations. Although this study did not find any sufficient method to predict cancer incidence, there are key points in this study that can be extrapolated to its successes and limitations.

**Successes and Limitations**

In the context of cancer incidence prediction rate, the most general method of multiple linear regression obtained the highest r-squared value of 0.4603 utilizing forward stepwise regression with BIC as the measure. This was the best performing model by far for predictive power; however, it failed all the goodness of fit tests and lacked explanatory power since it used data across several different years with no correlation as to why it was chosen. Additionally, the multiple linear regression, and the other

analyses, suffered from the problem of complexity and individuality for cancer rate prediction on the county level. Therefore, modeling this rate based on the selection of a few predictors would be near impossible, given its complicated nature.

With respect to the overall problem of modeling cancer incidence rates, the elastic net regression and the preliminary analysis supported several correlations that align with current literature on the topic. For example, the method proposed by Dong et al. that found correlations between cancer and poverty were supported by the elastic net model, as the poverty rate and household income had a significant effect on the prediction. So, although the MSE and the r-squared do not support a satisfactory prediction due to the nature of cancer prediction, the model did support conclusions made from other studies, with respect to poverty and socioeconomic status.

Overall, the methods were limited by the data that was chosen to predict cancer incidence. This data was not suitable for several of the analyses explored in this study. This effect can be seen in both the Poisson and the logistic regression models as it does not follow a Poisson distribution, nor did it contain data in the form of a binary response. And as a result, the goodness of fit tests and the evaluation metrics (i.e. deviances, overdispersion, and classification scores) did not support the possibility that any conclusions could be drawn from these methods.

**Next Steps**

Further analyses should consider employing robust regression to compensate for the relatively symmetric distribution of residuals with heavy tails. Another analytical method to examine in the future could also use regularized regression to improve the poor performance of stepwise regression. This method should be examined in the future because there are many variables that may be used in analyses. Future analyses could also look at how the data are prepared and use trends in predictors over time as a predictor to examine how the history of a county may impact its cancer incidence.

In addition to these analytical next steps, a further analysis should be performed on the outlier county identified in each sub analyses reported above. Union County Florida has been found by others to have high cancer rates among many different types of cancer (Mokdad et al., 2017). More investigation needs to be done to see why this county has such high cancer rates.

It's important to note that these analyses are not at an individual level and are instead at a county level. This means that conclusions about these analyses can't be extrapolated to say anything about a single individual's risk of developing cancer but instead the conclusiosn drawn from this analysis can say something about how cancer is distributed amongst counties in the United States.

Last, individual cancer types should also be examined. While others have attempted to do this (Henley et al., 2017; Mokdad et al., 2017), more work to examine how cancer incidence may be predicted and evaluated should be undertaken.

References

Centers for Disease Control and Prevention. (2024). *PLACES: Local Data for Better Health, County Data* [Dataset]. https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb/about_data

Clegg, L. X., Reichman, M. E., Miller, B. A., Hankey, B. F., Singh, G. K., Lin, Y. D., Goodman, M. T., Lynch, C. F., Schwartz, S. M., Chen, V. W., Bernstein, L., Gomez, S. L., Graff, J. J., Lin, C. C., Johnson, N. J., & Edwards, B. K. (2009). Impact of socioeconomic status on cancer incidence and stage at diagnosis: Selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. *Cancer Causes & Control*, *20*(4), 417–435. https://doi.org/10.1007/s10552-008-9256-0

Dong, W., Bensken, W. P., Kim, U., Rose, J., Fan, Q., Schiltz, N. K., Berger, N. A., & Koroukian, S. M. (2022). Variation in and factors associated with US county-level cancer mortality, 2008-2019. *JAMA Network Open*, *5*(9), e2230925–e2230925.

Economic Research Service, US Department of Agriculture. (2025). *County-level Economic Data* [Dataset]. https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data

Environmental Protection Agency. (2024a). *Air Quality—Cities and Counties* [Air Quality - Counties]. https://www.epa.gov/air-trends/air-quality-cities-and-counties

Environmental Protection Agency. (2024b). *Annual Air Quality Index by County* [Data & Tools]. EPA Air Data. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *International Statistical Review*, *90*(1), 118–145. https://doi.org/10.1111/insr.12469

Henley, S. J., Anderson, R. N., Thomas, C. C., Massetti, G. M., Peaker, B., & Richardson, L. C. (2017). *Invasive cancer incidence, 2004–2013, and deaths, 2006–2015, in nonmetropolitan and metropolitan counties—US*. https://stacks.cdc.gov/view/cdc/46535

Liu, B., Zhu, L., Zou, J., Chen, H.-S., Miller, K. D., Jemal, A., Siegel, R. L., & Feuer, E. J. (2021). Updated methodology for projecting US- and State-level cancer counts for the current calendar year: Part I: Spatio-temporal modeling for cancer incidence. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *30*(9), 1620–1626. https://doi.org/10.1158/1055-9965.EPI-20-1727

Mokdad, A. H., Dwyer-Lindgren, L., Fitzmaurice, C., Stubbs, R. W., Bertozzi-Villa, A., Morozoff, C., Charara, R., Allen, C., Naghavi, M., & Murray, C. J. L. (2017). Trends and Patterns of Disparities in Cancer Mortality Among US Counties, 1980-2014. *JAMA*, *317*(4), 388–406. https://doi.org/10.1001/jama.2016.20324

Moss, J. L., Wang, M., Liang, M., Kameni, A., Stoltzfus, K. C., & Onega, T. (2021). County-level characteristics associated with incidence, late-stage incidence, and mortality from screenable cancers. *Cancer Epidemiology*, *75*, 102033.

National Cancer Institute. (n.d.). *Joinpoint*. Joinpoint Help System. Retrieved March 16, 2025, from https://surveillance.cancer.gov/help/joinpoint

National Cancer Institute, & Centers for Disease Control and Prevention. (2024). *County Cancer Profiles Incidence Rates Table* [Dataset]. https://www.statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=00&areatype=county&cancer=001&race=00&sex=0&age=001&type=incd&sortVariableName=rate&sortOrder=default&output=0#results

O'Neil, M. E. (2019). Lung cancer incidence in nonmetropolitan and metropolitan counties—United States, 2007–2016. *MMWR. Morbidity and Mortality Weekly Report*, *68*. https://www.cdc.gov/mmwr/volumes/68/wr/mm6844a1.htm

Shi, L., Macinko, J., Starfield, B., Politzer, R., Wulu, J., & Xu, J. (2005). Primary Care, Social Inequalities, and All-Cause, Heart Disease, and Cancer Mortality in US Counties, 1990. *American Journal of Public Health*, *95*(4), 674–680. https://doi.org/10.2105/AJPH.2003.031716

Tabuchi, T. (2020). Cancer and socioeconomic status. *Social Determinants of Health in Non-Communicable Diseases: Case Studies from Japan*, 31–40.

US Department of Agriculture. (2024). *Food Environment Atlas* [Dataset]. https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads

Withrow, D. R., Berrington de González, A., Spillane, S., Freedman, N. D., Best, A. F., Chen, Y., & Shiels, M. S. (2019). Trends in mortality due to cancer in the United States by age and county-level income, 1999–2015. *JNCI: Journal of the National Cancer Institute*, *111*(8), 863–866.

Yu, B. (2013). Predicting county-level cancer incidence rates and counts in the USA. *Statistics in Medicine*, *32*(22), 3911–3925. https://doi.org/10.1002/sim.5833

Zhang, F. (2017). Windows of developmental susceptibility in reproduction and cancer. In *Translational toxicology and therapeutics*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=4yA-DwAAQBAJ&oi=fnd&pg=PT10&dq=Translational+Toxicology+and+Therapeutics:+Windows+of+Developmental+Susceptibility+in+Reproduction+and+Cancer&ots=zPEfrDA3_i&sig=gvx4_x5YiqzOz_hqJQu7ofPOWyc

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

# Appendix

## Contributions

### *Chris Friedman*

Chris Friedman ran the analysis of SES and Healthcare Access predicting Cancer Incidence Using Poisson Regression. He performed data exploration and produced figures and then ran progressively simpler Poisson and Quasi-Poisson models. He also found all the data sets used within this larger group of analyses described throughout this paper. Chris compiled this report from individual authors. Chris's work was completed using rstudio and rmarkdown and is available at:

https://github.gatech.edu/ISyE6414-SP2025/zealous_rogues/tree/main/cfriedman32

### *Michael Strange*

Michael focused on using Multiple Linear Regression in attempting to examine the relationship between the variables we had and cancer incidence. He started with a handful of self-selected variables for a few different attempts and ended up with very weak models. He then proceeded to try a handful of different variable selection techniques (forwards and backwards stepwise regression using AIC and BIX) and ended up with models with a stronger r-squared value, but all the models failed the normality assumption and did not look great for some of the other assumptions. The qqplot looked like one of the charts in Module 4 where the professor stated the distribution was "symmetric, but with heavy tails, possibly, more of a T distribution than a normal distribution". A suggestion was to use robust regression instead of least square regression but that is a more advanced topic that we did not have the knowledge or time to investigate further. His work was done in a Jupyter notebook through Anaconda Navigator. His scripts and data sources are located at:

https://github.gatech.edu/ISyE6414-SP2025/zealous_rogues/tree/main/mstrange30

If the notebook and datasets are downloaded to the same directory, then the code should run fine if you have all the packages used in the example codes provided in class loaded to your version of R.

Because he tried both forward and backward (and combination forward/backward) stepwise regression it can take a while (10-15 minutes) for the entire code to run but it does successfully complete.

***Toshan Doodnauth***

The method of elastic net regression produced results that did not satisfactorily explain the proportion of total variability in the cancer incidence rate. In addition, the mean square error (MSE) was too large to draw any strong conclusions from. In order to run the elastic net regression, all of the CSVs were analyzed to determine which columns are appropriate to use in this regression. Initially, the air quality dataset was to be used in the project, however the amount of data it had for each county was too sparse to be included. The observations (as seen in the R code) had only accounted for 38.27% of all counties in the other datasets. Use of this dataset would dramatically decrease the sample size, so it was decided that it would not be included. Next, some manipulation of county names and joining of another dataset was performed so that the PLACESTest.csv has corresponding FIPS codes to the respective counties. This was done so that the PLACES dataset could be joined together with the other datasets later in the analysis. Data selection was then executed for each dataset, this included the FIPS codes and the selected columns in the first step. All of the datasets were then joined by the FIPS codes, scaled, and split into a training and testing set, with the ratio of 8:2, respectively. The final joined dataset consisted of 2929 counties, each having 10 features (Age_Adjusted_Incidence_Rate_cases_per_100.000 (**Predictor**), percent_of_adults_who_are_not_high_school_graduates_2019_23, percent_of_adults_who_are_high_school_graduates_or_equivalent_2019_23, Percent_of_adults_completing_some_college_or_associate_degree_2019_23, percent_of_adults_with_a_bachelors_degree_or_higher_2019_23, Age_adjusted_Cancer_or_melanoma_among_adults_rate, PCTPOVALL_2023, Unemployment_rate_2023, Med_HH_Income_Percent_of_State_Total_2022, CENSUS_2020_POP). Finally, elastic net regression was run on the joined dataset with several different alpha values, and the

lambda value that minimizes MSE. First, an alpha of 0.5 was used as it is the midpoint between the bounds of alpha (0 being ridge regression and 1 being LASSO regression). Next, cross-validation was added to improve model accuracy. Finally, a stronger alpha value was determined by running the model at increasing iterations of alpha equaling 1/50. The table below outlines the results gathered. Overall, the elastic net regression performed poorly, most likely due to the model underfitting the 8 predictors of age-adjusted cancer incidence rate. This can be seen in the R code as all predictors were minimized 0, except for percent_of_adults_who_are_high_school_graduates_or_equivalent_2019_23 at 1.923474. https://github.gatech.edu/ISyE6414-SP2025/zealous_rogues/tree/main/tdoodnauth3/.

***Joshua Geiger***

This team member explored the potential of logistic regression as a method of meeting the teams objective of being able to predict cancer rate at the county level in the United Status using various social and economic predictors. Both continuous probability and binary response types were explored. Model assumptions, statistical significance, quality of fit and predictive power were examined using the suggested approaches from the course lectures. This team member also set up a Microsoft Teams for the group to enable O365 based collaboration and meeting scheduling. The R kernel jupyter notebooks for this team member can be found at the git link below. This team member also showed a passion for ending meetings with a collective THWG.

https://github.gatech.edu/ISyE6414-SP2025/zealous_rogues/tree/main/jg318

## Tables

**Table 1**

*Correlation of SES and Health-Related Predictors to Cancer Incidence*

| | Pearson Correlation | Mean | Standard Deviation |
|---|---|---|---|
| Disability Risk Factors | | | |
| Any Disability | 0.09 | 33.19 | 5.70 |
| Cognitive Disability | 0.10 | 16.63 | 3.01 |
| Hearing Disability | 0.06 | 7.42 | 1.22 |
| Independent Living Disability | 0.08 | 9.09 | 2.02 |
| Mobility Disability | 0.09 | 14.24 | 3.34 |
| Self-care Disability | 0.02 | 4.14 | 1.23 |
| Vision Disability | -0.05 | 5.87 | 1.77 |
| Health Outcomes Risk Factors | | | |
| All Teeth Lost | 0.17 | 16.47 | 6.02 |
| Arthritis | 0.29 | 26.21 | 3.14 |
| Cancer (non-skin) or Melanoma | 0.19 | 6.99 | 0.59 |
| COPD | 0.25 | 7.65 | 1.87 |
| Coronary Heart Disease | 0.13 | 6.55 | 0.91 |
| Current Asthma | 0.12 | 10.89 | 1.00 |
| Depression | 0.19 | 24.57 | 3.56 |
| Diabetes | 0.04 | 11.33 | 2.35 |
| High Blood Pressure | 0.19 | 33.02 | 4.87 |
| High Cholesterol | 0.11 | 30.80 | 2.34 |
| Obesity | 0.21 | 38.29 | 4.60 |
| Stroke | 0.10 | 3.47 | 0.68 |
| Health Behavior Risk Factors | | | |
| Binge Drinking | -0.05 | 18.48 | 3.06 |
| Current Cigarette Smoking | 0.22 | 18.06 | 3.93 |
| Physical Inactivity | 0.10 | 26.82 | 5.23 |
| Short Sleep Duration | 0.15 | 37.38 | 4.11 |
| Health Status Risk Factors | | | |
| General Health | 0.04 | 20.03 | 4.89 |
| Frequent Mental Distress | 0.11 | 18.74 | 2.26 |
| Frequent Physical Distress | 0.06 | 13.85 | 2.35 |
| Social Isolation | -0.13 | 34.43 | 2.44 |

| | | | |
|---|---|---|---|
| Food Insecurity | -0.03 | 16.60 | 6.20 |
| Housing Insecurity | -0.07 | 14.40 | 4.34 |
| Transportation Barriers | -0.03 | 9.87 | 2.83 |
| Lack of Social/Emotional Support | -0.08 | 26.04 | 4.00 |
| Food Stamps | 0.03 | 15.29 | 6.26 |
| Utility Services Threat | 0.02 | 9.75 | 3.11 |
| Health Needs Risk Factors | | | |
| Cholesterol Screening | 0.06 | 80.95 | 3.83 |
| Colorectal Cancer Screening | 0.22 | 57.16 | 4.54 |
| Health Insurance | -0.18 | 12.62 | 6.25 |
| Mammography | 0.16 | 73.10 | 4.45 |
| High Blood Pressure Medication | 0.33 | 60.21 | 3.99 |
| Dental Visit | -0.07 | 56.15 | 11.98 |
| Annual Checkup | 0.35 | 73.43 | 3.64 |
| Socio-Economic Status | | | |
| Unemployment | -0.03 | 3.70 | 1.30 |
| Poverty | 0.05 | 14.82 | 5.61 |
| Percent with No High School Diploma | -0.03 | 11.65 | 5.69 |
| Percent with High School Diploma or Equivalent | 0.23 | 33.92 | 7.51 |
| Percent with Some College or Associate Degree | -0.08 | 30.97 | 5.09 |
| Percent with Bachelors Degree or Higher | -0.12 | 23.46 | 9.84 |

*Table 2*

This table describes the results generated from all of the elastic net models. With each optimization, it can be seen that the MSE and the R-Squared were improved.

| Method | MSE | R-Squared | Alpha |
|---|---|---|---|
| 10-Fold CV Elastic Net with Calculated Alpha | 3110.888 | 0.07191892 | 0.38 |
| 10-Fold CV Elastic Net | 3112.637 | 0.07139686 | 0.5 |
| Elastic Net | 3119.596 | 0.06932086 | 0.5 |

*Table 3*

This table displays all of the coefficients generated from each of the elastic net models. As the model was optimized, all of the coefficients trended toward zero. This implied that the baseline model overfitted each of the predictors. In addition, this table describes the data that was used to predict age-adjusted cancer incidence per 100k. All of the data was scaled before implementation of the elastic net model.

| Variable | Coefficients of 10-Fold 0.38-Alpha | Coefficients of 10-Fold 0.5-Alpha | Coefficients of 0.5-Alpha |
|---|---|---|---|
| (Intercept) | 454.65 | 454.65 | 454.64 |
| percent_of_adults_who_are_not_high_school_graduates_2019_23 | 0.00 | 0.00 | -0.11 |
| percent_of_adults_who_are_high_school_graduates_or_equivalent_2019_23 | 15.57 | 15.82 | 17.17 |
| percent_of_adults_completing_some_college_or_associate_degree_2019_23 | 0.00 | 0.00 | 1.24 |
| percent_of_adults_with_a_bachelors_degree_or_higher_2019_23 | -0.18 | -0.35 | 0.00 |
| Age_adjusted_Cancer_or_melanoma_among_adults_rate | 7.20 | 7.67 | 8.73 |
| PCTPOVALL_2023 (Poverty Rate 2023) | 6.75 | 7.50 | 9.94 |
| Unemployment_rate_2023 | 0.44 | 0.53 | 0.86 |
| Med_HH_Income_Percent_of_State_Total_2022 | 9.33 | 10.13 | 12.37 |
| CENSUS_2020_POP(Population in 2020) | 0.42 | 0.69 | 1.62 |

## Figures

*Figure 1 - Correlation Chart of Variables Considered in Logistic Regression of Problem*

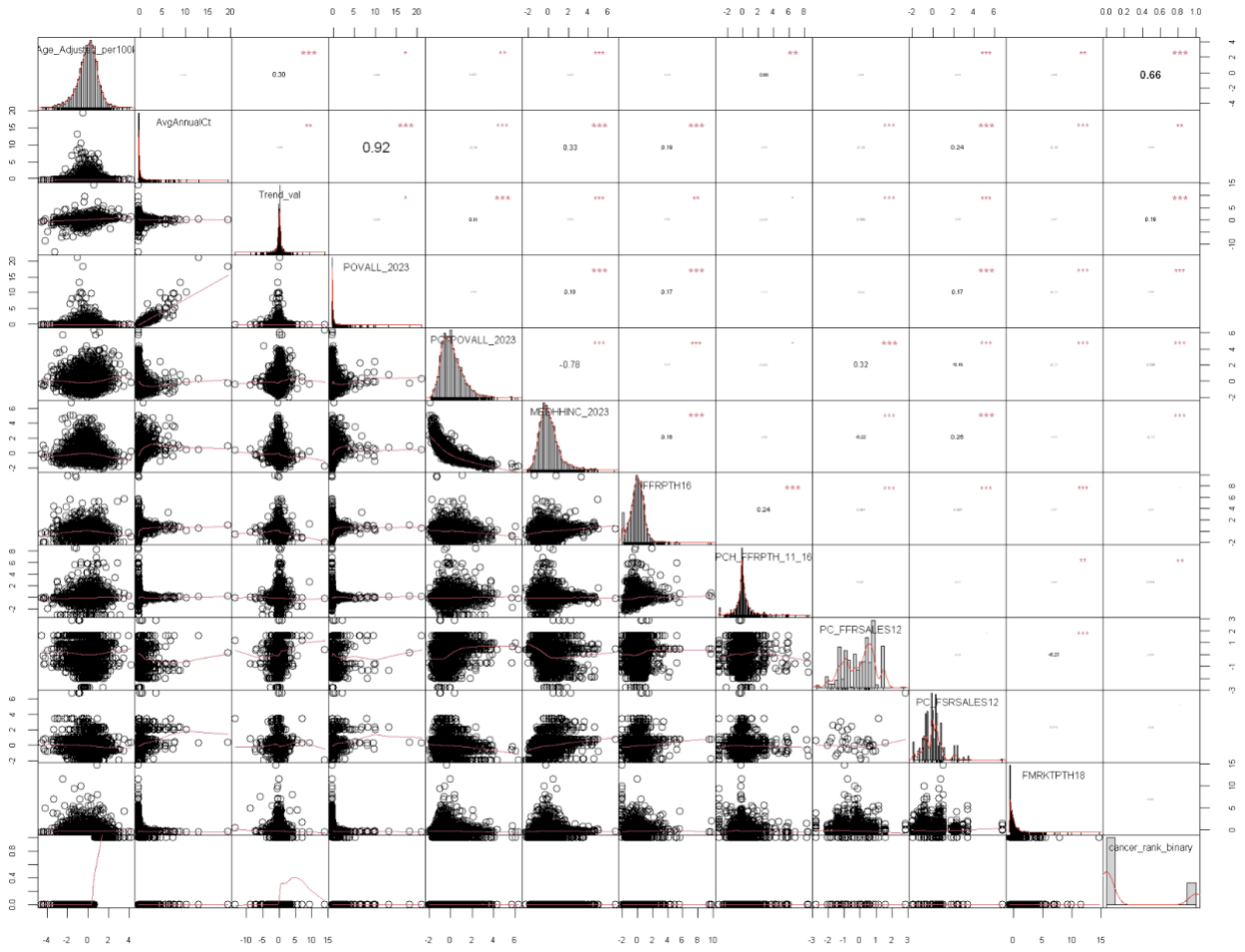

*Figure 2 – Distribution of Cancer Rate Before Outlier Removal*

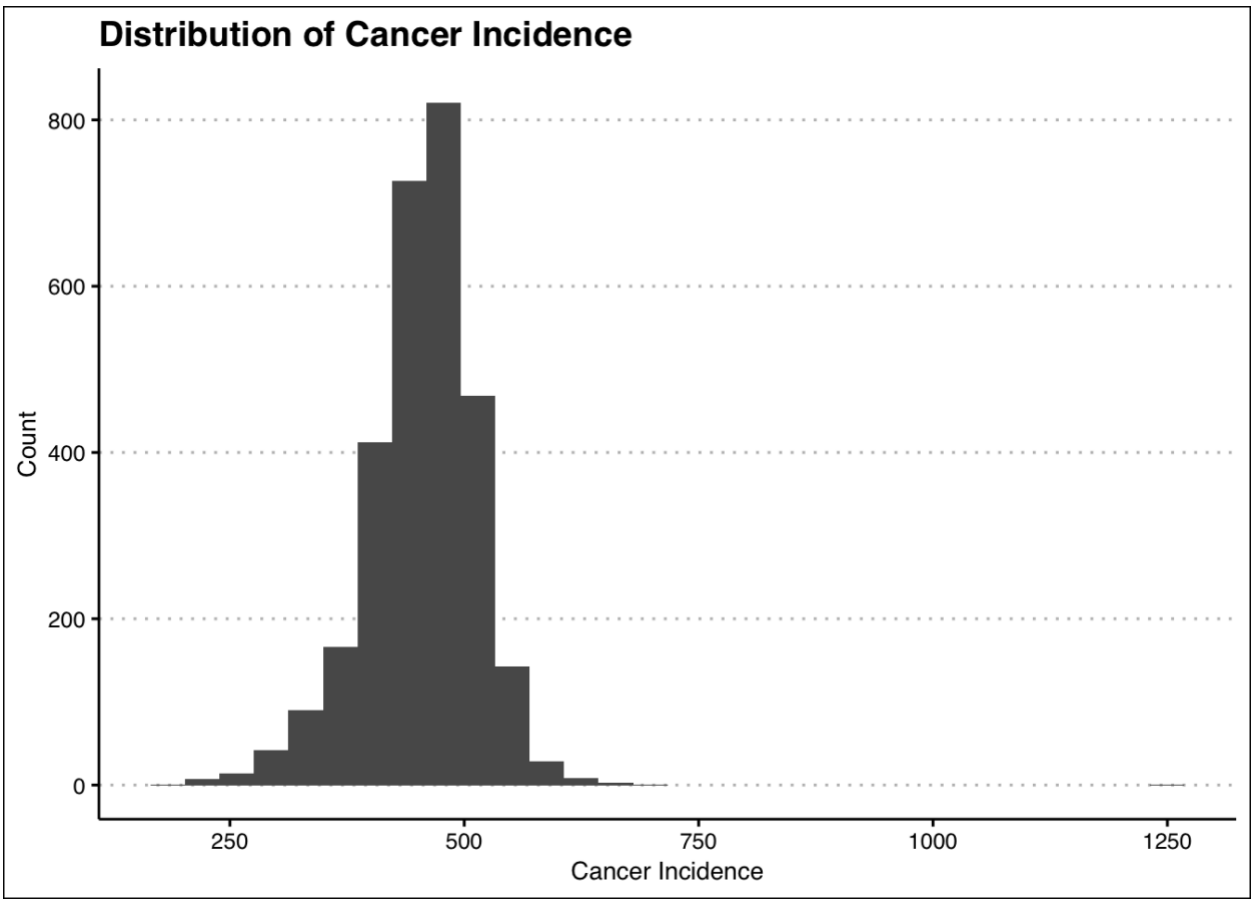

*Figure 3 – Boxplot of Cancer Rate compared to Rural/Urban County Classification*

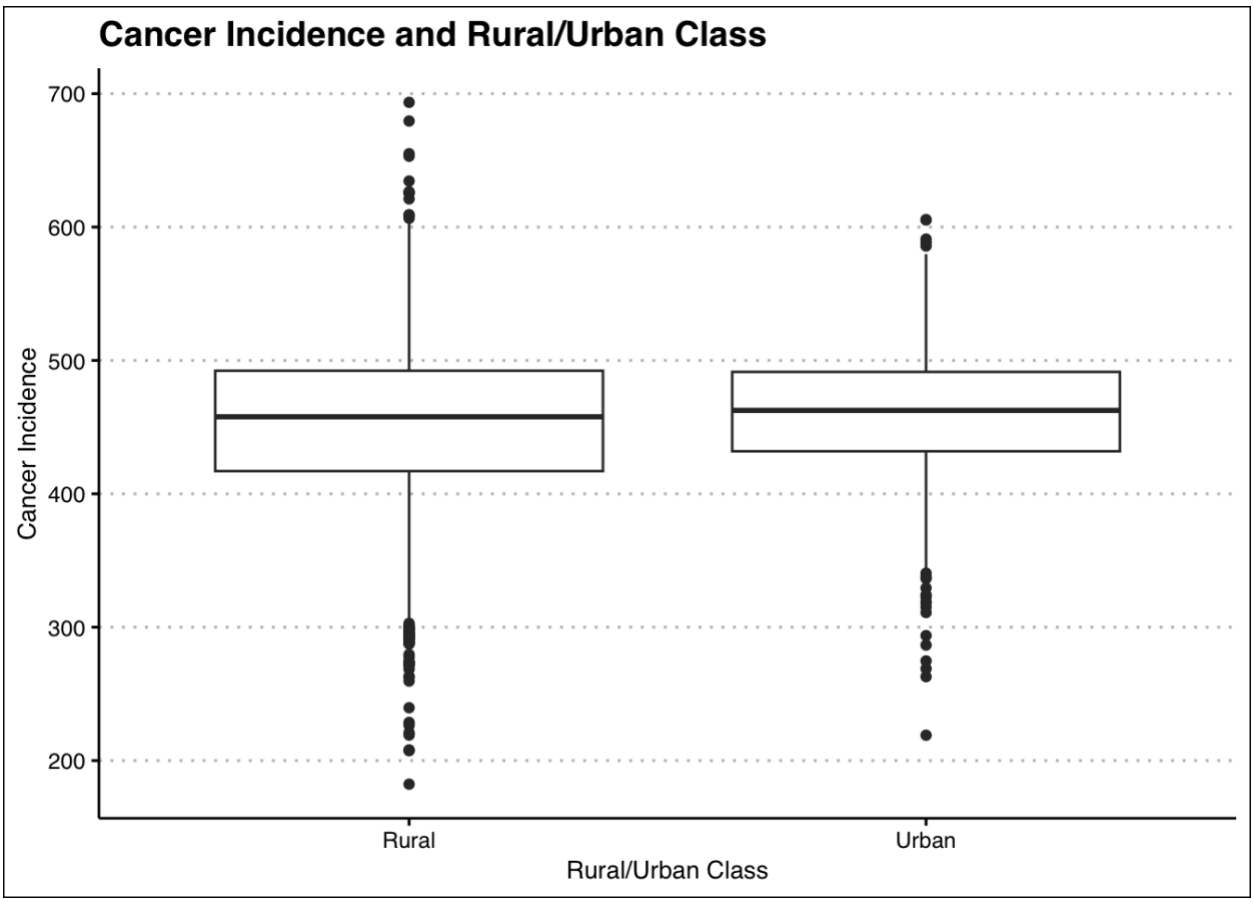

*Figure 4 – Boxplot of Cancer Rate compared to Recent Trend in Cancer Rate*

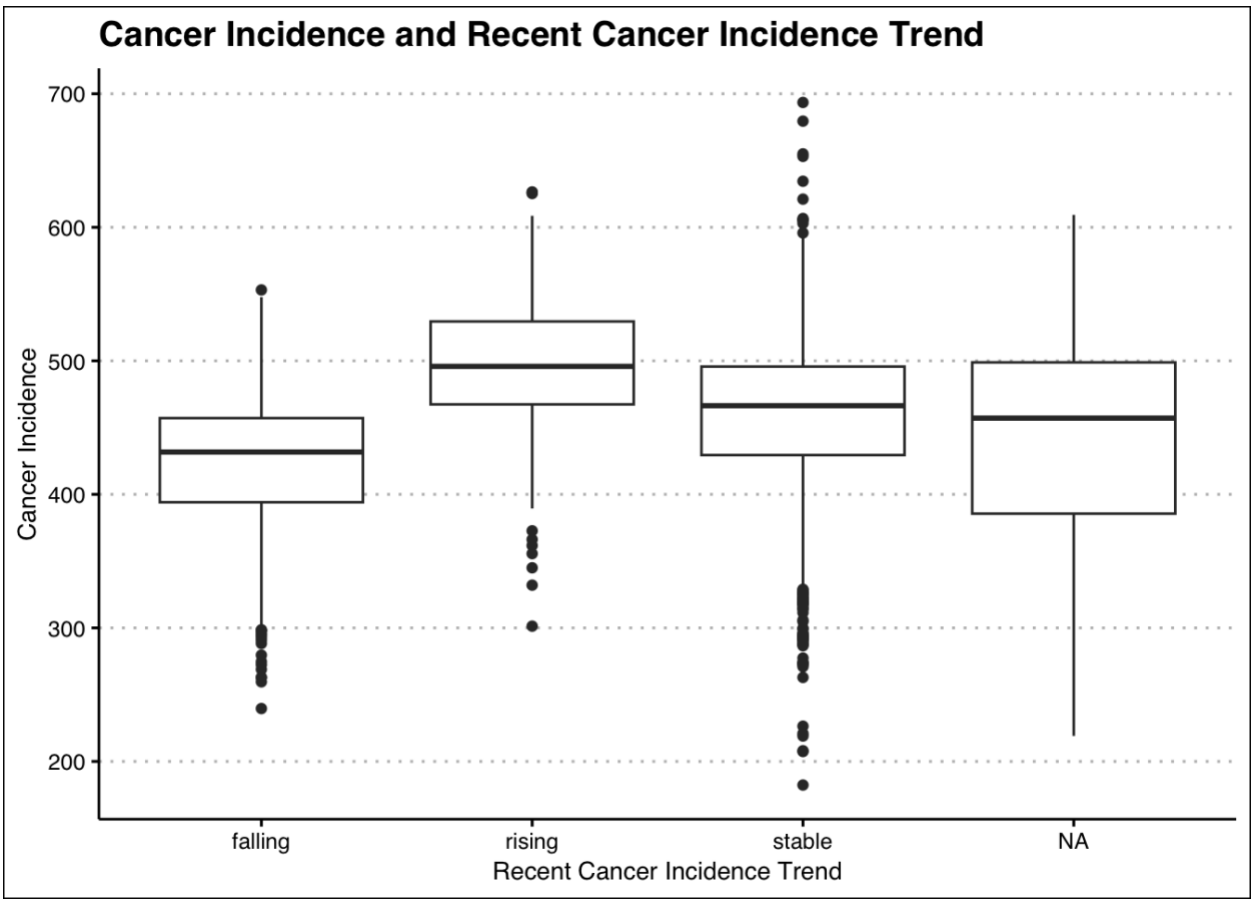

*Figure 5 – Map of the United States with Counties Colored by Cancer Rate*

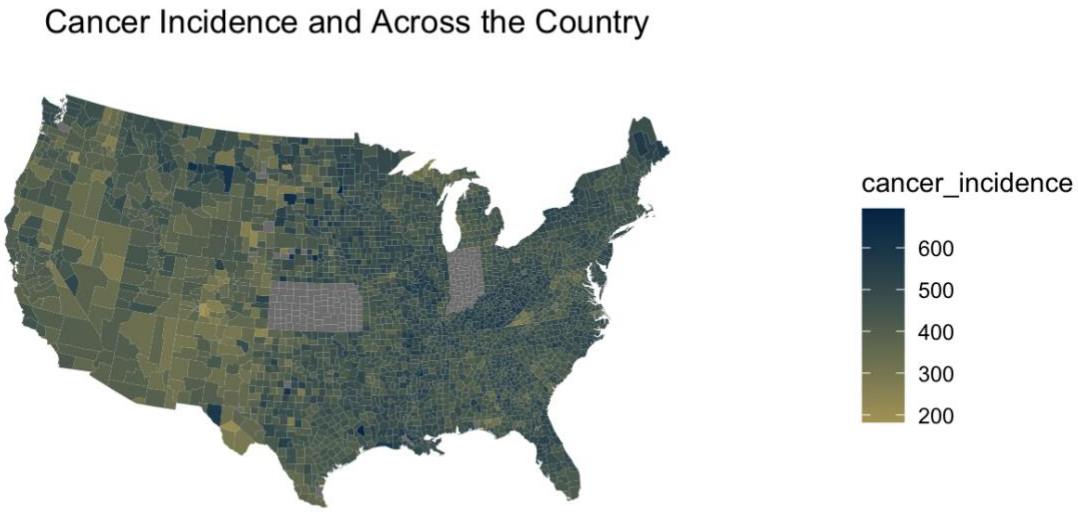

*Figure 6 – Boxplot of Cancer Rate Compared by State*

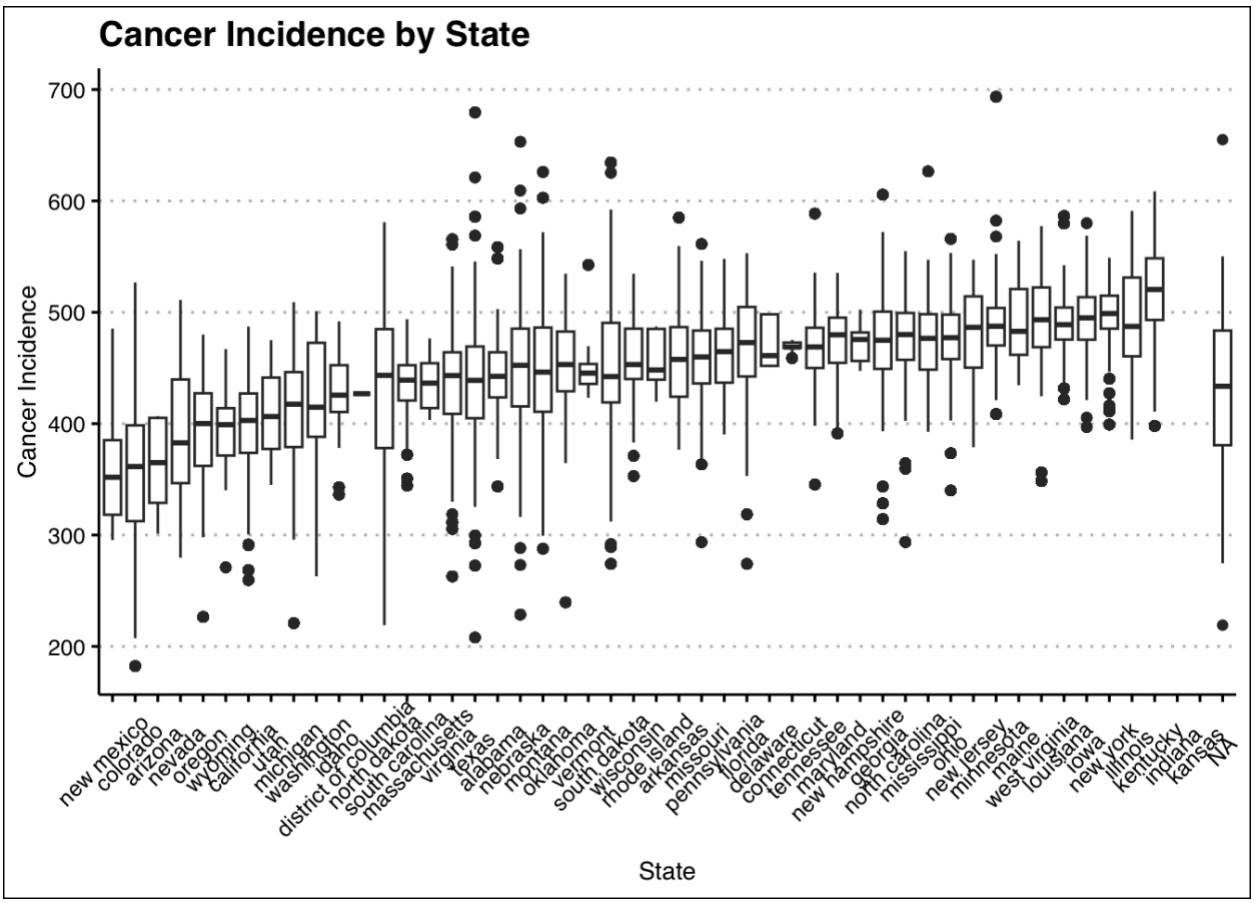

*Figure 7 – Cancer Rate Compared to Percent of Population Having Attained Different Education Levels*

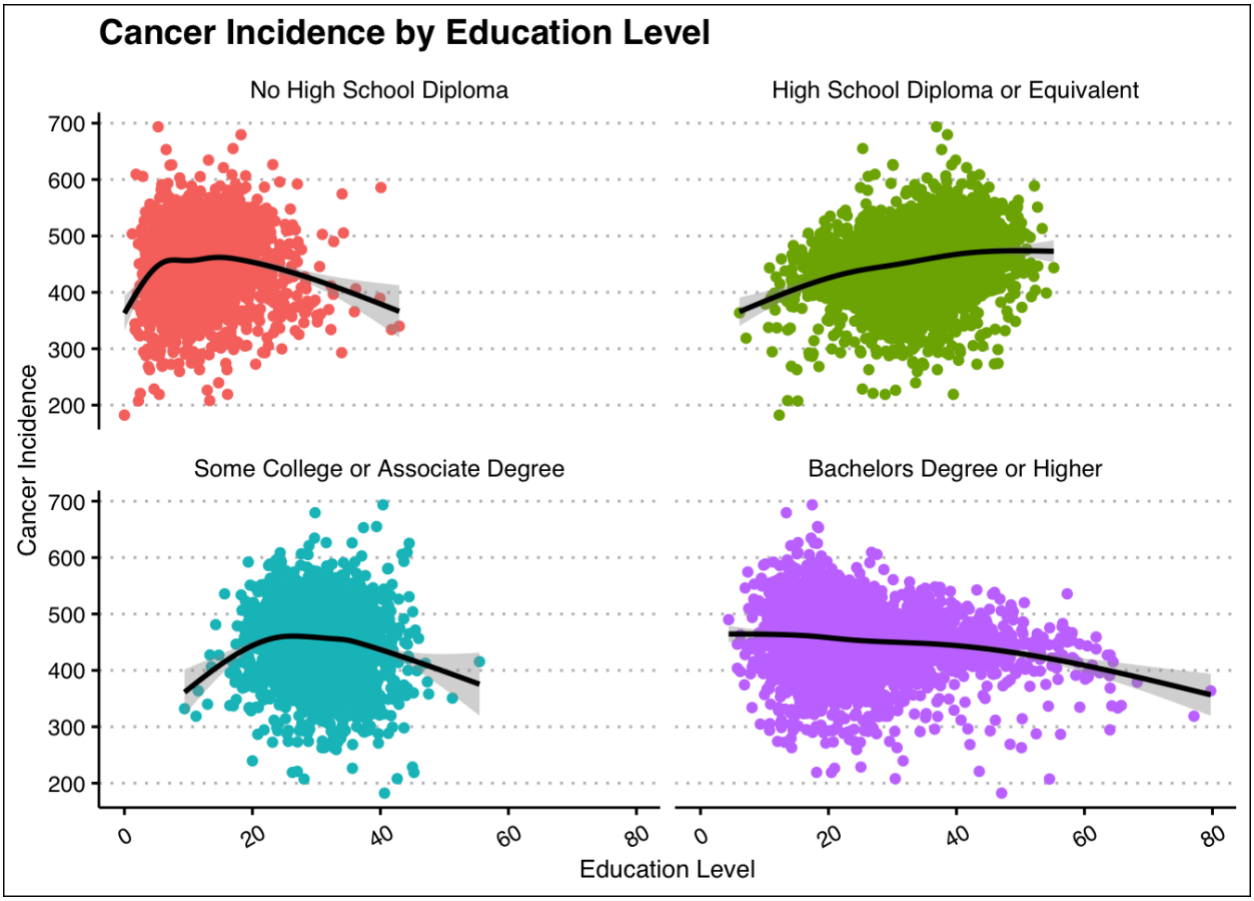

*Figure 8 – Cancer Rate Compared to Percent of Population in Poverty*

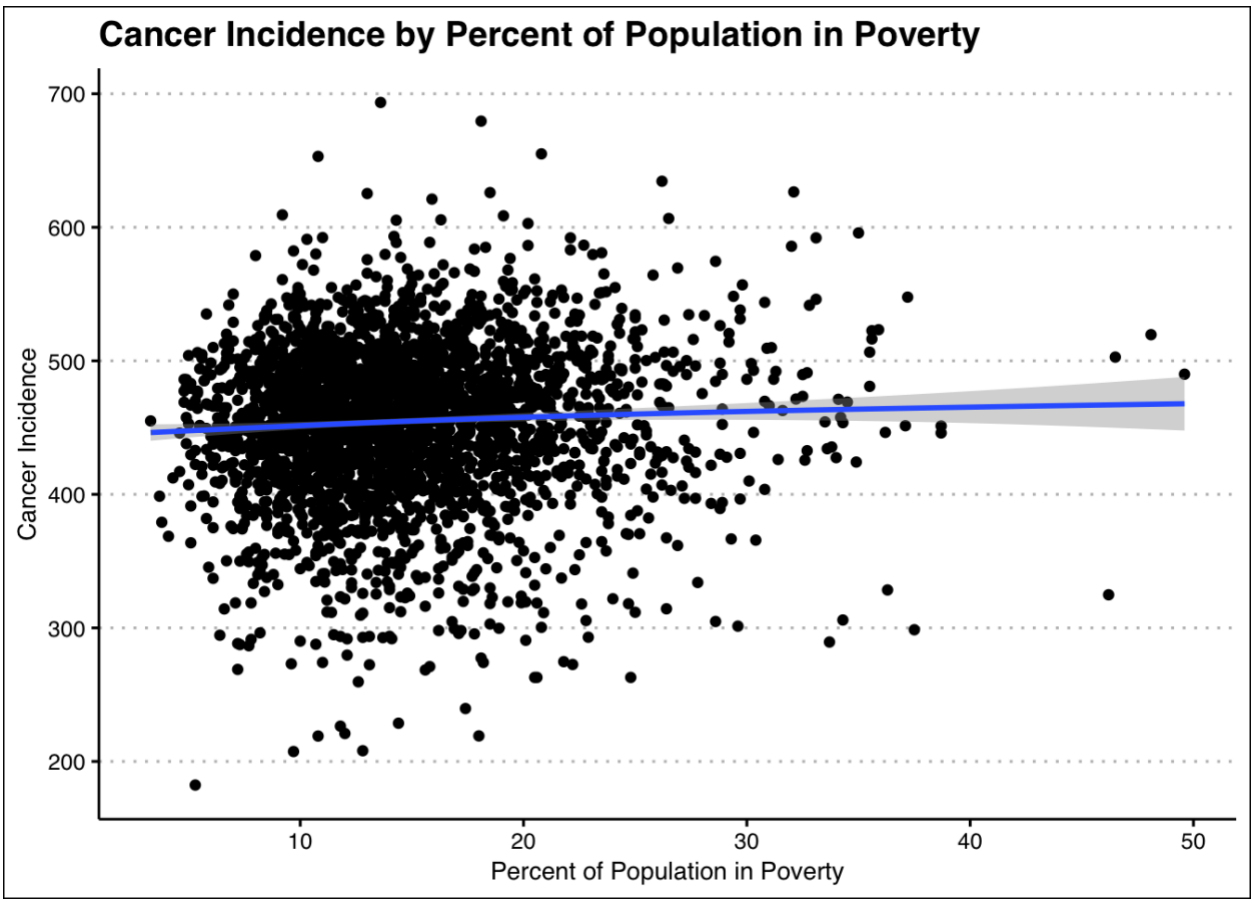

*Figure 9 – Cancer Rate Compared to Unemployment Rate*

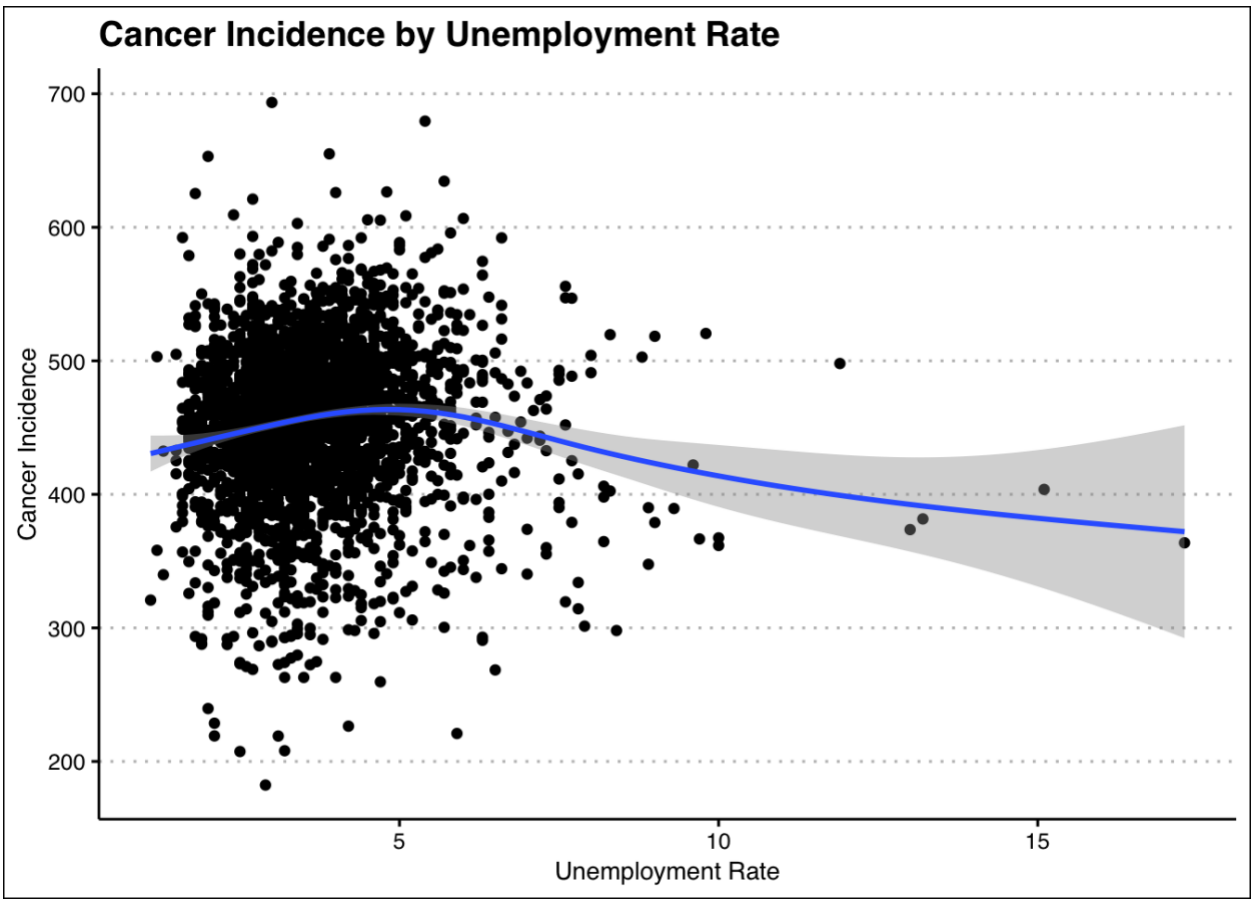

*Figure 10 – Distribution of Cancer Rate with Outlier County Removed*

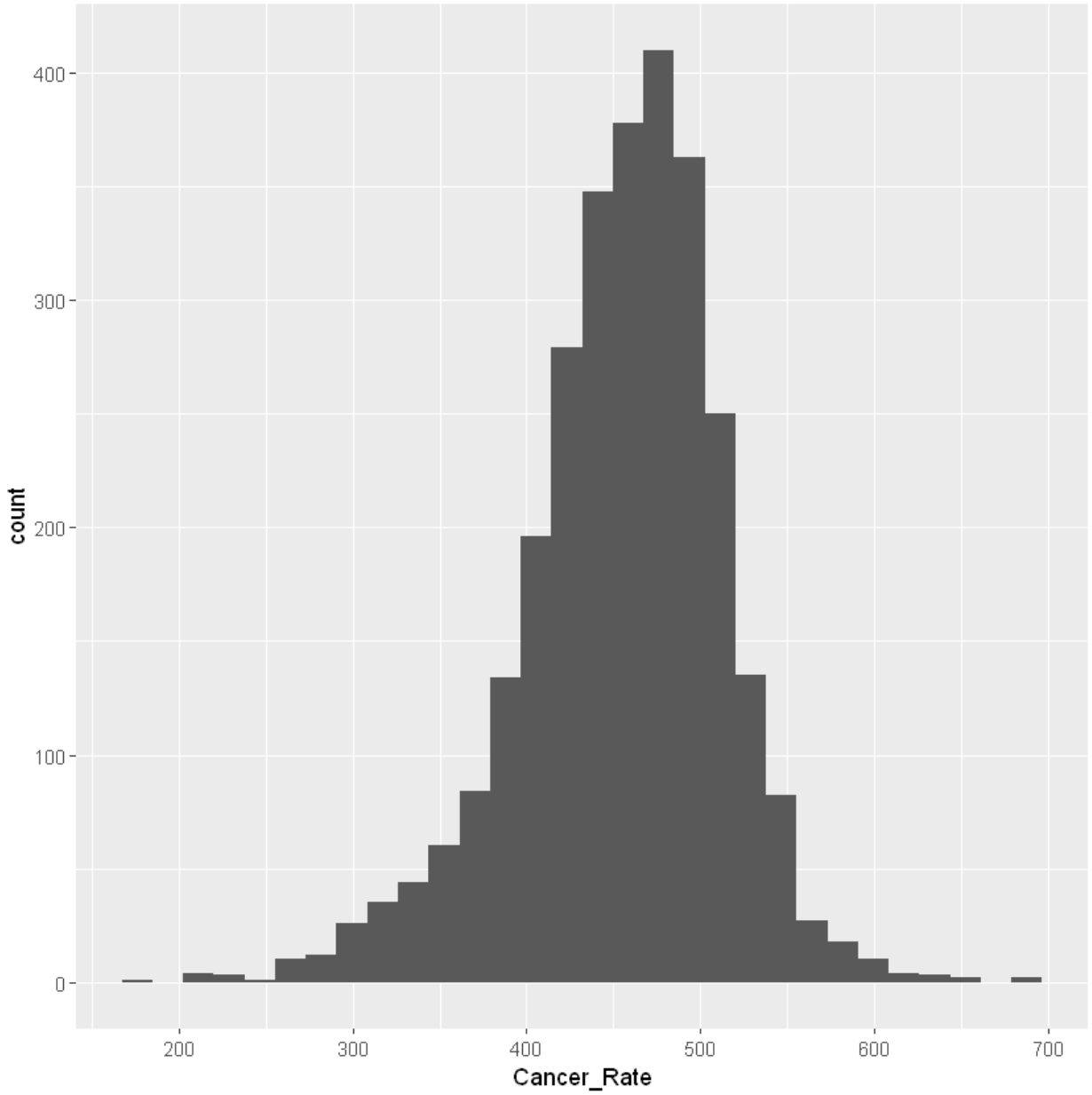

*Figure 11 – Boxplots for Selected Categorical Variables*

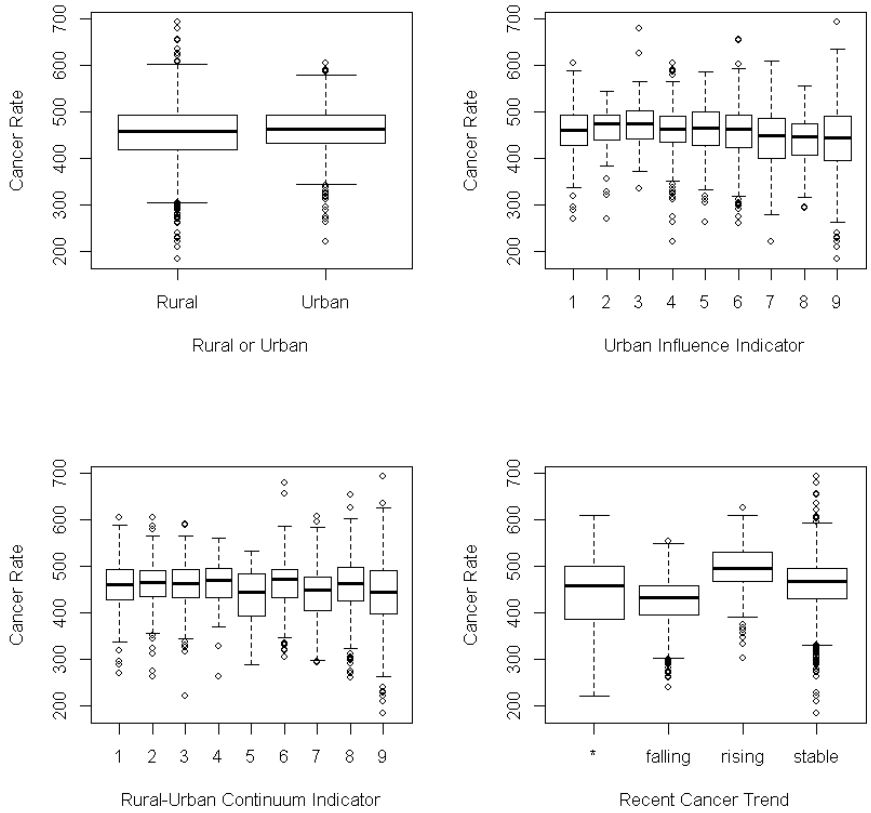

Figure 12 – Cancer Rate Compared to Select Quantitative Predictors

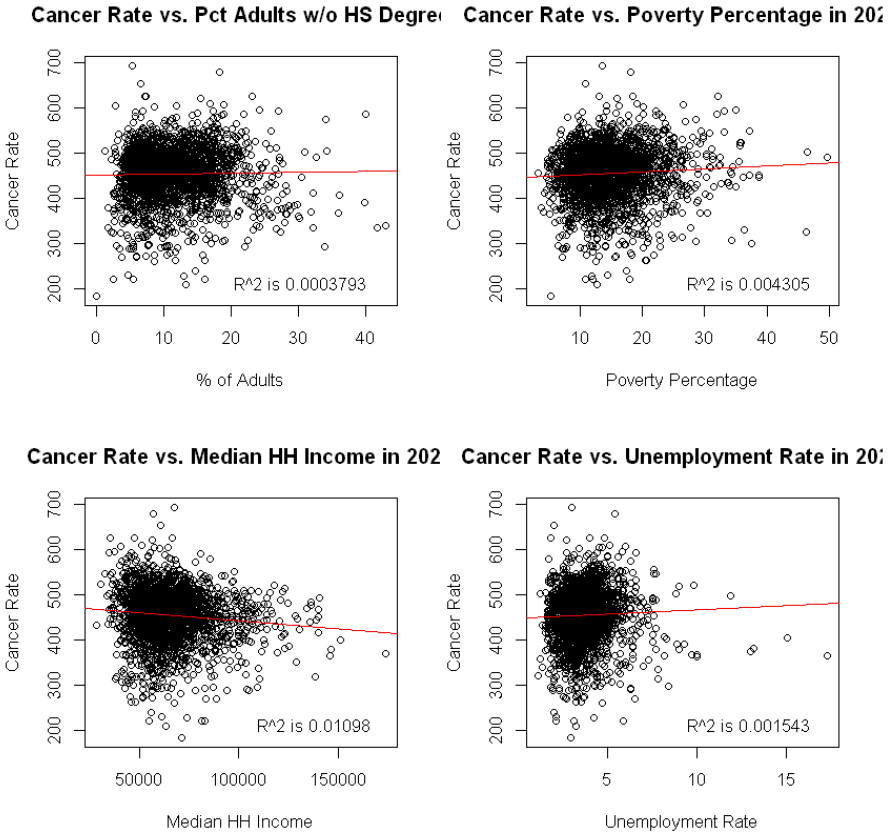


*Figure 13 – Diagnostic Plots for Model Before Variable Selection*

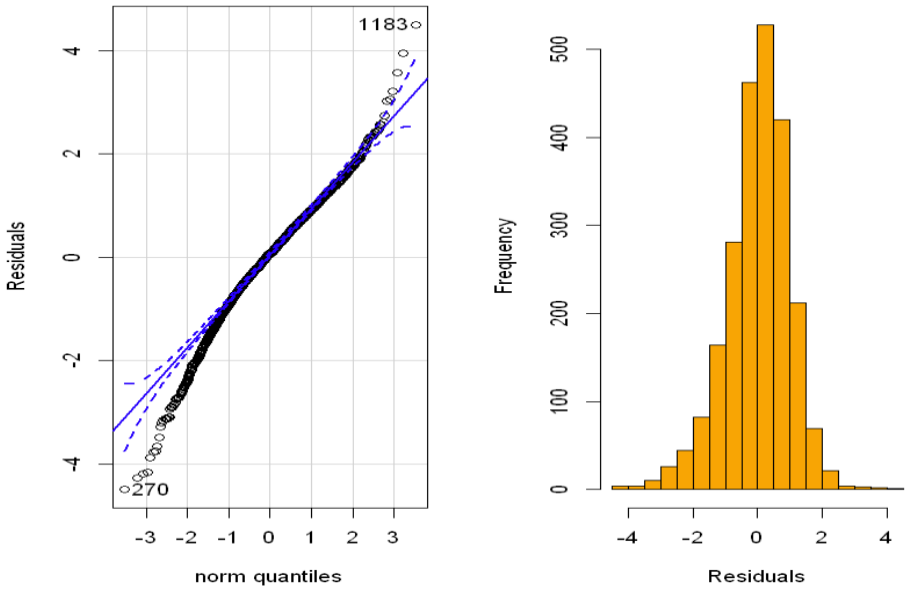

*Figure 14 – Diagnostic Plots for Goodness of Fit Testing*

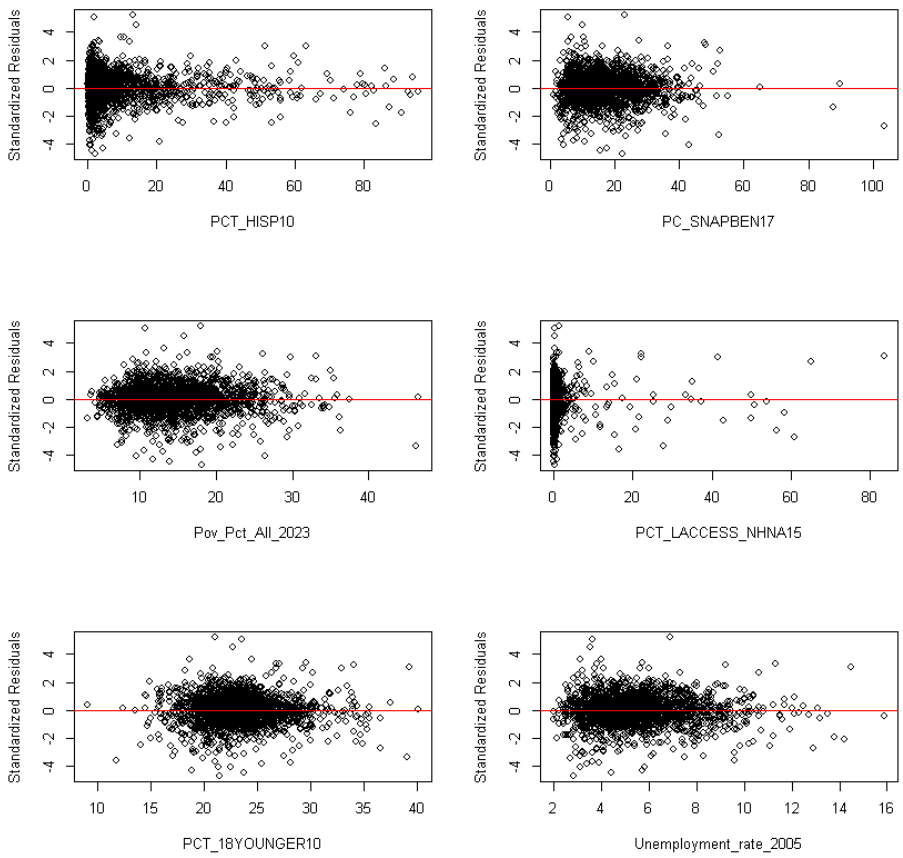


*Figure 15- Diagnostic Plots for Model After Variable Selection*

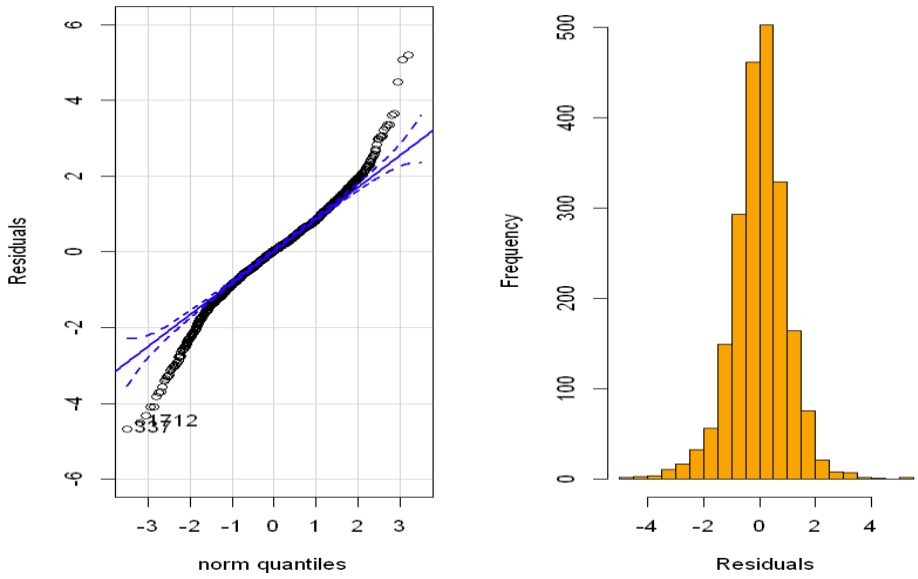

*Figure 16*

This figure shows the coefficient path for the 10-fold CV Elastic net with the calculated alpha of 0.38. The vertical dashed line plots the log of the lambda value that minimizes the MSE.

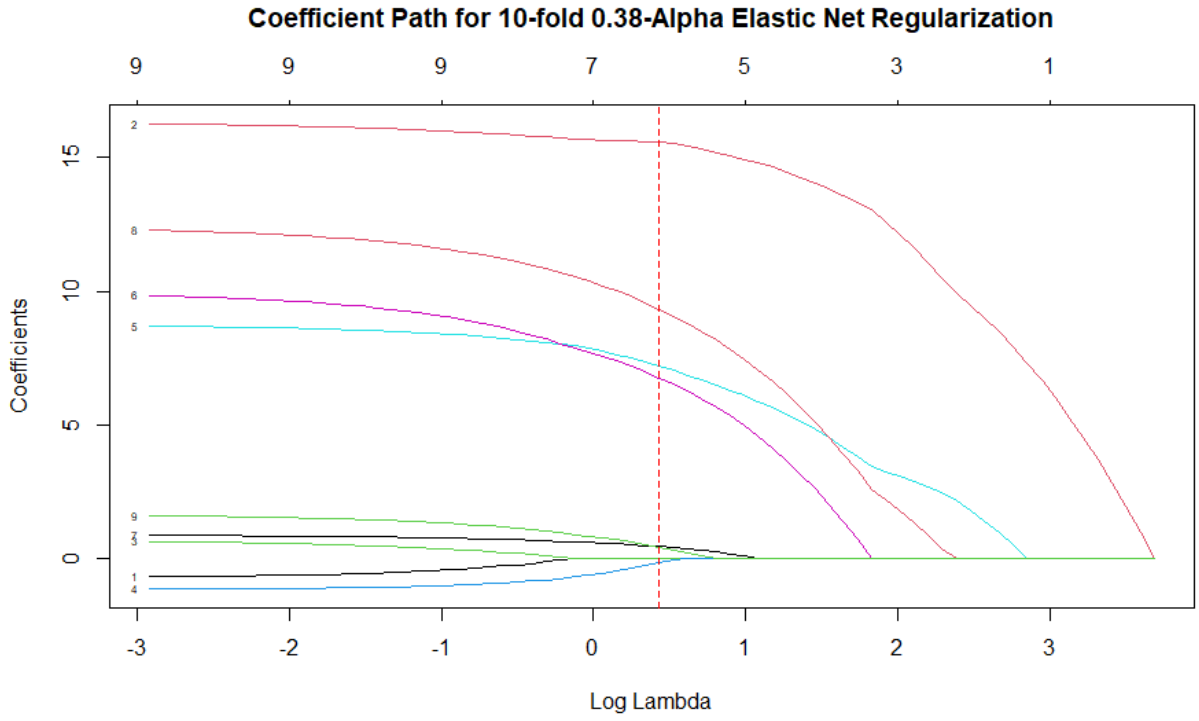

*Figure 17*

This figure denotes the coefficient path for the 10-fold 0.5 alpha elastic net. The vertical dashed line plots the log of the lambda value that minimizes the MSE.

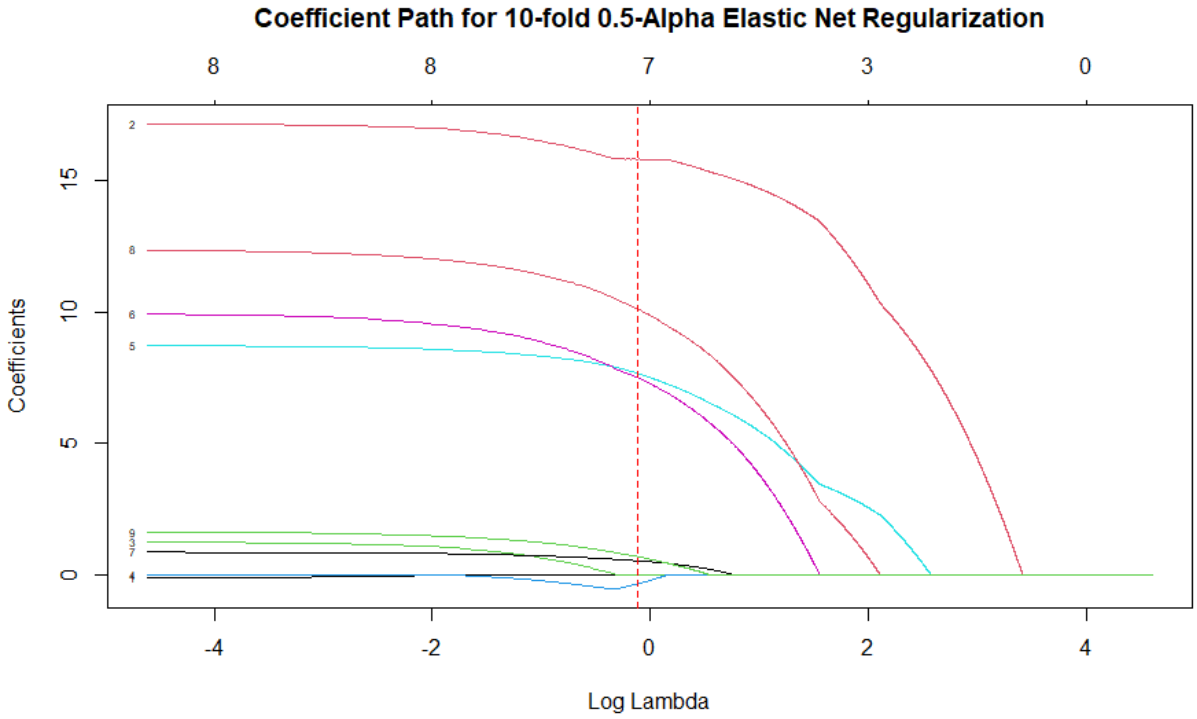

*Figure 18*

This figure visualizes the coefficient path for 0.5 alpha elastic net with a lambda of 0.01 so that the shrinkage performed is minimal. The vertical dashed line plots the log of the lambda value that minimizes the MSE.