Analysis of Effects of Access to Healthcare by Region on Modeling Cancer Mortality Rates

Selina Lu

STATS 401: Applied Statistical Methods II

December 12, 2023

Background

Medicine still has yet to find a cure to one of the leading causes of death today: cancer. There are many known causes of cancer in the biological realm as relating to the body such as carcinogens, lifestyle choices, other related diseases, and even without a known source. Outside of the biological field, there are also social factors that could contribute to cancer mortality rates in the United States. The American Community Survey, conducted by the Census, along with other organizations relating to clinical trials and cancer research have collected data from around the country at the county-level, aggregated into a dataset containing various demographic factors that may have a relationship with cancer mortality rate.

The variables that were chosen to be investigated in relation to cancer mortality rate for this study include cancer mortality rate, median income (in dollars), average household size (number of residents), percent of county residents with various healthcare coverage, and region. Cancer mortality rate is measured as deaths per 100,000 people. The health coverage plans explored include private, employer-provided, and government provided. Regions are separated into categories of Midwest, Northeast, Southeast, Southwest, and West. The goal of this analysis is to explore the different relations between these demographic factors and cancer mortality rate, using the results to create the best model to represent the relationship between them.

Analysis

Initial Model

The initial model displayed the response variable of cancer mortality rate based on quantitative predictor variables of percent of county residents with private health coverage, employer-provided coverage, and public (government-provided) coverage along with one categorical factor variable of region. From the summary of the model, Figure 1, the R^2 value is

0.2861, so approximately 28.61 percent of the variation within the model can be explained by the predictor variables mentioned before. Looking at the residuals vs. fitted values plot in Figure 2, it can be assumed that linearity is met as the points closely follow the zero line in a linear fashion and there is constant variance as well, meeting our assumptions of Ordinary-Least-Squares. Looking at the QQ plot in Figure 3, there are slight deviations from both tail ends of the plot. However, as the sample size is large (537 counties) and each region has 10 times the number of predictors, it can be assumed that the distributions of the predictors are approximately normal (Figure 4). Multicollinearity for the initial model was also checked through variance inflation factors, yielding moderate correlation with VIF values in the 3-4 range.

Data Exploration and Model Alterations

Creating a scatterplot matrix of the data from the initial model in Figure 4, very slight conical spread of points can be seen when examining the different quantitative predictors and comparing them to cancer mortality rates. Figures 5.1-5.3 show different log transformations done on the variable and the response, yielding no significant changes in the histograms or plots. After looking at these results, the percent employer-provided coverage variable was removed from the model to attempt to decrease multicollinearity between the predictor variables which appear to overlap and have moderate to strong relations with each other as seen in the scatterplot matrix (Figure 4) and with a VIF of 4 with our cutoff for high collinearity falling at a value of 5. This yielded a lower R^2 value of 0.2689 as well as lower VIFs, now in the 2 range as seen in Figure 6. Figure 7 displays the residuals and QQ plot for this updated linear model, confirming that the assumptions are still met.

As one predictor variable was removed, additional variables were tested to see if they would help explain the variation found in the model. Figure 8 displays the histograms of the additional tested variables of average household size, median income, birth rate, and median age along with their log transformations. As median income was the only variable with a moderate right skew whereas the others were approximately normal, the log transformation was applied to this variable, yielding a stronger linear plot in the scatterplot. The scatterplot matrix of data in Figure 9 compares the new predictor variables with cancer mortality rate and shows that only median income appears to have somewhat of a linear relationship whereas the others do not have clear linearity. Thus, the median income predictor variable was added to the linear model 3 in Figure 10, yielding a slightly higher R^2 of 0.2726 and slightly higher VIFs around 3. Figure 10 also features a residuals and QQ plot, confirming that this model still follows our assumptions of linearity, constance variance, and normality.

Interaction Attempt

An interaction between the quantitative variable percent of county residents with private coverage and the categorical variable of region was explored, with Figure 11 displaying a scatterplot of the variable data by region. It is evident in the scatterplot that there are multiple slope differences depending on the region, confirmed by Figure 12 where the predictors were modeled with interaction to be plotted. Figure 13 compares the models without interaction and with interaction, showing an increase in R^2 to 0.2794. Figure 14 confirms that this new linear model again follows the assumptions of linearity, constant variance, and normality. Although the increase is not large, it was decided to keep this interaction in context of the dataset as private health coverage policies differ state by state and could affect cancer mortality rates because of this difference in magnitude. One last exploration of variables was done in Figure 15, removing the categorical variable of region altogether. However, this resulted in a large decrease in R^2 to 0.17 and this change was thus not kept.

Final Model

Thus the final model is the fourth model which included cancer mortality rate as the response with predictors median income, percent public coverage, and an interaction term between percent private coverage and region. The residuals vs fitted plot shows the plot following the zero line, assuming linearity, and with constant variance throughout the plot. The QQ plot shows slight deviations from line with some additional outliers, but does not appear significant enough to violate normality. The assumption of independence may possibly be violated given this dataset has many predictor variables that may relate to each other, such as the amount of income one makes may affect the type of health coverage they are able to afford. When checking for multicollinearity, the VIFs were around 3.0, indicating some signs of multicollinearity, but as it has not broken the threshold of 5.0, the predictors are not strongly related to each other enough to break the assumption of independence. The most practically significant predictors is percent private insurance as it appears to have an almost direct 1:1 relationship with the response. For about every percent decrease in private insurance coverage, there is an additional person increase in the cancer mortality rate. Depending on the region, this magnitude of addition or subtraction to the cancer mortality rate may be slightly higher or lower.

Conclusion

Overall, the model's R^2 value was not as high as expected, indicating that although measures were taken to best improve the model, these demographic factors may only play a partial role in the cause of cancer mortality rates whereas other unknown predictors may have stronger or more direct relations. The RMSE for this final model was about 23, which on a scale of per 100,000 people is not strongly significant, but within the context of this data set, means that there is an average error of 23 to be added or subtracted from the cancer mortality rate.

Although the addition of some variables or interactions did not yield large changes in R^2 values, they were still important to include based on the premise of the varying healthcare coverages that exist across the United States. In the future, more interactions between regions and other predictor variables should be researched as the United States is a vast country with many different attitudes and policies towards healthcare and medicine depending on the region.