Is Employment the Cure to Cancer?

Andrew Wolfe and Morgan Huhndorff

STATS 401: Applied Statistical Methods II

April 25, 2022

Background

The dataset cancer contains county-level data from a random sample of 570 counties with 28 variables. Cancer mortality rates can be predicted by exploring the relationship between the response variable ‘cmRate’, the number of cancer mortalities per 100,000 people, and potential predictor variables. The variables record a variety of county data including economic and healthcare status, geographic location, racial diversity, and median ages. Median income by county, labeled ‘medianIncome’ in the data, is probably a strong predictor in addition to the percent of country residents with private health coverage, or ‘pctPrivateHC.’ These variables give insight into the residents’ access to healthcare and if they can afford the costs of treatment. The percent of residents who identify as White, ‘pctWhite,’ could be used to determine the socioeconomic diversity of the county. It is known that racial and ethnic minority groups experience higher rates of mortality, so calculating the percent of residents who do not identify as white is also helpful. Lastly, the categorical variable ‘region’ can be used to identify relationships that may exist between location, cancer mortality rates, and healthcare coverage. It may be a strong indicator of where we can improve as a country for healthcare purposes.

Analysis

Initial Model

The cancer dataset showed a wide range of cancer mortality rates by county, precisely from 94.4 to 293.9 per 100,000 people with a mean of 179 (Figure 1). To explain this variability in cancer mortality, the initial variables explored were median income, percentage of residents with private health coverage, and percentage of white residents. An initial linear regression model was created using these predictor variables without the use of any transformations or interactions. The summary of this initial model showed that median income had a strong effect on the model, given its statistically significant p-value (Figure1). The other predictor variables, however, show to be not statistically significant, so this initial model may benefit from the implementation of adjustments such as a quadratic term and transformations. Cancer mortality rates appear to have a strong linear relationship with private healthcare percentages but a non-linear, possibly quadratic, relationship with median income (Figure 2). With the addition of a quadratic term on median income in the linear regression, the resulting p-value was low enough to prove that a quadratic fit is more appropriate (Figure 5). The histogram of variable 'pctWhite' showed a heavy left-skewed distribution, requiring some transformations before it can be useful in a linear regression model (Figure 3). First, the variable can be flipped to create a new variable called 'pctNonWhite,' the percentage of residents who do not identify as white. This was completed by subtracting the 'pctWhite' variable from the value 100 because the data is recorded as percentages. Thus, the histogram becomes a distribution with a right skew, and a log transformation is appropriate to create an approximately normal sampling distribution (Figure 4). Lastly, the variable 'pctEmployed' which records the percentage of county residents that are employed was added to the model as a predictor because it has a very strong, negative relationship with cancer mortality rates (Figure 6).

Categorical Variables

As for a categorical variable, the counties were divided into regions that showed little differences in cancer mortalities in a side-by-side boxplot (Figure 7). This was further verified using the emmeans() and pairs() function in R that test for a difference in the estimated means of each region. When a model was run with an interaction term between private health coverage and region, there was only one statistically significant p-value of 0.000379 which tests for a difference in the slopes between Southwest and Midwest (Figure 8). A scatterplot of cancer mortalities versus private health coverage colored by groups further revealed that the slope for Southwest was positive unlike the others. Still, the slopes of the remaining regions did not differ by much. When comparing the two models with and without the interaction, the interaction model did have a higher adjusted r-squared value of 0.2635 compared to the value 0.162 of the non-interaction model (Figure 11). The interaction was ultimately dropped from the final model because it over-complicates the model and does not seem to be practically significant.

$$E_{cmRate}|X = \beta_0 + \beta_1 X_{medianIncome} + \beta_2 X^2_{medianIncome} + \beta_3 log(X_{pctNonWhite}) + \beta_4 X_{pctPrivateHC} + \beta_5 X_{pctEmployed}$$

Diagnostic Plots and Assumptions

A QQ and Residual Plot was used to check for assumptions regarding our linear model (Figure 11). The QQ plot, used to compare the data to a theoretical distribution, falls on the straight line with no deviations indicating that the assumption of normality is not violated. A Residual Plot was also created to confirm linearity and constant variance and shows a random scatter of residuals above and below zero in a constant width. With these assumptions confirmed and confirmation that the data comes from a random sample of the population, a linear regression model can be executed properly.

Final Model

For our final linear model, we decided to use the quadratic term on ‘medianIncome’, the log of our flipped variable ‘pctNonWhite’, ‘pctPrivateHC’, and ‘pctEmployed’ to predict the number of cancer mortalities per 100,000 people. The summary of this model results in an coefficient of determination, or R, value of 0.1694, indicating that about 17% of the variation in the number of cancer mortalities per 100,000 people can be explained by the linear relationship with the four predictor variables (Figure 12). In addition, the square root of the variance of the residuals (RMSE) is 24.93, representing the average error of cancer mortalities. Although there is strong evidence for multicollinearity on median income as the vif value is above 10, this can be ignored because of the quadratic term. The model follows the rule of thumb for overfitting in which $n \geq 10p$ (570 > 4*10). For interpretation, β_0 (the intercept) represents the estimated average value of cancer mortality rates when all explanatory variables are set to equal zero. β_1 is the slope for the explanatory variable median income while β_2 tells both the direction and curvature of the quadratic fit. The log transformation has a unique interpretation of β_3 in which a k% increase in 'pctNonWhite' results in an estimated average increase of $\beta_3 * log(1 + \frac{k}{100})$ in Y. β_4 is the slope for 'pctPrivateHC', and β_5 is the slope for 'pctEmployed'. The percentage of employed residents is both practically and statistically significant. For a 1% increase in employed residents, cancer mortalities decrease by 0.5601 per 100,000 people. Its low p-value of 0.00940 indicates strong statistical significance. As mentioned in the initial model, median income was also statistically significant. It would not be considered practically significant however because a $1 increase would only decrease mortalities by 0.001813.

Conclusion

To conclude, the final linear model revealed that the combination of 'medianIncome', 'pctNonWhite', 'pctPrivateHC', and 'pctEmployed' are useful predictors of cancer mortality per 100,000 people. A low-enough p-value was obtained to review that our linear model was significant, benefiting from a quadratic term of the 'medianIncome' variable. Employment percentage proved to be a strong predictor as it was both practically and statistically significant. Although the adjusted r-squared value is not ideal, future exploration could determine how these variables may interact with one another due to the natural correlation of income with employment and health coverage plans. Lastly, assumptions of normality, linearity, and constant variance were checked through residual plots and qq plots, indicating that OLS is reasonable for estimating a final linear model.