Final Report: Analyzing Socioeconomic Factors on USA Cancer Mortality

Jasen Zhang, Alexander Zhu, Christian Pascual

Introduction

Cancer ranks among the leading causes of death worldwide. According to a report published by the American Cancer Society, an estimated 1.8 million new cancer cases are estimated to be diagnosed in 2020 in the United States, and more than 600,000 are expected to die directly as a result of cancer [1]. Cancer also carries a significant economic burden, costing the United States an estimated $80.2 billion according to a 2015 report [2]. From both a humane and economic perspective, investigation into effective, affordable cancer cures and treatments represents a important front of research.

Cancer is inherently a disease of the genes, but a wide berth of research has demonstrated that a diverse set of environmental and socioeconomic factors contribute to increased risk of cancer incidence and mortality. For example, Adler et. al showed that higher socioeconomic status was associated with decreased cancer mortality [3]. Rawl et. al demonstrates a similar result in a statewide survey in Indiana that income and education were inversely related to cancer mortality. Rawl also discusses how race affected cancer mortality in their sample, finding that African-American participants worried less about cancer and were less likely to seek treatment [4]. Rohfling et. al found that uninsured patients or those under Medicaid were more likely to have more advanced tumors and poorer survival compared to peerse with private insurance [5]. Higher socioeconomic status gives people better access to healthcare resources that are potentially life-saving or will help increase survival, and the opposite effect has been seen the lower scale.

Similar research has documented race-based health disparities in cancer mortality. In an observational study in Philadelphia, Zeigler-Johnson found that black men were at the highest risk of prostate cancer relative to similar white counterparts [6]. Looking at data spanning from 1950 to 2014, a study by Singh showed that individuals from lower educational backgrounds experienced higher mortality of various types of cancer. Furthermore, African-Americans saw higher cancer mortality compared to their Asian and White counterparts in this group [7]. These are just a few examples of studies that have documented how race influences cancer mortality, highlighting that it is critical to consider in any cancer-related intervention.

One difficulty in researching cancer is that it is an incredibly diverse collection of diseases, as opposed to a monolithic set of symptoms. Further complicating this is that different cancers can occur at different rates across different regions of the United States. Mokdad et. al found that there were distinct clusters of counties in different regions with especially high cancer mortality. For example, breast cancers are highly prevalent in the southern belt, whereas liver cancer is the prevailing diagnosis along the Texas-Mexico border [8]. The heterogeneity of different cancers in the United States offers an interesting research avenue. Just as the aforementioned studies examined how socioeconomic factors and race affect cancer mortality on an individual, it could

be useful to understand how these factors relate to cancer mortality on a higher, geographic level. Understanding how these factors contribute to cancer mortality on a geographic level offers an opportunity for researchers to understand the factors that contribute to higher mortality in different states and possibly a better way to allocate health resources to areas that are harder hit.

Data

For our analysis, we will use a dataset aggregated from multiple sources including the American Community Survey, the National Census, clinicaltrials.gov, and cancer.gov. The data spans from 2010 to 2016 and includes information on 3047 counties and county-equivalents in the United States. Some counties are missing from the dataset since there are a total of 3,143 counties and county-equivalents in the United States.

Our data contains information on various demographic, socioeconomic, household, and cancer-related factors for each county, represented typically as percentages. Each row contains the percentage of each race (Asian, Black, White and Other) that live in the county. The data also contains information on educational acheivement as well, measured as the proportion of the county population who have achieved high school and college degrees. The proportion of people who have public and private health insurance is another notable variable in the data. In terms of important cancer-related factors, we have the incidence rate of *all* cancer diagnoses in the county, measured in terms of *mean per capita (100,000 people)* and the average number of cancer cases reported annually from 2010 to 2016.

Our target response is cancer mortality, also measured as mean per capita. In this dataset, cancer mortality ranges from 59.7 to 362.8, with a median value of `cancer$target_death_rate %>% median`. The outcome is also reasonably bell-shaped, so we don't think any transformation will be necessary for the analyses.

Initial Analyses & Objectives

Given our literature and what is present in our dataset, we aim to explore how different socioeconomic and demographic factors are associated with cancer mortality on a county level.

After some initial exploratory analyses. we found that

- exploratory analyses:
  - relationship between outcome and:
    - race
    - education
    - cancer incidence
    - income
    - insurance status
  - correlation between the different predictors

Based on the results of our exploratory analyses, we propose two analytic questions:

1. Can we identify county-level factors that contribute to a significant difference in cancer mortality? If so, can we identify any significant interaction between these factors that also contribute to increased cancer mortality?
2. Can we use any significant findings from (1) to improve the predictive ability of a potential predictive model?

Methodology

Explanatory Model

To select the best possible candidate model,

- model selection
  - ANOVA
  - how did we choose which subset of the columns to use?
  - found that interactions are significant and should be considered
- show the selected model (latex)
- choice of confounders in explanatory model
- Using a Wald Test
- checking regression diagnostics
- using bootstrap to double check our inferences

Prediction Model

- choosing which model to use
  - PCA, regularization
- how to split the data

Results

Explanatory Model

- Results based on the wald Test
- regression diagnostics results
  - qqplot, residual plot
- inferences using the bootstrap

Prediction Model

- test MSE (final model, and final model including interactions)

PREXISTING MATERIAL BELOW

Inference Model

One of our goals was to assess whether there was any significant interaction between race and education that contributed to cancer mortality in a county. The dataset contains different percentages of educational attainment in each county and the percentage of four race categories (Asian, African American, Other and White). We elected for a model that looked at high school achievement and bachelor degrees among adults 25 years and older because we felt that this population would be more reflective of the education levels of the working population in each county. We experimented with a few different interaction models to see what combination of education and race predictors would make for both a sensible and useful model. Our final inference model includes 2 educational variables (high school achievement rate and bachelor degree achievement rate), the aforementioned 4 race percentages, 8 interaction variables based on each race and education pair, and also controls for cancer incidence, median income, county population size, and median age. We use the Wald

test to check if any of the interaction coefficients are non-zero. Our dataset is reasonably large, so we feel the test is appropriate for our needed hypothesis with $\alpha = 0.05$. As another way to assess the significance of the interaction coefficients, we used 5000 bootstrap samples and examined the 95% bootstrap percentile interval to see if its results agree with the test.

Interaction Between Education & Race

Our interaction model found that 3 of the interaction coefficients were significant. We found that the interaction between both educational variables and African-Americans was significant, as well as the interaction between high school achievement rate and white Americans. The coefficients associated with these interactions are all negative (albeit small), indicating that they help *reduce* cancer mortality. Figure XXX shows the estimated interaction coefficients, along with their 95% confidence intervals.

These results were encouraging and were subsequently supported by the results of the Wald test, which yielded a p-value less than 0.05. This test only indicates that at least one of the interaction variables is non-zero, but we look to the bootstrap 95% percentile intervals to get an idea of which were. Figure YYY shows the results of 5000 bootstrap interaction models.

The bootstrap 95% confidence intervals indicate that only the high school interactions with African- and white Americans were significant. The direction of these bootstrap estimates matches that of the overall model, which further supports our findings. We feel confident that our model supports our research hypothesis.

Prediction Model

- results of cross-validation
  - optimal hyper parameters
- final test prediction score

Conclusion

Our analysis found at least one significant interaction between race and education that served to help reduce cancer mortality per capita on a county level. Although this effect was small, their significance was supported simultaneously by the model itself, the Wald test, and the 95% bootstrap percentile intervals. These results support the hypothesis that higher education has a beneficial effect on cancer mortatlity, at least on the county scale. Furthermore, our findings are in line with a wealth of literature that suggest that education is beneficial to improving health outcomes.

These findings are limited by the fact that the observational unit of the data was a county. The model suggests a beneficial benefit, but does not elucidate any reason as to why or how education interacts with race to reduce cancer mortality among African- and white Americans. The educational and race data come from 2013, so the model may be limited in its generalizability to future years, especially in the context of changing demographics and new educational environments created by COVID-19. In light of these weaknesses, we feel that our model offers an interesting perspective on how different demographic factors interact to affect a health-related outcome. Being so prevalent, it's important to understand these factors and interactions so that we can design proper interventions.

- results

- significance
- weaknessses

References

3. ADLER, N.E. and OSTROVE, J.M. (1999), Socioeconomic Status and Health: What We Know and What We Don't. Annals of the New York Academy of Sciences, 896: 3-15. https://doi.org/10.1111/j.1749-6632.1999.tb08101.x
4. Rawl SM, Dickinson S, Lee JL, Roberts JL, Teal E, Baker LB, Kianersi S, Haggstrom DA. Racial and Socioeconomic Disparities in Cancer-Related Knowledge, Beliefs, and Behaviors in Indiana. Cancer Epidemiol Biomarkers Prev. 2019 Mar;28(3):462-470. doi: 10.1158/1055-9965.EPI-18-0795. Epub 2018 Nov 28. PMID: 30487135.
5. Rohlfing ML, Mays AC, Isom S, Waltonen JD. Insurance status as a predictor of mortality in patients undergoing head and neck cancer surgery. Laryngoscope. 2017 Dec;127(12):2784-2789. doi: 10.1002/lary.26713. Epub 2017 Jun 22. PMID: 28639701; PMCID: PMC5688011.
6. Zeigler-Johnson C, Keith S, McIntire R, Robinson T, Leader A, Glanz K. Racial and Ethnic Trends in Prostate Cancer Incidence and Mortality in Philadelphia, PA: an Observational Study. J Racial Ethn Health Disparities. 2019 Apr;6(2):371-379. doi: 10.1007/s40615-018-00534-z. Epub 2018 Dec 5. PMID: 30520002.
7. Singh, Gopal & Jemal, Ahmedin. (2017). Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States, 1950–2014: Over Six Decades of Changing Patterns and Widening Inequalities. Journal of Environmental and Public Health. 2017. 1-19. 10.1155/2017/2819372.
8. Mokdad AH, Dwyer-Lindgren L, Fitzmaurice C, et al. Trends and Patterns of Disparities in Cancer Mortality Among US Counties, 1980-2014. JAMA. 2017;317(4):388–406. doi: 10.1001/jama.2016.20324