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, which we note later. The data spans from 2010 to 2016 and includes information on 3047 counties in the United States. There are a total of 3,143 counties and county-equivalents in the United States, so 3.1% of them are represented in the data. Our group did not aggregate the data ourselves, it is publically available and can be found here.

Our data contains information on various demographic, socioeconomic, household, and cancer-related factors for each county, represented typically as percentages. These data are gathered from the 2013 Census. 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.

The cancer data has been aggregated from the American Community Survey, cancer.gov and clinicaltrials.gov, spanning from 2010 to 2016. 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.

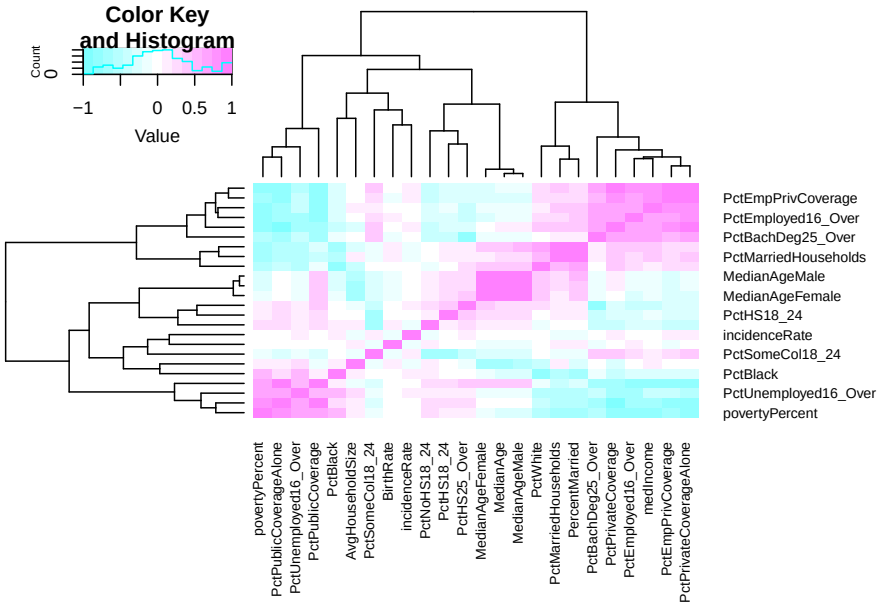
Initial Analyses & Objectives

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.

Correlation Between Predictors

Many of the predictors are highly correlated, so we created a heat map to keep track of these intercorrelations. Figure 1 below shows this heat map.

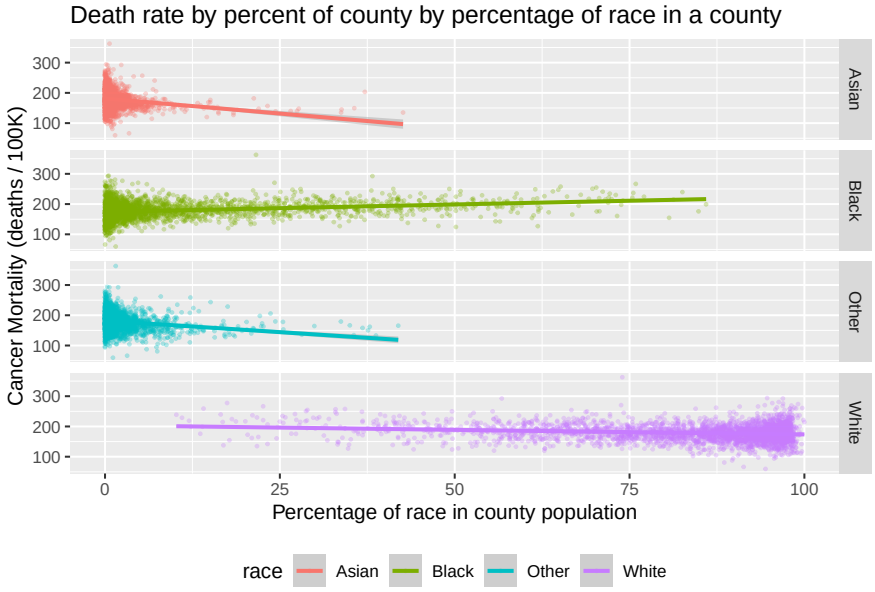
```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(all_vars)` instead of `all_vars` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```



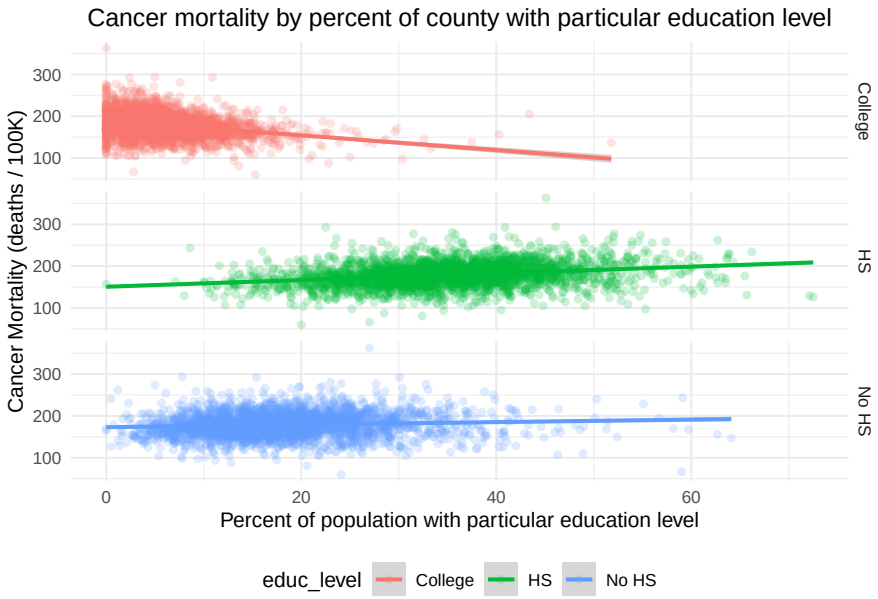
The `povertyPercent` variable correlates highly with other predictors that deal with being under public insurance of being unemployed. Conversely, poverty is highly negatively correlated with having private insurance. The median ages of men and women are highly correlated as well, so we will probably opt for just one of these. Since the educational variable represent highest achievement, many of them are correlated as well.

In making our future models, we now know that we'll have to deal with this correlation and perform some model selection on our variables. We performed an initial PCA to figure out what variables explained most of the variance in average cancer mortality, and this will help guide our final model selection.

Race & Average Cancer Mortality

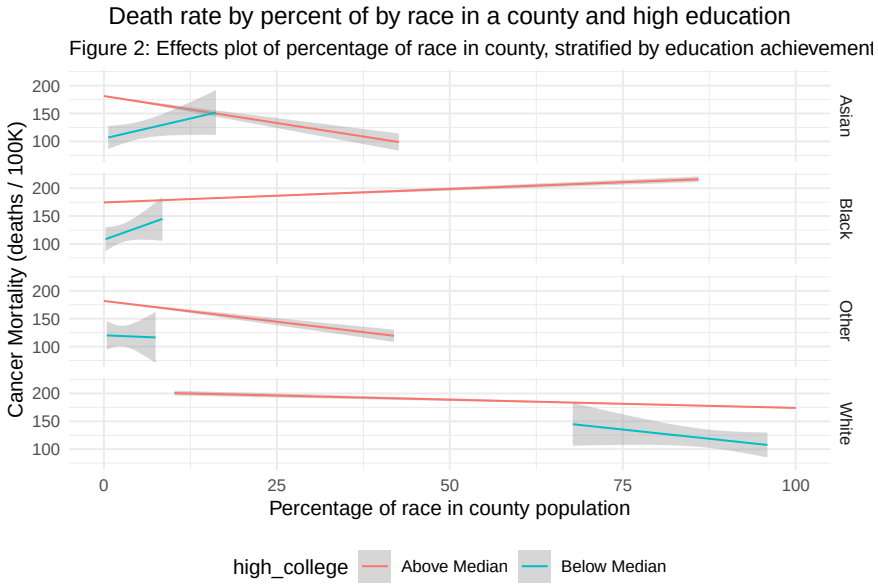


Education & Average Cancer Mortality



Interaction Between Race & Education

One interesting trend that we found in the data was that there seemed to be an interaction between race and education in relation to cancer mortality. According to an article by the US Department of Education, the median percentage of adults 25 and over completing a bachelor's degree was 34%. We divided the counties by if they fell below or were equal to higher to the median value and investigated how cancer mortality changed with percentage of race (Figure 2).



The change is most drastic in Asians, but the trend holds over all races present in the data. For African-Americans, we also see an attenuation of the average cancer mortality desite it not converting to a negative correlation.

- 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 create an effective predictive model from the data? If we can find any significant interactions from our explanatory model, might they be helpful in increasing predictive ability?

Our two questions is motivated by the nature of the data. The data is contains most of the counties in the United States, so any estimates we find apply directly to these counties during this time period. Since the data is a "snapshot" of the population from 2011 - 2016, we must acknowledge that the estimates we get in any explanatory model will reflect population characteristics. We also acknowledge that some variation is introduced because the data is come different points in time, we assume that these measurements do not change drastically on the scale of a few years. With the nature of the data in mind, we want to try to use the explanatory model to see if any features or combination of features might be useful in a prediction model, based on the estimated coefficients.

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 & Discussion

Explanatory 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 1000 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.

- 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)

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.

References

1.

2.

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

9. https://nces.ed.gov/pubs93/93442.pdf

Appendix

```r
library(tidyverse)
library(knitr)
library(gplots)
set.seed(1)

knitr::opts_chunk$set(
  fig.align = "center",
  out.width = "70%"
)

# Load the data and do some processing
cancer = read_csv("cancer_registry.csv") %>%
  # Split the geography variable
  separate(Geography, into = c("county", "state"), sep = ", ") %>%
  # Split up binnedInc into a lower and upper decile
  mutate(
    binnedInc = str_remove_all(binnedInc, "[(\\]]"),
    # also try to group states by region
    region = case_when(
      state %in% c("California", "Oregon", "Washington", "Nevada", "Idaho",
                   "Montana", "Wyoming", "Colorado", "Utah", "Alaska", "Hawaii") ~ "West",
      state %in% c("Arizona", "New Mexico", "Texas", "Oklahoma") ~ "Southwest",
      state %in% c("North Dakota", "South Dakota", "Nebraska", "Kansas",
                   "Minnesota", "Iowa", "Missouri", "Wisconsin", "Illinois",
                   "Indiana", "Ohio", "Michigan") ~ "Midwest",
      state %in% c("Arkansas", "Louisiana", "Mississippi", "Alabama", "Georgia",
                   "Florida", "South Carolina", "North Carolina", "Tennessee",
                   "Kentucky", "Virginia", "West Virginia", "District of Columbia",
                   "Delaware") ~ "Southeast",
      state %in% c("Maryland", "Pennsylvania", "New Jersey", "New York", "Rhode Island",
                   "Connecticut", "Massachusetts", "New Hampshire", "Vermont", "Maine") ~ "Nort
      TRUE ~ "Southwest" # Weird formatting means a single NM is NA in state
    )
  ) %>%
  separate(binnedInc, into = c("inc_dec_low", "inc_dec_high"), sep = ",") %>%
  janitor::clean_names() %>%  # Convert all column names to lowercase'
  mutate(
    high_college = pct_bach_deg25_over > median(pct_bach_deg25_over), # median(pct_bach_deg18_
    high_hs = pct_hs18_24 > median(pct_hs18_24)
  )
df <- read.csv('cancer_registry.csv') %>%
  mutate(PctSomeCol18_24 = 100 - PctNoHS18_24 - PctHS18_24 - PctBachDeg18_24) %>%
  filter(incidenceRate < 1000) %>%
  filter(avgAnnCount < 20000) %>%
  filter(MedianAge < 200) %>%
```

```
  filter(AvgHouseholdSize > 1)

df <- na.omit(df)
vars <- colnames(df)
misc_vars <- c('binnedInc', 'Geography', 'TARGET_deathRate')
vars_1 <- setdiff(vars, misc_vars)
log_vars <- c('avgAnnCount', 'avgDeathsPerYear', 'popEst2015', 'studyPerCap', 'PctBachDeg18_24
log_names <- c()

all_vars <- setdiff(vars_1, log_vars)
all_vars <- append(all_vars, log_names)
all_vars <- sort(all_vars)

df_temp <- df %>% select(all_vars)
df_temp <- na.omit(df_temp)

heatmap2 <- cor(df_temp)
heatmap.2(heatmap2, trace = 'none', margins = c(10,10), col = 'cm.colors', cexRow=0.7, cexCol =
cancer %>%
  pivot_longer(
    cols = c("pct_white",
             "pct_black",
             "pct_asian",
             "pct_other_race"),
    values_to = "pct",
    names_to = "race"
  ) %>%
  mutate(
    race = case_when(
      race == "pct_white" ~ "White",
      race == "pct_black" ~ "Black",
      race == "pct_asian" ~ "Asian",
      race == "pct_other_race" ~ "Other"
    )
  ) %>%
  ggplot(aes(x = pct, y = target_death_rate, color = race)) +
  geom_point(size = 0.5, alpha = 0.3) +
  geom_smooth(method = "lm") +
  facet_grid(race ~ .) +
  theme(legend.position = "bottom") +
  labs(
    title = "Death rate by percent of county by percentage of race in a county",
    x = "Percentage of race in county population",
    y = "Cancer Mortality (deaths / 100K)")
cancer %>%
  pivot_longer(
    cols = c("pct_no_hs18_24", "pct_hs18_24", "pct_bach_deg18_24"),
```

```
    values_to = "pct",
    names_to = "educ_level"
  ) %>%
  mutate(
    educ_level = case_when(
      educ_level == "pct_no_hs18_24" ~ "No HS",
      educ_level == "pct_hs18_24" ~ "HS",
      educ_level == "pct_bach_deg18_24" ~ "College"
    )
  ) %>%
  ggplot(aes(x = pct, y = target_death_rate, color = educ_level)) +
  facet_grid(educ_level ~ .) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm") +
  labs(
    title = "Cancer mortality by percent of county with particular education level",
    x = "Percent of population with particular education level",
    y = "Cancer Mortality (deaths / 100K)")
cancer %>%
  mutate(
    high_college = pct_bach_deg25_over > 34, # median(pct_bach_deg18_24),
  ) %>%
  pivot_longer(
    cols = c("pct_white",
             "pct_black",
             "pct_asian",
             "pct_other_race"),
    values_to = "pct",
    names_to = "race"
  ) %>%
  mutate(
    race = case_when(
      race == "pct_white" ~ "White",
      race == "pct_black" ~ "Black",
      race == "pct_asian" ~ "Asian",
      race == "pct_other_race" ~ "Other"
    )
  ) %>%
  ggplot(aes(x = pct, y = target_death_rate, color = high_college)) +
  geom_smooth(method = "lm", size = 0.5) +
  facet_grid(race ~ .) +
  theme_minimal() +
  theme(
```

```
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5)
  ) +
  labs(
    title = "Death rate by percent of by race in a county and high education",
    subtitle = "Figure 2: Effects plot of percentage of race in county, stratified by educatio
    x = "Percentage of race in county population",
    y = "Cancer Mortality (deaths / 100K)") +
  scale_color_discrete(labels = c("Above Median", "Below Median"))
# Trying out different set of education vars
model2 = lm(target_death_rate ~ pct_hs25_over + pct_bach_deg25_over +
              pct_white + pct_black + pct_asian + pct_other_race +
              # Interactions
              pct_white*pct_hs25_over + pct_black*pct_hs25_over + pct_asian*pct_hs25_over + pct
              pct_white*pct_bach_deg25_over + pct_black*pct_bach_deg25_over + pct_asian*pct_bac
              # Confounders
              incidence_rate + med_income + pop_est2015 + poverty_percent + median_age,
            data = cancer)

# Create nice table for coefficients
broom::tidy(model2) %>%
  mutate(
    left_bound = estimate - qnorm(0.975) * std.error / sqrt(nrow(cancer)),
    right_bound = estimate + qnorm(0.975) * std.error / sqrt(nrow(cancer)),
    conf = paste0(round(left_bound, 3), ", ", round(right_bound, 3))
  ) %>%
  filter(
    term %in% (model2$coefficients %>% names %>% .[13:20])
  ) %>%
  mutate(
    estimate = round(estimate, 3),
    term = factor(term,
                  levels = c(
                    "pct_hs25_over:pct_asian", "pct_bach_deg25_over:pct_asian",
                    "pct_hs25_over:pct_black", "pct_bach_deg25_over:pct_black",
                    "pct_hs25_over:pct_other_race", "pct_bach_deg25_over:pct_other_race",
                    "pct_hs25_over:pct_white", "pct_bach_deg25_over:pct_white"
                  ),
                  labels = c(
                    "HS x Asian", "College x Asian",
                    "HS x Black", "College x Black",
                    "HS x Other", "College x Other",
                    "HS x White", "College x White"
                  ))
  ) %>%
  select(
    Term = term,
```

```
    Estimate = estimate,
    `95% CI` = conf
  ) %>%
  kable(
    caption = "Table XXX: Estimated interaction coefficients (95% CI)",
    align = "c")
create_int_model = function(data) {

  lm(target_death_rate ~ pct_hs25_over + pct_bach_deg25_over +
               pct_white + pct_black + pct_asian + pct_other_race +
               # Interactions
               pct_white*pct_hs25_over + pct_black*pct_hs25_over + pct_asian*pct_hs25_over + pct_
               pct_white*pct_bach_deg25_over + pct_black*pct_bach_deg25_over + pct_asian*pct_bach
               # Confounders
               incidence_rate + med_income + pop_est2015 + poverty_percent + median_age,
             data = data)
}

bs_n = 1000

# Terms to keep
# terms = model$coefficients %>% names %>% .[15:30]
terms = model2$coefficients %>% names %>% .[13:20]

# Create the bootstrap datasets and models
bs = tibble( idx = 1:bs_n ) %>%
  mutate(
    bs_data = map(idx, function(i) {
      sample_n(cancer, size = nrow(cancer), replace = TRUE)
    }),
    bs_model = map(bs_data, function(bsd) {
      create_int_model(bsd)
    }),
    bs_results = map(bs_model, broom::tidy)
  ) %>%
  select(idx, bs_results) %>%
  unnest(bs_results) %>%
  group_by(term) %>%
  summarize(
    n = n(),
    bs_mean = mean(estimate),
    bs_var = var(estimate),
    left_bound = quantile(estimate, 0.025),
    right_bound = quantile(estimate, 0.975),
  ) %>%
  filter(term %in% terms) %>%
  # Convert terms to factors for easier reordering
```

```
  mutate(
    term = factor(term,
                  levels = c(
                    "pct_hs25_over:pct_asian", "pct_bach_deg25_over:pct_asian",
                    "pct_hs25_over:pct_black", "pct_bach_deg25_over:pct_black",
                    "pct_hs25_over:pct_other_race", "pct_bach_deg25_over:pct_other_race",
                    "pct_hs25_over:pct_white", "pct_bach_deg25_over:pct_white"),
                  labels = c(
                    "HS x Asian", "College x Asian",
                    "HS x Black", "College x Black",
                    "HS x Other", "College x Other",
                    "HS x White", "College x White"
                  ))
  )

# Visualize the bootstrap confidence intervals
bs %>%
  ggplot(aes(x = term, y = bs_mean)) +
  geom_pointrange(aes(ymin = left_bound, ymax = right_bound,
                      color = if_else(left_bound > 0 | right_bound < 0, "y", "n"))
                  ) +
  geom_hline(yintercept = 0, color = "red", alpha = 0.5) +
  coord_flip() +
  labs(
    title = "Bootstrap confidence intervals for 5000 resamples",
    x = "Interaction Term",
    y = "Bootstrap Interaction Coefficient Estimate",
    caption = "Figure XXX"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "none"
  ) +
  scale_color_manual(values = c("#DC143C", "#2E8B57"))
```

Prediction Model

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