

BST210 HW4 Project Check-In 1

Group 7: Regression Heroes

Ryan Wang

Stella Nam

Hongkai Wang

1. What is the general domain/subject area of this project?

The general domain/subject area of this project is cancer prevalence with respect to socioeconomic status in the United States.

2. What data will you use, and what is the source?

We are using data from “OLS Regression Challenge - dataset by nrppner | data.world”, which aggregates socioeconomic and clinical data from census.gov, clinicaltrials.gov, and cancer.gov. These cancer outcomes data were aggregated from cancer trials during 01/01/2010 through 06/01/2016 and socioeconomic and demographic data were aggregated from 2013 U.S. census data.

Source: <https://data.world/nrppner/ols-regression-challenge>

3. What primary questions will you seek to answer?

We are trying to explore the relationship between cancer-related death and the socioeconomic status of US counties. We will use predictors such as median income, age, and employment status to create a model that could predict the number of cancer diagnoses.

4. What secondary questions will you seek to answer?

Some secondary questions we are considering revolve around analysis of demographic data from the dataset. We are proposing to explore the relationship between race, income, and available medical resources in each county.

5. What outcome(s)/endpoint(s) will you use? (could be continuous, binary, polytomous, Poisson, survival,...and you may be considering more than one—and this may be updated/added to, as the semester progresses)

Our main outcome will be the cancer-related death rates in US counties from 2010 to 2016. This outcome is transformed from discrete count (count of cancer-related deaths) type, and we will use poisson distribution to model. For our secondary questions, we might

6. What is your draft Statistical Analysis Plan? (should be a very thorough, detailed, bullet point outline, demonstrating that you have broadly thought this through and included details - this may be updated of course as the semester proceeds, and with feedback) ** Note that we will be discussing all forms of outcome/endpoint data in this course, and at present have not yet covered each of these...so this plan may be updated/added to as the semester progresses, but you still should be able to plan out the structure and significant details of your plan. If your outcome data is other than continuous, you can still include for instance 'Regression modeling involving 'Y' outcome data of interest, involving these variables (list them)...' and any other concerns or methods of interest (listing potential confounders, effect modifiers, potential use of splines or additive modeling, potential missing data considerations, data reduction methods, regularization methods, etc,...or none of these—you will want to consider what is most appropriate for your data and questions at hand). Recall the BST 210 Regression Models Overview Table from which most extensions arise.

Analysis plan:

- Poisson regression (within a generalized linear framework) will be our main modeling approach. GAMs and splines could also be added to the model as we progress into the project.
- The cancer-related death counts could be seen as a Poisson distribution.
- Data cleaning and computing additional predictors (e.g., regions, climate ...)
- Create comprehensive visualization of our demographic data to illustrate differences between US counties.
- There are 3046 empty data points in our data. Currently, we are planning on learning and exploring different amputation techniques to adjust for these data points.
- Identify and adjust for potential confounders (on the association of SES to cancer prevalence) such as, but not limited to, county and income.

7. What are the biggest challenges you foresee in answering your proposed questions and completing this project? (logistical, statistical, etc, if there are any)

A challenge we could face is that our data is predominantly collected on the caucasian population rather than populations of ethnic minority groups. The lack of data of these underrepresented groups may lead to potential bias in our findings. The data on socioeconomic status is not stratified on ethnicity, making it difficult to determine outcomes by race. Additionally, a lot of the data is in percentages. Therefore, when we convert the data into counts, the data may not be independent between different variables.

8. Will you seek domain expertise? Why or why not? If so, from whom?

No, we will not seek domain expertise because the data is fairly general to all cancer cases. All our group members have some background in life sciences, molecular biology, and/or economics, which we believe is sufficient for the scope of the project. However, if need be, we can further scope the literature for additional information on cancer biology or economic terminology.

9. What software package(s) will you use to complete this project? (It is absolutely fine for different group members to use different packages; in fact, some tasks are easier in some packages over others and vice versa.)

We will use the R statistical programming language to analyze these data and complete this project. We plan on using tidyverse for transforming and cleaning data. Furthermore, this library of packages allows us to create effective EDA plots in order to explore our data efficiently. For modeling, we plan on using the `lm` and `glm` functions in base R for standard linear regression and Poisson regression (or other generalized models we may approach later), respectively.

10. Complete an initial round of exploratory analyses on your data that would be relevant to your plan and responses above, and include any plots, summaries, code and output. Please include exploratory analysis for outcome(s) of continuous form however/wherever possible even if your ultimate goals/questions involve a different form of outcome data such as binary, polytomous, etc. (You may consider this initial analysis as a potential sub-analysis later on.)

Exploratory Data Analysis

```
library(tidyverse)


## — Attaching packages — tidyverse
1.3.2 —
## ✓ ggplot2 3.3.6      ✓ purrr 0.3.4
## ✓ tibble 3.1.8      ✓ dplyr 1.0.10
## ✓ tidyr 1.2.0       ✓ stringr 1.4.1
## ✓ readr 2.1.2      ✓ forcats 0.5.2
## — Conflicts —
tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag() masks stats::lag()

library(stringr)
library(viridisLite) # nice colours
```

What's the structure of our data?

```
dat <- read_csv("cancer_reg.csv")

## Rows: 3047 Columns: 34
## — Column specification
## Delimiter: ","
## chr (2): binnedInc, Geography
## dbl (32): avgAnnCount, avgDeathsPerYear, TARGET_deathRate, incidenceRate,
me...
##
## ⓘ Use `spec()` to retrieve the full column specification for this data.
```

 Specify the column types or set `show_col_types = FALSE` to quiet this message.

str(dat)

```
## spec_tbl_df [3,047 × 34] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ avgAnnCount          : num [1:3047] 1397 173 102 427 57 ...
## $ avgDeathsPerYear     : num [1:3047] 469 70 50 202 26 152 97 71 36
1380 ...
## $ TARGET_deathRate     : num [1:3047] 165 161 175 195 144 ...
## $ incidenceRate        : num [1:3047] 490 412 350 430 350 ...
## $ medIncome            : num [1:3047] 61898 48127 49348 44243 49955 ...
## $ popEst2015           : num [1:3047] 260131 43269 21026 75882 10321
...
## $ povertyPercent       : num [1:3047] 11.2 18.6 14.6 17.1 12.5 15.6
23.2 17.8 22.3 13.1 ...
## $ studyPerCap          : num [1:3047] 499.7 23.1 47.6 342.6 0 ...
## $ binnedInc            : chr [1:3047] "(61494.5, 125635]" "(48021.6,
51046.4]" "(48021.6, 51046.4]" "(42724.4, 45201]" ...
## $ MedianAge            : num [1:3047] 39.3 33 45 42.8 48.3 45.4 42.6
51.7 49.3 35.8 ...
## $ MedianAgeMale        : num [1:3047] 36.9 32.2 44 42.2 47.8 43.5 42.2
50.8 48.4 34.7 ...
## $ MedianAgeFemale      : num [1:3047] 41.7 33.7 45.8 43.4 48.9 48 43.5
52.5 49.8 37 ...
## $ Geography            : chr [1:3047] "Kitsap County, Washington"
"Kittitas County, Washington" "Klickitat County, Washington" "Lewis County,
Washington" ...
## $ AvgHouseholdSize     : num [1:3047] 2.54 2.34 2.62 2.52 2.34 2.58
2.42 2.24 2.38 2.65 ...
## $ PercentMarried       : num [1:3047] 52.5 44.5 54.2 52.7 57.8 50.4
54.1 52.7 55.9 50 ...
## $ PctNoHS18_24        : num [1:3047] 11.5 6.1 24 20.2 14.9 29.9 26.1
27.3 34.7 15.6 ...
## $ PctHS18_24          : num [1:3047] 39.5 22.4 36.6 41.2 43 35.1 41.4
33.9 39.4 36.3 ...
## $ PctSomeCol18_24      : num [1:3047] 42.1 64 NA 36.1 40 NA NA 36.5 NA
NA ...
## $ PctBachDeg18_24      : num [1:3047] 6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4
7.1 ...
## $ PctHS25_Over        : num [1:3047] 23.2 26 29 31.6 33.4 30.4 29.8
31.6 32.2 28.8 ...
## $ PctBachDeg25_Over    : num [1:3047] 19.6 22.7 16 9.3 15 11.9 11.9
11.3 12 16.2 ...
## $ PctEmployed16_Over   : num [1:3047] 51.9 55.9 45.9 48.3 48.2 44.1
51.8 40.9 39.5 56.6 ...
## $ PctUnemployed16_Over : num [1:3047] 8 7.8 7 12.1 4.8 12.9 8.9 8.9
10.3 9.2 ...
## $ PctPrivateCoverage   : num [1:3047] 75.1 70.2 63.7 58.4 61.6 60 49.5
55.8 55.5 69.9 ...
```

```

## $ PctPrivateCoverageAlone: num [1:3047] NA 53.8 43.5 40.3 43.9 38.8 35
33.1 37.8 NA ...
## $ PctEmpPrivCoverage      : num [1:3047] 41.6 43.6 34.9 35 35.1 32.6 28.3
25.9 29.9 44.4 ...
## $ PctPublicCoverage       : num [1:3047] 32.9 31.1 42.1 45.3 44 43.2 46.4
50.9 48.1 31.4 ...
## $ PctPublicCoverageAlone : num [1:3047] 14 15.3 21.1 25 22.7 20.2 28.7
24.1 26.6 16.5 ...
## $ PctWhite                : num [1:3047] 81.8 89.2 90.9 91.7 94.1 ...
## $ PctBlack                : num [1:3047] 2.595 0.969 0.74 0.783 0.27 ...
## $ PctAsian                : num [1:3047] 4.822 2.246 0.466 1.161 0.666 ...
## $ PctOtherRace            : num [1:3047] 1.843 3.741 2.747 1.363 0.492 ...
## $ PctMarriedHouseholds    : num [1:3047] 52.9 45.4 54.4 51 54 ...
## $ BirthRate               : num [1:3047] 6.12 4.33 3.73 4.6 6.8 ...
## - attr(*, "spec")=
## .. cols(
## ..   avgAnnCount = col_double(),
## ..   avgDeathsPerYear = col_double(),
## ..   TARGET_deathRate = col_double(),
## ..   incidenceRate = col_double(),
## ..   medIncome = col_double(),
## ..   popEst2015 = col_double(),
## ..   povertyPercent = col_double(),
## ..   studyPerCap = col_double(),
## ..   binnedInc = col_character(),
## ..   MedianAge = col_double(),
## ..   MedianAgeMale = col_double(),
## ..   MedianAgeFemale = col_double(),
## ..   Geography = col_character(),
## ..   AvgHouseholdSize = col_double(),
## ..   PercentMarried = col_double(),
## ..   PctNoHS18_24 = col_double(),
## ..   PctHS18_24 = col_double(),
## ..   PctSomeCol18_24 = col_double(),
## ..   PctBachDeg18_24 = col_double(),
## ..   PctHS25_Over = col_double(),
## ..   PctBachDeg25_Over = col_double(),
## ..   PctEmployed16_Over = col_double(),
## ..   PctUnemployed16_Over = col_double(),
## ..   PctPrivateCoverage = col_double(),
## ..   PctPrivateCoverageAlone = col_double(),
## ..   PctEmpPrivCoverage = col_double(),
## ..   PctPublicCoverage = col_double(),
## ..   PctPublicCoverageAlone = col_double(),
## ..   PctWhite = col_double(),
## ..   PctBlack = col_double(),
## ..   PctAsian = col_double(),
## ..   PctOtherRace = col_double(),
## ..   PctMarriedHouseholds = col_double(),
## ..   BirthRate = col_double()

```

```

## .. )
## - attr(*, "problems")=<externalptr>

summary(dat)

##      avgAnnCount      avgDeathsPerYear TARGET_deathRate incidenceRate
## Min.   :    6.0    Min.   :    3    Min.   : 59.7    Min.   : 201.3
## 1st Qu.:   76.0    1st Qu.:   28    1st Qu.:161.2    1st Qu.: 420.3
## Median :  171.0    Median :   61    Median :178.1    Median : 453.5
## Mean   :   606.3    Mean   :  186    Mean   :178.7    Mean   : 448.3
## 3rd Qu.:  518.0    3rd Qu.:  149    3rd Qu.:195.2    3rd Qu.: 480.9
## Max.   :38150.0    Max.   :14010    Max.   :362.8    Max.   :1206.9
##
##      medIncome      popEst2015      povertyPercent      studyPerCap
## Min.   : 22640    Min.   :    827    Min.   : 3.20    Min.   :  0.00
## 1st Qu.: 38882    1st Qu.:  11684    1st Qu.:12.15    1st Qu.:  0.00
## Median : 45207    Median :   26643    Median :15.90    Median :  0.00
## Mean   : 47063    Mean   :  102637    Mean   :16.88    Mean   : 155.40
## 3rd Qu.: 52492    3rd Qu.:   68671    3rd Qu.:20.40    3rd Qu.:  83.65
## Max.   :125635    Max.   :10170292    Max.   :47.40    Max.   :9762.31
##
##      binnedInc      MedianAge      MedianAgeMale      MedianAgeFemale
## Length:3047    Min.   : 22.30    Min.   :22.40    Min.   :22.30
## Class :character 1st Qu.: 37.70    1st Qu.:36.35    1st Qu.:39.10
## Mode  :character Median : 41.00    Median :39.60    Median :42.40
##              Mean   : 45.27    Mean   :39.57    Mean   :42.15
##              3rd Qu.: 44.00    3rd Qu.:42.50    3rd Qu.:45.30
##              Max.   :624.00    Max.   :64.70    Max.   :65.70
##
##      Geography      AvgHouseholdSize PercentMarried      PctNoHS18_24
## Length:3047    Min.   :0.0221    Min.   :23.10    Min.   : 0.00
## Class :character 1st Qu.:2.3700    1st Qu.:47.75    1st Qu.:12.80
## Mode  :character Median :2.5000    Median :52.40    Median :17.10
##              Mean   :2.4797    Mean   :51.77    Mean   :18.22
##              3rd Qu.:2.6300    3rd Qu.:56.40    3rd Qu.:22.70
##              Max.   :3.9700    Max.   :72.50    Max.   :64.10
##
##      PctHS18_24      PctSomeCol18_24 PctBachDeg18_24      PctHS25_Over
## Min.   : 0.0    Min.   : 7.10    Min.   : 0.000    Min.   : 7.50
## 1st Qu.:29.2    1st Qu.:34.00    1st Qu.: 3.100    1st Qu.:30.40
## Median :34.7    Median :40.40    Median : 5.400    Median :35.30
## Mean   :35.0    Mean   :40.98    Mean   : 6.158    Mean   :34.80
## 3rd Qu.:40.7    3rd Qu.:46.40    3rd Qu.: 8.200    3rd Qu.:39.65
## Max.   :72.5    Max.   :79.00    Max.   :51.800    Max.   :54.80
##              NA's   :2285
##      PctBachDeg25_Over PctEmployed16_Over PctUnemployed16_Over
PctPrivateCoverage
## Min.   : 2.50    Min.   :17.60    Min.   : 0.400    Min.   :22.30
## 1st Qu.: 9.40    1st Qu.:48.60    1st Qu.: 5.500    1st Qu.:57.20
## Median :12.30    Median :54.50    Median : 7.600    Median :65.10

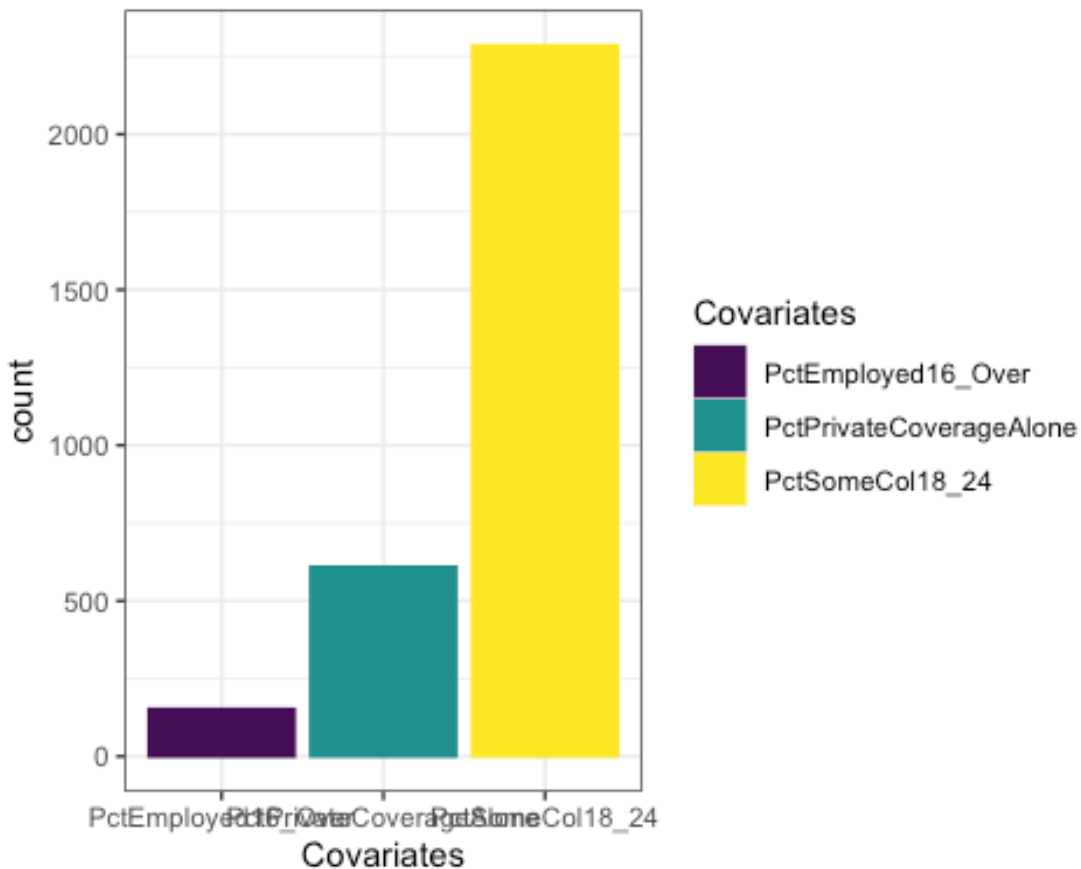
```

```
## Mean :13.28      Mean :54.15      Mean : 7.852      Mean :64.35
## 3rd Qu.:16.10    3rd Qu.:60.30    3rd Qu.: 9.700    3rd Qu.:72.10
## Max. :42.20      Max. :80.10      Max. :29.400      Max. :92.30
## NA's :152
## PctPrivateCoverageAlone PctEmpPrivCoverage PctPublicCoverage
## Min. :15.70      Min. :13.5      Min. :11.20
## 1st Qu.:41.00    1st Qu.:34.5    1st Qu.:30.90
## Median :48.70    Median :41.1    Median :36.30
## Mean :48.45      Mean :41.2      Mean :36.25
## 3rd Qu.:55.60    3rd Qu.:47.7    3rd Qu.:41.55
## Max. :78.90      Max. :70.7      Max. :65.10
## NA's :609
## PctPublicCoverageAlone PctWhite PctBlack PctAsian
## Min. : 2.60      Min. : 10.20    Min. : 0.0000    Min. : 0.0000
## 1st Qu.:14.85    1st Qu.: 77.30    1st Qu.: 0.6207    1st Qu.: 0.2542
## Median :18.80    Median : 90.06    Median : 2.2476    Median : 0.5498
## Mean :19.24      Mean : 83.65     Mean : 9.1080     Mean : 1.2540
## 3rd Qu.:23.10    3rd Qu.: 95.45    3rd Qu.:10.5097    3rd Qu.: 1.2210
## Max. :46.60      Max. :100.00     Max. :85.9478     Max. :42.6194
##
## PctOtherRace PctMarriedHouseholds BirthRate
## Min. : 0.0000    Min. :22.99     Min. : 0.000
## 1st Qu.: 0.2952    1st Qu.:47.76     1st Qu.: 4.521
## Median : 0.8262    Median :51.67     Median : 5.381
## Mean : 1.9835     Mean :51.24      Mean : 5.640
## 3rd Qu.: 2.1780    3rd Qu.:55.40     3rd Qu.: 6.494
## Max. :41.9303     Max. :78.08      Max. :21.326
##
```

From the summary above, we can observe that we have a lot of variables to consider to answer our main questions. Additionally, it tells us about the complexity about the data itself in that cancer prevalence is a factor of many socioeconomic factors.

How much missing data is there?

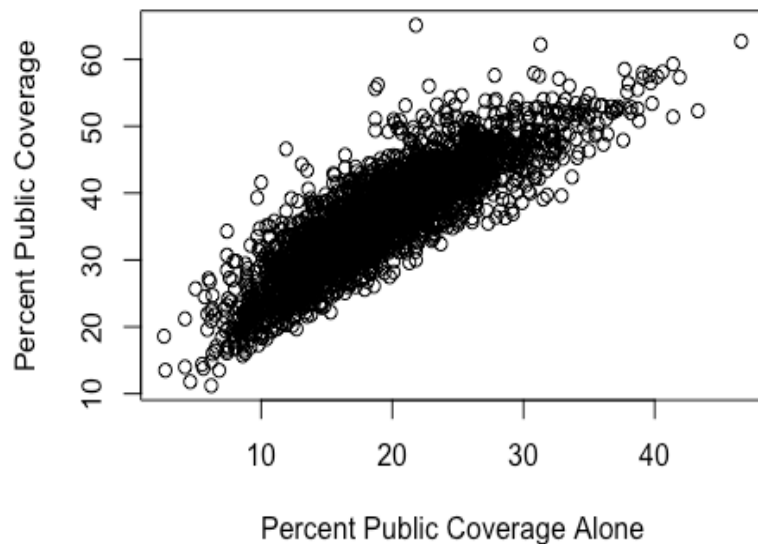
```
no_geodat <- dat %>% select(-c("Geography", "binnedInc")) # no missing
counties
no_geodat[!complete.cases(no_geodat),] %>% # keep rows with NAs
  pivot_longer(colnames(no_geodat), names_to = "Covariates", values_to =
"Values") %>% # pivot into long table
  filter(is.na(Values)) %>% # filter out all the non-na's
  ggplot(aes(x = Covariates)) +
  geom_bar(position = "dodge", aes(col = Covariates, fill = Covariates)) +
  theme_bw() +
  scale_fill_viridis_d() +
  scale_color_viridis_d()
```



Interestingly, there is some missing data in the reporting of county employment for residents age 16 and over. We observe more missing data in percentages of county residents with only private healthcare coverage, and missing data in the majority of counties for reports of the percent of county residents between 18 and 24 years old with with some college as their highest attained education. One hypothesis for why there may be a large amount of missing data in educational reporting for 18 to 24 year olds with some college may be because some counties may not have high emphasis on higher education and many individuals in this age range that are pursuing college will probably be in more college-oriented counties.

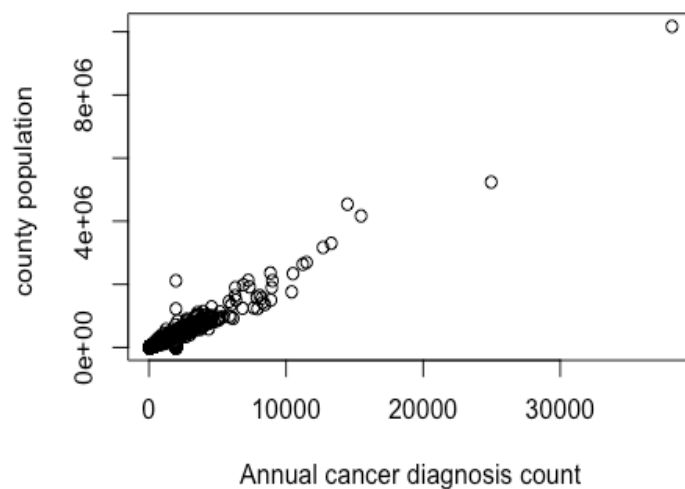
Exploring data patterns and modeling

```
pairs(dat %>% select(where(is.numeric)))
```

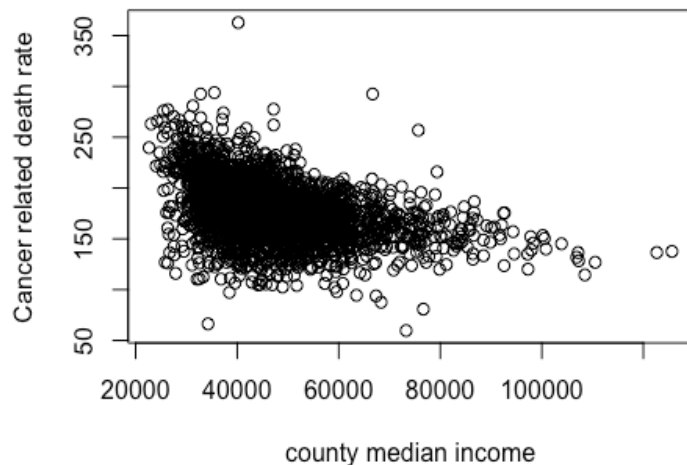
It is self explanatory that the higher the percentage of the county population is under public healthcare, the higher the percentage of the county single persons is under public healthcare.

```
plot(dat$avgAnnCount, dat$popEst2015, xlab = "Annual cancer diagnosis count",
ylab = "county population ")
```



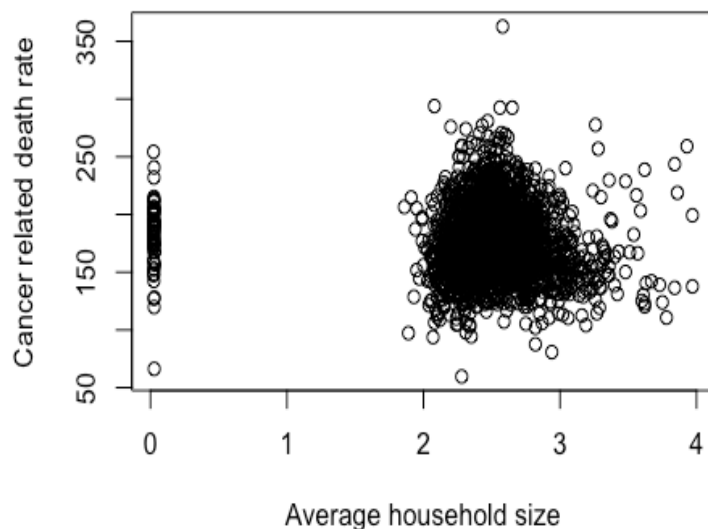
This is another expected correlation. The higher the population is, the more cancer occurrence will be in the population.

```
plot(dat$medIncome, dat$TARGET_deathRate, ylab = "Cancer related death rate",  
xlab = "county median income")
```



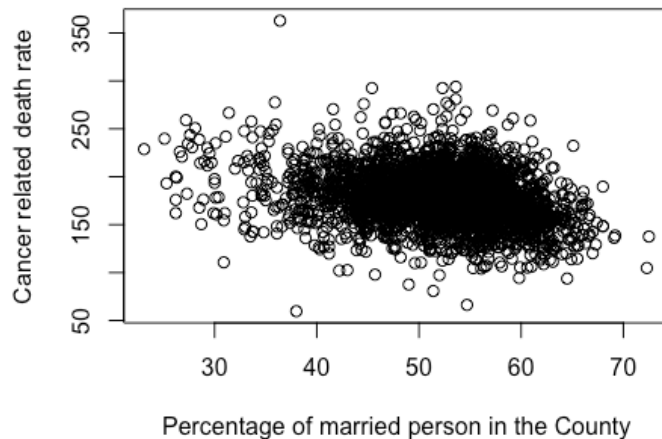
This is an example of significant predictors for our model. Annual income is likely very correlated with the # of cancer related death rate. However, we need to further explore what kind of model will be best fit for the data.

```
plot(dat$AvgHouseholdSize, dat$TARGET_deathRate, ylab = "Cancer related death  
rate", xlab = "Average household size")
```



As the plot suggests, there isn't a clear relationship between cancer-related death rate and the average household size. However, further work could help illustrate some latent meanings.

```
plot(dat$PercentMarried, dat$TARGET_deathRate, ylab = "Cancer related death rate", xlab = "Percentage of married person in the County")
```



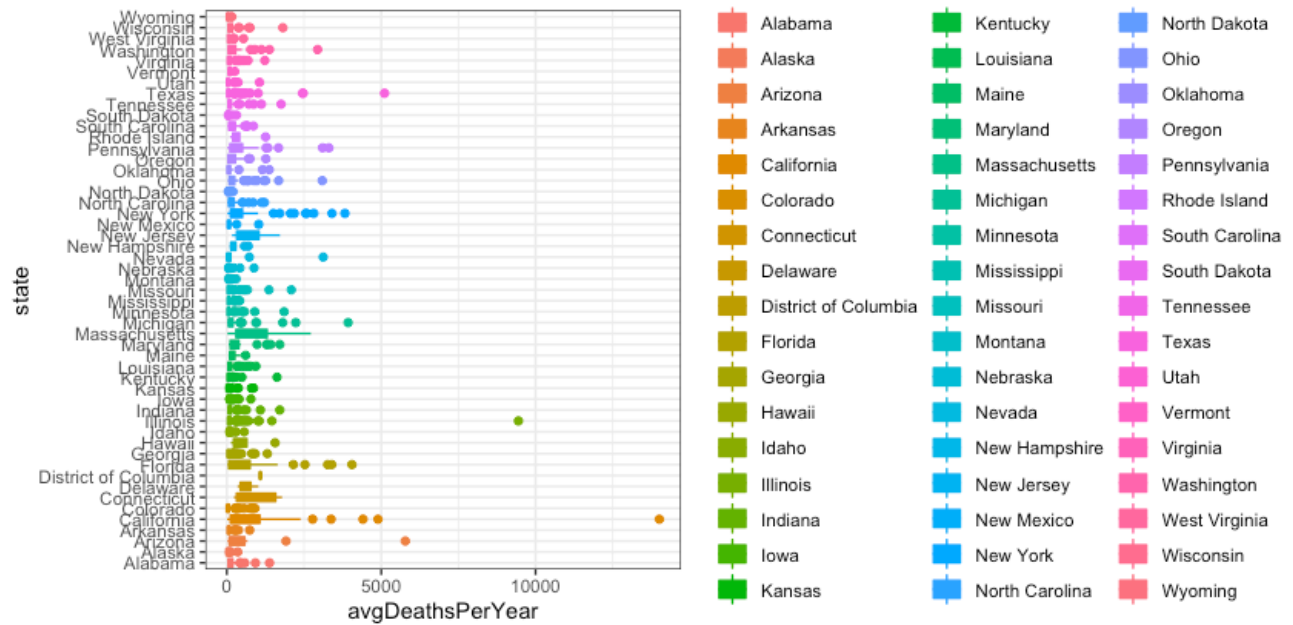
It seems that there is a negative association between percent of people married and cancer-related death rate. This seems to be another valuable predictor that we could use in our model.

These graphs are a brief look into the possible associations within our data set. We need to be careful that some of these relationships might be confounded by other co-variables. We also need to conduct confounding analysis (both classical, and statistical), effect modifier analysis to figure out the nature of the associations.

[Look at the state-wide view](#)

```
dat <- dat %>%
  mutate(state = str_extract(Geography, "[^,]+$")) %>% # regex to select
  # everything after ','
  mutate(state = str_trim(state))

dat %>%
  ggplot(aes(x = state, y = avgDeathsPerYear, colour = state, fill =
state)) +
  geom_boxplot() +
  theme_bw() +
  coord_flip()
```



It appears that there are some potential outliers when studying the relationship between states and average deaths per year due to cancer. Particularly, there may be potential outliers in the data for California, Illinois, and Arizona. We would need to further examine the reasons for such a large deviation from the rest of the data. Additionally, it appears that the more populated states, such as California and New York, have a higher average death rate compared to less populated states like Wyoming.

11.

We attest that no member of this group is using these data or same/similar questions in any other course or course project, at HSPH.

Hongkai Wang, Stella Nam, Ryan Wang