Project Proposal

Revised Proposal

Probability Pandas

10/17/2021

Load Packages

```
library(tidyverse)
```

Load Data

```
initial_data = read.csv("~/Probability Pandas Project/data/500_Cities.csv")

cities <- initial_data %>%
  select(StateAbbr, PlaceName, PlaceFIPS, Geolocation, ACCESS2_AdjPrev,
         ACCESS2_Adj95CI, CANCER_AdjPrev, CANCER_Adj95CI, CHD_AdjPrev,
         CHD_Adj95CI, CHECKUP_AdjPrev, CHECKUP_Adj95CI, COPD_AdjPrev,
         COPD_Adj95CI, COLON_SCREEN_AdjPrev, COLON_SCREEN_Adj95CI,
         COREM_AdjPrev, COREM_Adj95CI, COREW_AdjPrev, COREW_Adj95CI,
         KIDNEY_AdjPrev, KIDNEY_Adj95CI, MAMMOUSE_AdjPrev, MAMMOUSE_Adj95CI,
         PAPTEST_AdjPrev, PAPTEST_Adj95CI) %>%
  rename(state = StateAbbr, city = PlaceName,
         health_access = ACCESS2_AdjPrev,
         health_access_CI = ACCESS2_Adj95CI, cancer = CANCER_AdjPrev,
         cancer_CI = CANCER_Adj95CI, heart_disease = CHD_AdjPrev,
         heart_disease_CI = CHD_Adj95CI, checkup = CHECKUP_AdjPrev,
         checkup_CI = CHECKUP_Adj95CI, chronic_lung_disease = COPD_AdjPrev,
         chronic_lung_disease_CI = COPD_Adj95CI,
         colon_screen = COLON_SCREEN_AdjPrev,
         colon_screen_CI = COLON_SCREEN_Adj95CI,
         men_colorectal_cancer_screen = COREM_AdjPrev,
         men_colorectal_cancer_screen_CI = COREM_Adj95CI,
         women_colorectal_cancer_screen = COREW_AdjPrev,
         women_colorectal_cancer_screen_CI = COREW_Adj95CI,
         chronic_kidney_disease = KIDNEY_AdjPrev,
         chronic_kidney_disease_CI = KIDNEY_Adj95CI,
         mammogram = MAMMOUSE_AdjPrev,
         mammogram_CI = MAMMOUSE_Adj95CI, pap_test = PAPTEST_AdjPrev,
         pap_test_CI = PAPTEST_Adj95CI)
```

Introduction and Data, including Research Questions

This data is from the US Centers for Disease Control and Prevention, Epidemiology and Surveillance Branch. It contains data of the 500 largest cities in the United States. The cases are each of the 500 cities and the variables include prevalence of lack of health insurance, prevalence of clinical preventive services, screening services, and routine doctor visits, and prevalence of various chronic diseases, including but not limited to cancer, coronary heart disease, obstructive pulmonary disease, and kidney disease. The data set also includes the 95% confidence interval for the various variables in the data.

How does health insurance correlate to the prevalence of screening for various chronic diseases? Our hypothesis is that increased health insurance is correlated to increased screening for chronic diseases. The inverse also holds: decreased health insurance is correlated to decreased screening.

Answering this research question may indicate to policy makers that wider access to health insurance would prevent deaths due to chronic disease. Lower mortality from chronic disease is correlated to high prevalence of screening, and vice versa, as seen in recent literature. By showing the relationship of health insurance to screening, we may by extension show a link from health insurance to mortality of various chronic diseases.

The link between screening and the mortality of various chronic diseases is seen in the following examples:

"Women who participated in mammography screening had a statistically significant 41% reduction in their risk of dying of breast cancer within 10 years and a 25% reduction in the rate of advanced breast cancers" according to a July 1, 2020 article in the American Cancer Society Journal Cancer, titled "Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women". According to a study published in the peer-reviewed journal PLOS One, in an article titled "Colonoscopy reduces colorectal cancer mortality: A multicenter, long-term, colonoscopy-based cohort study," the colorectal cancer mortality of patients who received a colonoscopy was significantly less than those who did not receive a colonoscopy. According to a study published in the journal BMC Public Health, cervical cancer mortality was shown to have declined following the implementation of a papanicolaou test screening program. "Fecal occult blood testing reduces colorectal cancer mortality by 16 percent," according to a study published in a June 2007 article of the peer-reviewed journal American Family Physician. A 2014 study published in World Journal of Gastroenterology showed "a 28% risk reduction in overall CRC [colorectal cancer] mortality" among those who underwent sigmoidoscopy screening.

Glimpse

```
glimpse(cities)
```

```
## Rows: 500
## Columns: 26
## $ state                          <chr> "CA", "FL", "CA", "CA", "FL", "FL", ~
## $ city                           <chr> "Folsom", "Largo", "Berkeley", "Napa~
## $ PlaceFIPS                      <int> 624638, 1239425, 606000, 650258, 126~
## $ Geolocation                    <chr> "(38.67504943280, -121.147605753)", ~
## $ health_access                  <dbl> 7.7, 20.9, 7.1, 12.7, 23.3, 22.0, 25~
## $ health_access_CI               <chr> "( 7.2,  8.2)", "(20.4, 21.5)", "( 6~
## $ cancer                         <dbl> 6.2, 6.3, 6.0, 6.1, 5.8, 5.6, 5.4, 5~
## $ cancer_CI                      <chr> "( 6.1,  6.3)", "( 6.3,  6.4)", "( 5~
## $ heart_disease                  <dbl> 4.4, 6.7, 4.3, 5.3, 5.9, 5.3, 7.9, 6~
## $ heart_disease_CI               <chr> "( 4.3,  4.6)", "( 6.5,  6.8)", "( 4~
## $ checkup                        <dbl> 65.3, 73.6, 66.8, 62.8, 76.8, 76.3, ~
## $ checkup_CI                     <chr> "(65.1, 65.6)", "(73.4, 73.8)", "(66~
## $ chronic_lung_disease           <dbl> 4.2, 8.1, 4.1, 5.6, 6.5, 5.3, 9.3, 6~
## $ chronic_lung_disease_CI        <chr> "( 4.0,  4.4)", "( 7.8,  8.3)", "( 4~
## $ colon_screen                   <dbl> 77.7, 62.6, 74.6, 69.3, 59.9, 62.2, ~
```

```
## $ colon_screen_CI                   <chr> "(76.9, 78.4)", "(61.9, 63.3)", "(74~
## $ men_colorectal_cancer_screen      <dbl> 37.5, 33.9, 38.1, 38.3, 30.6, 31.2, ~
## $ men_colorectal_cancer_screen_CI   <chr> "(35.4, 39.5)", "(32.7, 35.1)", "(37~
## $ women_colorectal_cancer_screen    <dbl> 34.2, 34.4, 37.5, 31.4, 27.4, 28.0, ~
## $ women_colorectal_cancer_screen_CI <chr> "(32.5, 35.8)", "(33.3, 35.4)", "(36~
## $ chronic_kidney_disease            <dbl> 2.2, 2.8, 2.4, 2.6, 3.0, 2.7, 4.0, 3~
## $ chronic_kidney_disease_CI         <chr> "( 2.1,  2.2)", "( 2.8,  2.9)", "( 2~
## $ mammogram                         <dbl> 78.8, 71.3, 78.2, 73.4, 79.0, 80.3, ~
## $ mammogram_CI                      <chr> "(77.9, 79.7)", "(70.6, 72.0)", "(77~
## $ pap_test                          <dbl> 82.9, 71.5, 81.3, 81.1, 77.0, 79.6, ~
## $ pap_test_CI                       <chr> "(82.3, 83.5)", "(70.9, 72.2)", "(80~
```

Data Analysis Plan

The table shows the prevalence of lack of health care access in each state, as well as the prevalence for chronic disease screening in each state. This table is general summarization of the data given by city in the data set to the states that they cities reside in. The first plot explores the relationship between lack of health care access and at least yearly check ups with a doctor. While there doesn't seem to be any correlation between these two variables, we can see a correlation between lack of health care access and also a decrease in mammograms. This suggests that the two variables may be positively related.

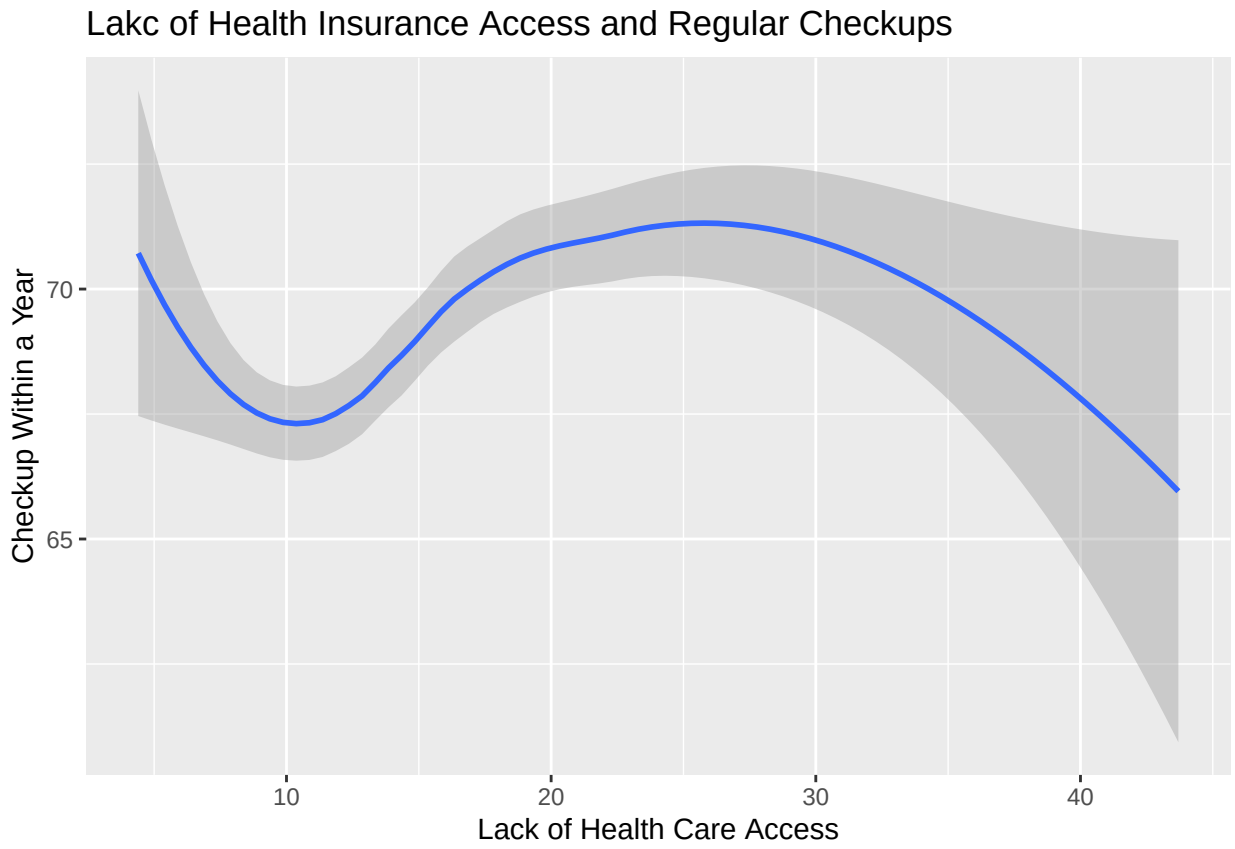
```
# Prevalence of Lack of Health Insurance Access and Chronic Disease Screening by State

cities %>%
  group_by(state) %>%
  summarise(state_prev = mean(health_access),
            colon_screen = mean(colon_screen),
            men_colorectal_cancer_screen = mean(men_colorectal_cancer_screen),
            women_colorectal_cancer_screen = mean(women_colorectal_cancer_screen),
            mammogram_screen = mean(mammogram, na.rm = TRUE),
            cervical_cancer_screen = mean(pap_test, na.rm = TRUE))
```

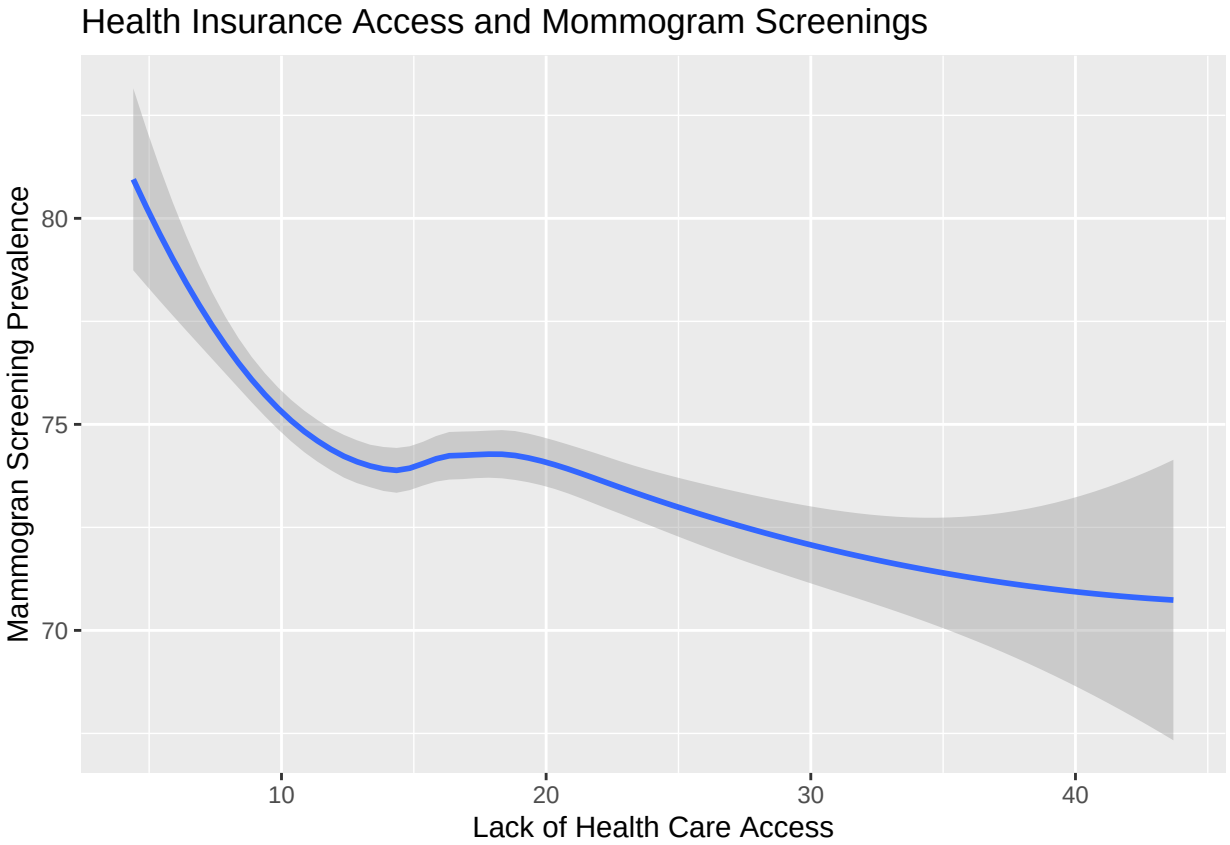
```
## # A tibble: 51 x 7
##    state state_prev colon_screen men_colorectal_cancer_screen women_colorectal_~
##    <chr>      <dbl>        <dbl>                        <dbl>              <dbl>
##  1 AK          12.6         63.8                         39.7               33.4
##  2 AL          16.3         67.0                         37.7               33.5
##  3 AR          13.7         61.4                         36.2               29.6
##  4 AZ          14.4         63.3                         36.0               32.5
##  5 CA          13.6         66.6                         32.2               32.7
##  6 CO          12.6         63.9                         37.2               34.8
##  7 CT          14.1         67.4                         30.1               29.7
##  8 DC           7.7         67.4                         30.6               28.4
##  9 DE          17.1         60.1                         32.2               31.4
## 10 FL          23.1         60.9                         31.5               30.2
## # ... with 41 more rows, and 2 more variables: mammogram_screen <dbl>,
## #   cervical_cancer_screen <dbl>
```

```
# Exploratory Visualizations

ggplot(cities) +
  geom_smooth(aes(x = health_access, y = checkup)) +
  labs(title = "Lakc of Health Insurance Access and Regular Checkups",
       x = "Lack of Health Care Access",y = "Checkup Within a Year")
```




```
ggplot(cities) +
  geom_smooth(aes(x = health_access, y = mammogram)) +
  labs(title = "Health Insurance Access and Mommogram Screenings",
       x = "Lack of Health Care Access", y = "Mammogran Screening Prevalence")
```



The predictor would be the lack of health care access and the outcome would be the prevalence for various chronic disease screenings.

We will be looking at a potential correlation between lack of health insurance access and screening access. The predictor variable we will use is the prevalence of lack of health care in the various cities. The outcome variables we will use are the prevalence of the various screening methods for chronic diseases. To answer our questions, we will use the statistical methods of modeling/visualizing and two sided t-tests.

The results from our modeling/visualizing would support our hypothesis if they show a negative correlation between prevalence of lack of health insurance and prevalence of the many screening methods. We are able to draw statistical conclusions from the predictor variable, lack of health insurance access, because as indicated from the article, "Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators From the Behavioral Risk Factor Surveillance System" by Zhang et al., the data is drawn from a survey question that asks, "Do you have any kind of health-care coverage, including health insurance, prepaid plans such as health maintenance organizations, or government plans such as Medicare or the Indian Health Service?". This indicates that this is an independent variable and new statistics can be drawn when comparing to other data within the dataset, because its value was not extrapolated from those variables which we will be comparing it to. We would use the 2-sided t-test to observe any significant differences in the prevalence of health care with a variety of screening methods and their relation to chronic diseases.