Cardiovascular Death Rate Capstone: Statistical Analysis

Zach Wimpee

November 5, 2020

1 Overview

Recall, **the primary goal of this project is to build a machine learning model that predicts the change in cardiovascular death rates (CDR) at the county level.** Some insight was gained from the data story about possible relationships between change in CDR (the dependent variable) and a set of census variables (independent variables). Statistical analysis techniques can be used to further explore these relationships, and can help determine if the data are appropriate for building a predictive model.

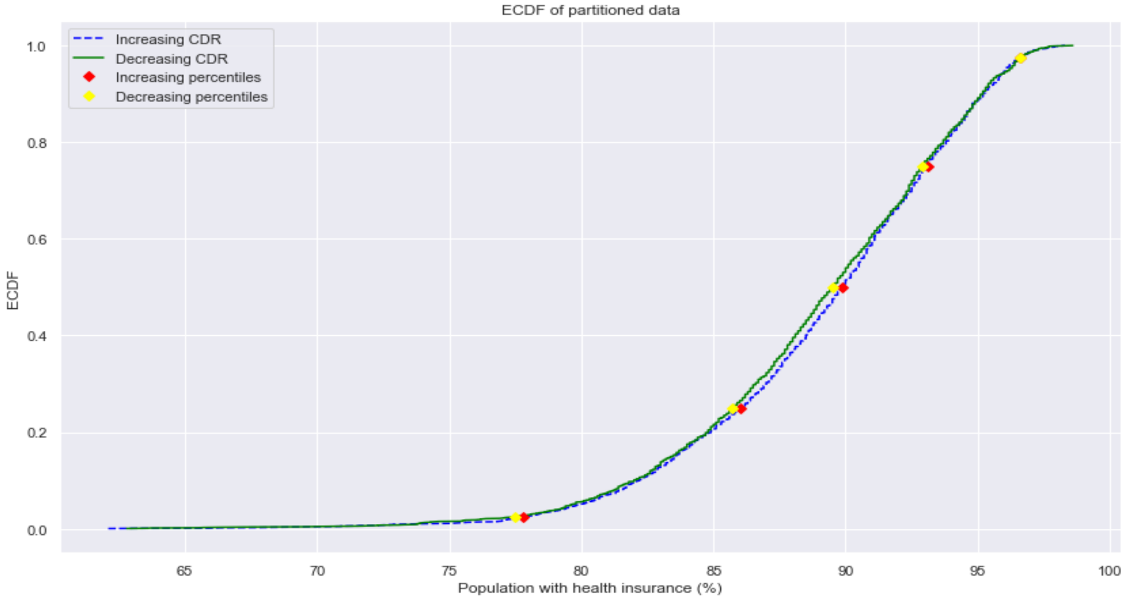
2 Hypothesis Testing

2.1 Null Hypothesis

Partition the data into two sets: One being the set of counties that experience an increase in CDR from the previous year they appear in the data, and the other being the set of counties that do not experience an increase in CDR. Both sets have the same set of feature columns (census variables). **I hypothesize that the distribution of each feature is identical between these two data sets.**

2.1.1 Example Feature: Percent with Health Insurance

Consider the distribution of a single census variable, *percent of population with health insurance.* Below is the ECDF plot for the data partitioned by CDR change.



2.2 Test Statistic

The null hypothesis is rejected if any feature has different distributions for the partition. To compare the feature distributions, the **sample statistic** T^F for feature F can be defined as:

$$T^F = \overline{F^{incr}} - \overline{F^{decr}}$$

where,

- $\overline{F^{incr}} \equiv$ Mean feature value for set of CDR increases
- $\overline{F^{decr}} \equiv$ Mean feature value for set of CDR decreases

For $F \equiv$ *percent of population with health insurance*, the observed value of this sample statistic is $T^F_{obs} = 0.211$. How reasonable is this value under the assumption of the null hypothesis?

2.2.1 Bootstrap Replicates

A new partition $\{\widetilde{F_i^{incr}}, \widetilde{F_i^{decr}}\}$, $i \in Z^+$, can be constructed by generating replicates of partition $\{F^{incr}, F^{decr}\}$ under the assumption of the null hypothesis. The sample statistic T^F_i can be calculated on this replicate:

$$T^F_i = \overline{\widetilde{F_i^{incr}}} - \overline{\widetilde{F_i^{decr}}}$$

Repeating this process N times results in a set of values for the sample statistic:

$$T^F_{bs} = \left\{ T^F_i \middle| i = 1, 2, \ldots, N \right\},$$

This set describes the distribution test statistic bootstrap replicates under the assumption of the null hypothesis. How does T^F_{obs} compare to this distribution for $F \equiv$ *percent of population with health insurance*?

2.3 p-value: Feature - Percent with Health Insurance

The p-value, denoted p, of T^F_{obs} for feature $F \equiv$ *percent of population with health insurance* is computed by:

$$p = \frac{|T'^F_{bs}|}{N}$$

Where,

$$T'^F_{bs} = \left\{ T^F_i \in T^F_{bs} \middle| T^F_i > T^F_{obs} \right\}$$

For $N = 10,000$, the p-value is $p = 0.0909$, and therefore feature $F \equiv$ *percent of population with health insurance* does not reject the null hypothesis.

2.4 p-value: Other Features

Repeating the process described above on the other features ...

- *mean household Social Security income*: $p = 0.0001$
- *SMOCAPI less than 10 percent*: $p = 0.0002$

These features result in p-values that suggest rejection of the null hypothesis. There are numerous other features that also reject the null hypothesis, but even just one such feature would suffice.