Cancer Regression Project

Paul Sciarpelletti, Angel Yu, Bibiana Cortes

26/02/2022

Description of dataset:

Source: https://data.world/nrippner/ols-regression-challenge

This dataset contains lung cancer related data from 2010-2016 based on regions of the United States. The data source is the 2013 US census, government clinical trials and the National Cancer Institute . The cancer related variables we will focused on are "avgAnnCount", "avgDeathsPerYear", "incidenceRate" and the created population adjusted variable "adjustedDeathsPerYear". Below is the list of some variables in the dataset; description (a) refers to cancer related variables, and description(b) refers to unrelated to cancer explanatory variables:

- TARGET_deathRate: Response variable. Mean per capita (100,000) cancer mortalities(a)

(This response variable is different than the *adjustedDeathsPerYear response variable, even though they should theoretically be the same. The meta data does not describe many differences and we believe, based on the name "TARGET_deathrate", it might differ based on expected population growth in the respective region. We will model and compare both variables through statistical methods and decide on which response variable is optimal.)

- avgAnnCount: Mean number of reported cases of cancer diagnosed annually(a)
- avgDeathsPerYear: Mean number of reported mortalities due to cancer(a)
- incidenceRate: Mean per capita (100,000) cancer diagnosis(a)
- medianIncome: Median income per county (b)
- povertyPercent: Percent of populace in poverty (b)
- studyPerCap: Per capita number of cancer-related clinical trials per county (a)
- MedianAge: Median age of county residents (b)
- PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
- PctPrivateCoverage: Percent of county residents with private health coverage (b)
- PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance) (b)
- PctPublicCoverage: Percent of county residents with government-provided health coverage (b)
- PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone (b)

- PctWhite: Percent of county residents who identify as White (b)
- PctBlack: Percent of county residents who identify as Black (b)
- PctAsian: Percent of county residents who identify as Asian (b)
- PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian (b)
- *adjustedDeathsPerYear: number of reported mortalities due to cancer divided by population then times by 100 000 to be mean number of deaths per capita. (a)

Analysis questions

The aim of this project is to answer the question of what regressor or combination of regressors best predicts the death rate due to cancer.

Regression techniques

We started our project by viewing the raw data. From a first look, we noticed that the data is based on regions with varying populations. Because many population adjusted variables come in the form of percentages, we needed to adjust the death rate in order to perform regression. Therefore, we created adjustedDeathsPerYear as a relative variable which is based on cancer deaths per capita (100 000 people). This variable and incidenceRate will be the key response variables. We will use a pair plot diagram as a primary investigation into linearity in the data.

We will fit some simple and multilinear regression models to find the best model. Next, we will asses the accuracy of the coefficient estimates by performing hypothesis testing using the p-value method. Also, for assessing accuracy of the model and comparing models we will use the adjusted R-squared statistics and stepwise selection techniques.

Next, we will access all assumptions for linear models: linearity between independent and dependent variables, independence of errors, homoscedasticity, normality of errors distribution and no multicollinearity. For this, we will use plots such as scatter plots, residuals versus fitted values, QQ plots, Scale-location plot. Also, we will use Cook-distance plots and residuals versus leverage plot to detect outliers and influential points and remove outliers if we see fit. For multicollinearity, we will use scatter plots to check correlation between variables, we will check for high standard errors for the regression coefficients and large changes in coefficients when adding predictors, and also we can use the variance inflation factor VIF.

Ultimately, we will seek to prove linearity through OLS regression and minimize the mean squared error. By doing so, we hope to also see insight into some of the (b) variables that are unrelated to cancer but provide statistical significance in estimating the cancer incident rate and cancer deaths. Such insight will show some root causes of cancer mortalities beyond biological explanations and expose potential previously overlooked at risk populations. This will hopefully provide directions for future public health investigations in the fight against lung cancer and lung cancer mortalities.