Regression hero

Hongkai Wang, Ryan Wang, Stella Nam

12/10/2022

**Introduction:**

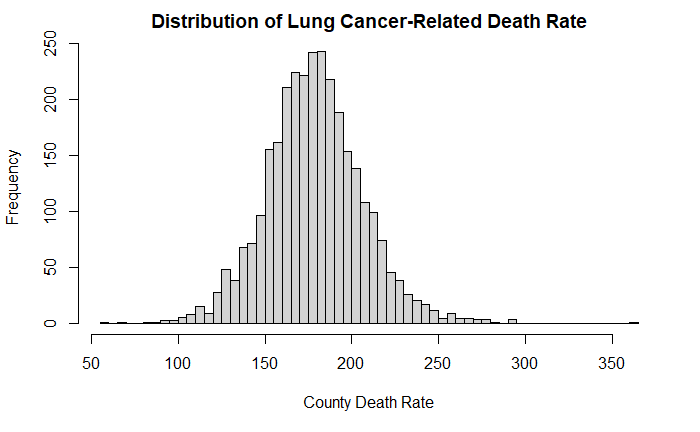
**Review of literature and domain expertise**

**Research and analysis methods**

For linear models, attempts were made to create models using GAM and smoothing methods. However, there is very little merit to using the more complicated GAM or smoothing methods, as more of the significant relationships appear to be in a fairly linear relationship.

**Findings and analysis**

**linear regression model:**

Conventionally, count data is best modeled after a Poisson regression model due to the presence of high zero values count. Disease death rate is a transformation of death count data, and thus, its distribution could still be modeled with a Poisson distribution. Specifically, our data set shows a normal distribution of the United States County level lung cancer-related death rates.

As a result, our group is curious whether a linear regression model could adequately describe the county death rate based on many demographic predictor values. We first fitted a multivariate linear regression model with describing county death count using county median income, median age, demographic percentage compositions, education level percentages, health insurance coverage status, and marriage status. These predictor variables were chosen by a stepwise and bottom-up approach.

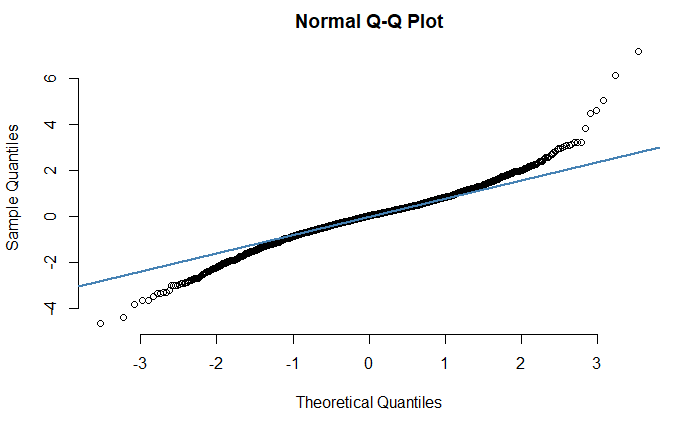
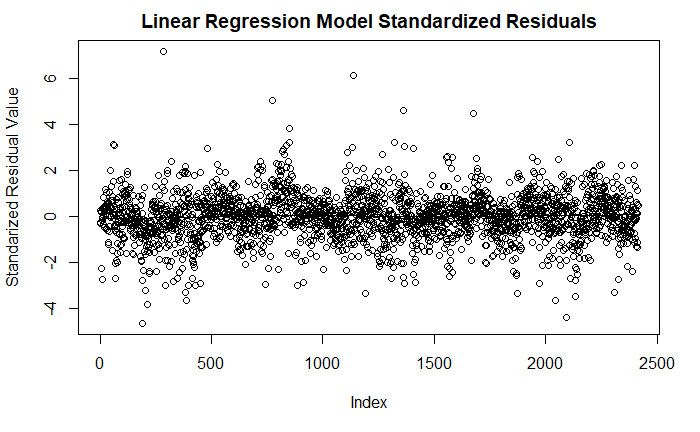
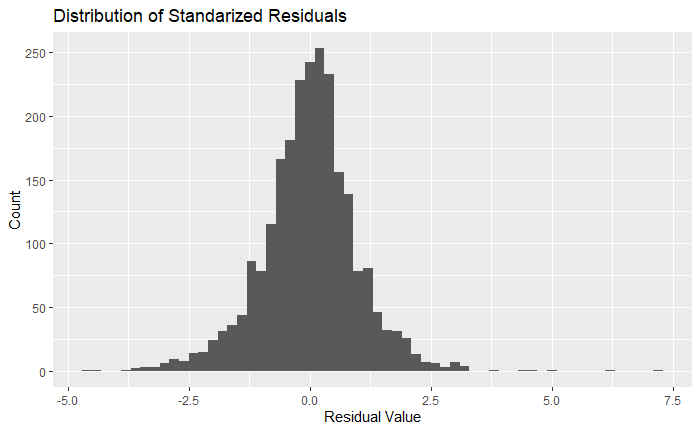
The fitted model has the form:

|  |  |  |
| --- | --- | --- |
|  | Variable Name | Estimate |
|  |  | 1.332 \* 102 |
|  |  | -1.137 \*10­-3 |
|  |  | 8.626\*10-9 |
|  |  | 1.250 |
|  |  | -1.533\*10-1 |
|  |  | 3.818\*10-1 |
|  |  | 1.167 |
|  |  | -1.531\*10-1 |
|  |  | -1.456\*10-1 |
|  |  | 8.555\*10-1 |
|  |  | 3.123\*10-2 |
|  |  | -2.955\*10-2 |

The overall regression model was statistically significant (R2 = 0.349, F(12, 2402) = 108.7, p < 2.2\*10-16).

During the model fitting process, we thoroughly considered the potential for confounders and effect modifiers for each of the predictor variables. Most notably, we found that percentage of the married population is an effective modifier of the median age covariate. We also noted that the percentage of private and public health insurance coverage were significant predictors, but a more appropriate model could be established by adding a product term of the two covariates.

During the model fitting process, we were also interested in whether generalized additive models could be used to increase the model fit. However, after using LOWESS flexible methods on the predictor variables, we failed to notice the need for smoothing models like splines. The relationships between the predictor variables and county death rates could be mostly explained by some simple linear relationships, and we decided to avoid usage of complex smoothing models and maintain the expandability of our model.

****

Diagnostic plots for our multiple linear regression model is produced.

Text

Description automatically generated

**Discussion.**

**Limitations**

**Future scope:**

**Tables and figures.**

**References.**

**Appendix.**