DATA 410 Project:

Investigating Social Factors on Lung Cancer Death Rate

An in depth dive into the potential causes of lung cancer deaths based on socioeconomic variables.

By Paul Sciarpelletti, Angel Yu, Bibiana Cortes

Table of Contents

Dataset Introduction

This dataset was created by Noah Rippner, who aggregated information between the years 2010 to 2016 from clinicaltrials.gov, cancertrials.gov, cancer.gov, and census.gov. The challenge issued by the original author was to examine lung cancer mortality's correlation with social factors such as education, income, race, etc. The dataset is organised so that each row describes a county. There are 3047 counties in this dataset. During our initial analysis, we followed the author's challenge in trying to find regressors that would predict the death rate per capita, which is 100,000 people in this case. However, with further analysis, we further cleaned the dataset and created a new target variable to regress onto which is deathRatePerIncidence (Target death rate/ incidence rate). The target variable represents the survival rate once a person is diagnosed; deathRatePerIncidence would best represent the counties' effectiveness in treating lung cancer. Given the broadness of the variables collected in this dataset and socio economic nature of the variables, this project aims to discover potential patterns in what groups are most at risk versus others with regards to lung cancer related death.

Data Cleaning

There were many discrepancies in how the data was aggregated, such as the fact that some factors were reported as normalised percentages, but others were reported per capita or as raw numbers. Because of this, we dropped quite a few variables that were definitely collinear. This was confirmed by looking at their correlation values and plots. An example of this would be dropping PctMarriedHouseholds and keeping PercentMarried after seeing an almost perfect correlation plot. We used a weighted average for the four variables describing education levels in ages 18 to 24 into a Likert scale. The variables describing percentages of different races in each county were normalised and assumed to be mutually exclusive; we kept only PctWhite.

During our initial analysis where our target variable was just deathRate, we found a high VIF value for incidenceRate. This, combined with the discrepancies in units in the variables, led us to combine incidenceRate with deathRate to become deathRatePerIncidence. This is also a more interesting target variable because we can now make hypotheses and inferences based on per diagnoses. At the end of this data cleaning, we ended up with the following variables in our dataset, all of which are continuous.

Table 1. This table lists the variables and their corresponding description in the cleaned dataset.

Variable	Description
deathRatePerIncidence	Mean number of deaths per number of diagnoses
medIncome	Median income per that county
povertyPercent	Percent of populace in poverty
studyperCap	Number of cancer related clinical trials per capita (100,000)
MedianAge	Median age of county residents
AvgHouseholdSize	Mean household size of county
PercentMarried	Percent of county residents who are married
edScore	Education scale- Likert scale of highest education level attained in populace between the ages of 18 to 24: 1 = No high school diploma 2 = High school diploma 3 = Some college 4 = Bachelor's degree
PctHS25_Over	Percent of county residents ages 25 and over with the highest education attained being a high school diploma
PctBachDeg25_Over	Percent of county residents ages 25 and over with the highest education attained being a bachelor's degree
PctEmployed16_Over	Percent of county residents ages 16 and over employed
PctUnemployed16_Over	Percent of county residents ages 16 and over unemployed
PctEmpPrivCoverage	Percent of county residents with employee-provided private health coverage
PctPublicCoverageAlone	Percent of county residents with only government provided health coverage
PctWhite	Percent of county residents who identify as White
BirthRate	Number of live births relative to number of women in the county

Hypothesis

The variables described in this dataset are, in reality, complex social and cultural constructs. Race and socioeconomic status are intersectional. For example, a high median income could be due to the county having a higher educational score. However, it could also be due to that county having a higher percentage of caucasian residents as racial and economic disparities are often cofounded (Mode et al., 2016). Black Americans have been found to

have the highest incidence rates when it comes to lung cancer and less likely to receive timely and appropriate treatment when compared to White Americans after diagnosis (Schabath et al., 2017). With this background knowledge, we hypothesized that PctWhite, PctPublicCoverageAlone, PctEmployed16_Over, and PctUnemployed16_Over are among the significant regressors.

We included the other three variables in addition to race because an area with high PctPublicCoverageAlone likely means a lower income area, which likely means an area with poor access to care. PctEmployed16_Over and PctUnemployed16_Over were included because they also speak to the area's economic status.
More generally, we hypothesized that the full model of this data with all variables as regressors would be a decent model (with an R squared value above 0.6) because all variables can individually be argued to contribute to quality of care and therefore, contribute to explaining the death rate per incidence of lung cancer in a county.
Nonetheless, the goal of this project is not only to investigate our specific hypotheses but also to uncover an optimal set of predictors.

Assessing Multicollinearity

The data cleaning process was largely to prevent multicollinearity in the model. Before changing our target variable, we investigated a few models that predicted deathRate. However, from the analysis of those models, we found an especially large VIF value in incidenceRate. Thus, we moved the incidenceRate into the response variable to become DeathRate per incidenceRate (deathRatePerIncidence). In the variables we chose to keep, there may still be correlation (such as in empPrivateInsurance and publicInsuranceAlone). However, looking at the VIF of the remaining variables (Exhibit A), none of the variables exhibit values above 10 so we chose to keep them. Because of the thorough data clean, we did not seem to run into multicollinearity issues. We acknowledge, however, that there are 3 variables in the "warning zone" which would be a VIF greater than five. These three variables are povertyPercent, medIncome and PctBachDeg25_over. As we'll discuss later in this report, PovertyPercent and PctBachDeg25_over turn out to be the two most significant explanatory variables, so it would not be in our best interest to remove either one of them.

Model Selection

First, we looked at a full model with deathRatePerIncidence regressed on all variables from the cleaned dataset (Exhibit B). This resulted in an adjusted R-squared of 0.2716. We then tested our hypothesis and created a model with only PctWhite, PctPublicCoverageAlone,

PctEmployed16_Over, and PctUnemployed16_Over as predictors. The adjusted R-squared for this model is 0.1989. Neither of these are great R-squared values so we moved onto variable selection methods discussed in class.

We tried several methods including stepwise regression, stepwise with Akaike's Information Criteria (AIC), stepwise with Bayesian Information Criteria (BIC), and Least Absolute Shrinkage and Selection Operator (LASSO) with minimum λ and the λ that is 1 standard deviation away. From the LASSO plot (Exhibit C), we see that there is a big drop off in the number of variables in the 1 standard deviation λ, this suggests that the minimum λ may not have produced the most optimal model if those regressors can be driven to 0 that quickly. Because the diagnostic plots of all the models look to be relatively similar and the different models all had similar R-squared values in the 0.2 range, we looked for a parsimonious model with regressors that also showed up repeatedly in other models.

Ultimately, we chose stepwise regression with BIC as the best model (Exhibit D); the best linear predictors of deathPerIncidence are PctBachDeg25_Over, povertyPercent, PctRace, edScore, and BirthScore. We chose this model as it has a lower number of variables with a similar R-squared value to AIC models, satisfying parsimony, and the diagnostic plots do not indicate any major concerns. (See Model Diagnostics section for more detail.) The adjusted R-squared value for this model is 0.2692. All of the regressors in this model are significant and all these variables also appear in the best AIC model. PctBachDeg25_Over and povertyPercent also appear in the LASSO model with λ one standard deviation away. Therefore, the variables in this model are strong candidates to be a part of a best model.

Table 2. Variables selected in each model and each model's adjusted R-squared values.

Variable	LASSO (min λ)	LASSO (1se λ)	AIC	BIC
medIncome	—			
povertyPercent	**	***	***	***
studyperCap	—		—	
MedianAge	**		**	
AvgHouseholdSize	*		*	
PercentMarried				
edScore	***		***	***
PctHS25_Over	**		**	
PctBachDeg25_Over	***	***	***	***

PctEmployed16_Over	*	**	*	
PctUnemployed16_Over				
PctEmpPrivCoverage	*		**	
PctPublicCoverageAlone		—		
PctWhite				
BirthRate	**		**	
Adjusted R-Squared Value	0.2684	0.2578	0.2688	0.2608

not selected , — selected but not significant , Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Model Diagnostics

In Exhibit E, we show the diagnostic plots for our chosen best model, which is step regression with BIC; forward and backward selection results in the same model due to the fact that they both use the same selection criteria. Our first analysis is the residuals vs fitted values where we see a consistent dispersion scattered above and below the fitted line which is fairly even. We note that the scatter from the fitted line is substantial which makes sense given the low r-squared value.. The fitted line is straight and smooth and after reviewing the plot, we can conclude that we may still assume linearity in the model. If we look at the QQ-plot, we see some deviation in the tails with heavier tails on the right side but over the bulk of data, it is fitting pretty well. We also see that the histogram for the residuals (Exhibit F) follows a normal distribution with the heavy tails from the QQ-plot seemingly not as influential as we first thought, so the assumption of normality still holds. For the Scale-Location graph, we can see the heavy variance in the model by how far the scatter deviates from the fitted line. The residuals might be increasing with the fitted values slightly, but as a whole the red line is still fairly horizontal and smooth. Overall, we can state that our standardised residuals are fairly consistent and conclude that the assumption that our variance is constant (homoscedastic) still holds. Lastly, with the theoretical quantiles, we assess influential points, leverage points and outliers. We see that there are a few leverage points but with next to no influence on the fitted model. It is likely that this is the result of our thorough data cleaning. Therefore, when looking at the assumptions of a linear relationship, we can state that the residuals are correctly specified, homoscedasticity holds, the residuals are independent and uncorrelated, and the residuals are indeed normally distributed.

While we state that no assumptions have been violated, it is still likely that this data is slightly non-normal since our r-squared is low at 0.26; We also have to keep in mind that there is a slight increasing trend in the Scale-Location plot, and the QQ plot showed heavy tail distribution. We did a box cox transformation (Exhibit G) to investigate if it would have any impact. We calculated λ to be -0.303 with a short confidence interval that did not include 0 or -0.5, so we used our exact λ value to make the transformation. The r-squared value improved a bit to 0.28 but our data was fairly normal before so this result is not surprising. It is more likely that this dataset requires another method of modelling that is not linear regression.

Model Discussion

Our hypothesis of the full model being a decent model, while incorrect in the R-squared prediction, was not fully incorrect. In comparison to the other models that we explored, it still had a relatively high adjusted R-squared value and is actually higher than the adjusted R-squared value of the stepwise BIC model. However, as discussed in class, we want to consider parsimony when selecting the best model. Therefore, the slightly higher R-squared value is not enough for this model to be chosen. Also, because the R-squared values were all low, we were not too concerned with the slight differences among the models.

Our hypothesis model with PctWhite, PctPublicCoverageAlone, PctEmployed16_Over, and PctUnemployed16_Over as predictors was a weaker model in terms of the adjusted R-squared value. However, this was to be expected as this model was created with only subject matter expertise. It is interesting to note that PctWhite was not selected as a regressor in any of the models we created using variable selection methods learned in class. PctPublicCoverageAlone was selected in the LASSO model using 1se λ but it was non-significant. However, PctEmpPrivCoverage was selected and significant in the LASSO model with minimum λ and stepwise regression with AIC, so an insurance variable is an important variable. PctEmployed16_Over was not selected in any of the models but PctUnemployed16_Over was selected and significant in all but the stepwise regression with BIC model. If our models had higher R-squared values, we could make a statement about how this would suggest that race does not play a role in the death rate once a person is diagnosed and it would be an interesting discussion as we know this to be untrue in the real world. It is well documented that Black people have lower survival rates in various types of cancers, including lung cancer which is what this dataset is based off of. (Esnaola & Ford, 2012). However, because our models and selected best model does not do a great job of explaining the variance in the data, we expect it to be wrong in finding the important underlying variables.

Conclusion

This dataset was really interesting because it included so many variables that touch on societal demographics, economic conditions, level of education, and more. In something as complicated and individualistic as cancer, doctors often have to look beyond biological science when treating patients. Cancer and socioeconomic status has been a crucial public health topic in the past decade. As researchers develop more cutting edge treatments, those in lower socioeconomic classes, which is intersectional with education levels and race among other factors, are getting left behind in terms of treatment and therefore, have higher mortality rates. (Tabuchi, 2020). Our goal in this project was to apply this subject matter background to this data and contribute to this important topic in public health.

This investigation also informed us how important data collection and data aggregation methods are. Many variables in this dataset were aggregated in different ways, which contributed to our need to do such a thorough clean, and likely also affected our modelling results. If we were to do this again in the future and we were able to aggregate the data ourselves, we would definitely be mindful of keeping the units consistent as the current dataset had percentages, per capita, and raw numbers, which made it difficult to compare across different variables. As well, it was odd to us that sex only came up in the age category and nowhere else. Gender bias in medicine is also a big topic in public health so it would be interesting to include gender in more variables as well in a dataset like this. However, we do applaud the original author of the dataset; collecting health care data is challenging since there is not a universal health care system from which information can be gathered. Gathering the data through census likely impacted the quality of the data and therefore, the data analysis and model building.

While we weren't able to come up with a better model, we can conclude it is likely that linear regression is not the most optimal way to model this dataset. However, as with many scientific investigations, getting negative results does not mean failure. The data cleaning we did definitely resolved multicollinearity issues that existed in the original dataset. We contribute the knowledge that linear regression and the methods such as LASSO, stepwise regression with AIC and BIC, and Box-Cox transformations are not good approaches to modelling this data when death rate per incidence is the target.

Exhibits

Exhibit A: Variance inflation factor (VIF) Plot.

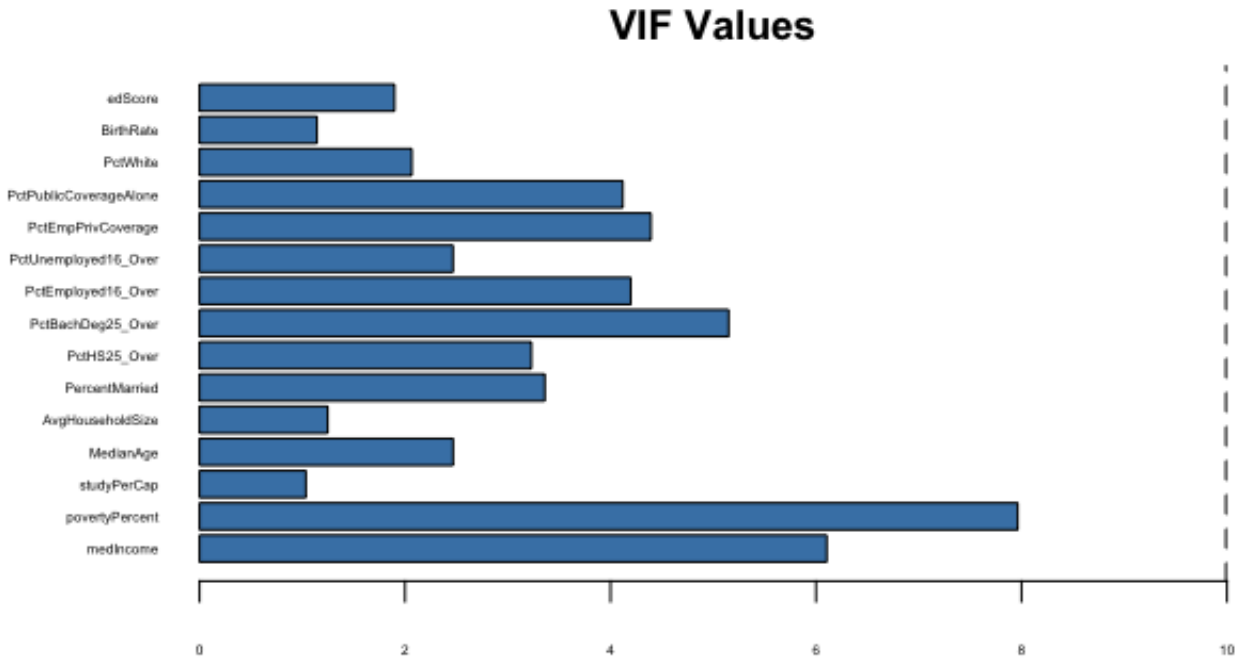

Exhibit B: Full model of cleaned data.

```
## Call:
## lm(formula = cancer$deathPerIncidence ~ ., data = cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22382 -0.03136 -0.00370  0.02483  0.62427
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.973e-01  3.702e-02  13.431  < 2e-16 ***
## medIncome              -2.355e-08  2.056e-07  -0.115  0.90879
## povertyPercent          1.422e-03  4.401e-04   3.232  0.00124 **
## studyPerCap            -3.232e-06  1.878e-06  -1.721  0.08544 .
## MedianAge              -8.810e-04  3.035e-04  -2.903  0.00373 **
## AvgHouseholdSize        5.890e-03  2.629e-03   2.241  0.02512 *
## PercentMarried          1.368e-04  2.668e-04   0.513  0.60810
## PctHS25_Over            7.461e-04  2.570e-04   2.903  0.00372 **
## PctBachDeg25_Over      -2.328e-03  4.243e-04  -5.487 4.45e-08 ***
## PctEmployed16_Over     -5.073e-04  2.470e-04  -2.054  0.04008 *
## PctUnemployed16_Over   -2.619e-05  4.567e-04  -0.057  0.95427
## PctEmpPrivCoverage     -5.078e-04  2.232e-04  -2.275  0.02296 *
## PctPublicCoverageAlone  2.060e-04  3.326e-04   0.619  0.53570
## PctWhite                2.118e-03  9.194e-03   0.230  0.81782
## BirthRate              -1.545e-03  5.365e-04  -2.879  0.00402 **
## edScore                -2.025e-02  6.300e-03  -3.215  0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05374 on 2852 degrees of freedom
## Multiple R-squared:  0.2716, Adjusted R-squared:  0.2677
## F-statistic: 70.88 on 15 and 2852 DF,  p-value: < 2.2e-16
```

Exhibit C: LASSO Plot

This plot displays minimum λ and one standard deviation away.

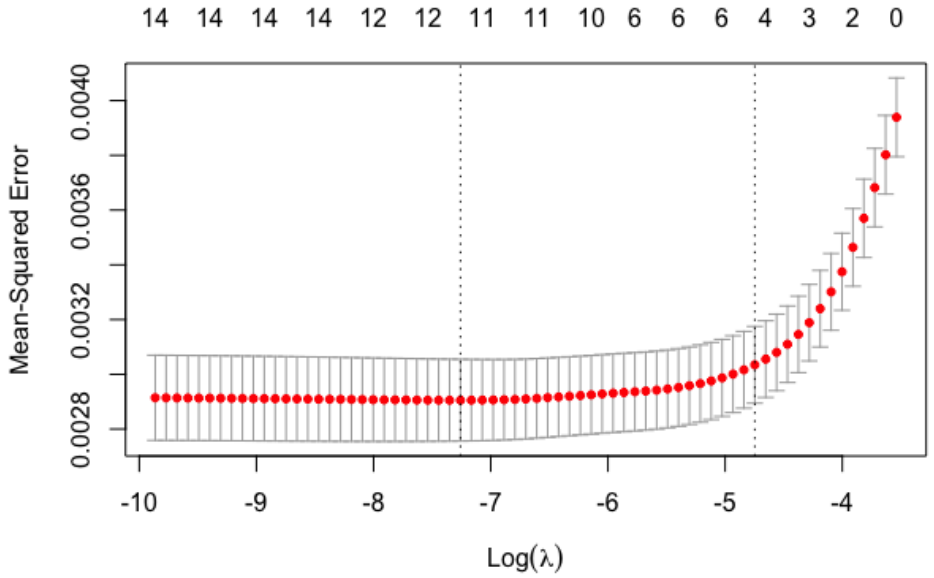

Exhibit D: Variable selection plot for BIC

```
Call:
lm(formula = cancer$deathPerIncidence ~ PctBachDeg25_Over + povertyPercent +
    edScore, data = cancer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21740 -0.03118 -0.00371  0.02618  0.61429

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.4676726  0.0128394  36.425  < 2e-16 ***
PctBachDeg25_Over -0.0033095  0.0002588 -12.789  < 2e-16 ***
povertyPercent     0.0023215  0.0001852  12.536  < 2e-16 ***
edScore           -0.0262253  0.0056775  -4.619 4.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05399 on 2864 degrees of freedom
Multiple R-squared:  0.2616,    Adjusted R-squared:  0.2608
F-statistic: 338.1 on 3 and 2864 DF,  p-value: < 2.2e-16
```

Exhibit E: Diagnostic plots for stepwise with BIC model

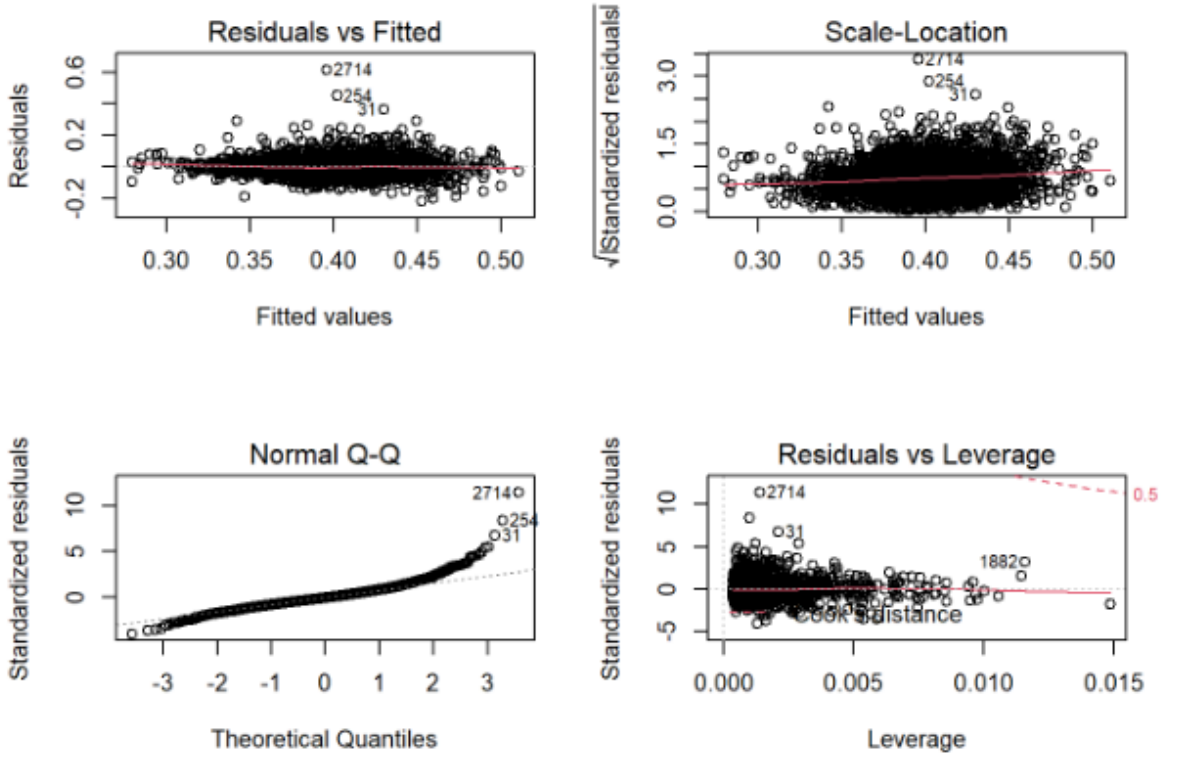

Exhibit F: Histogram of residuals for BIC model

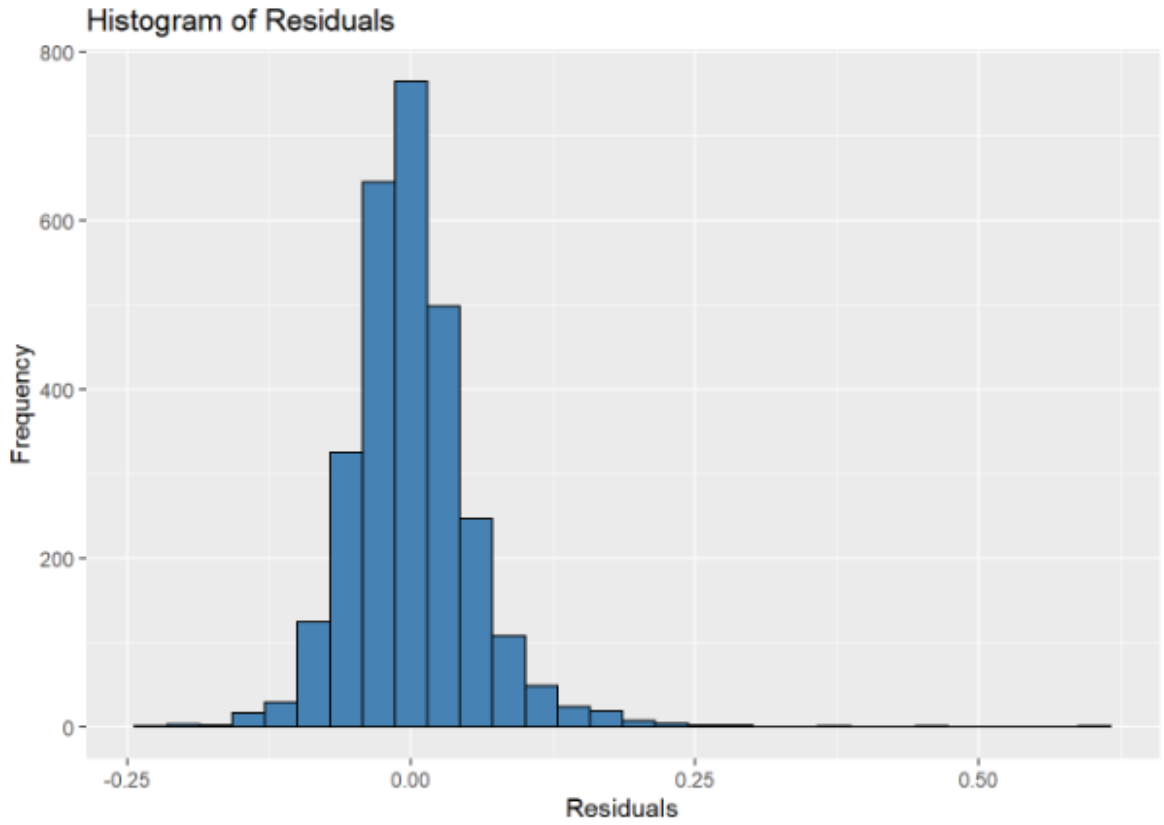

Exhibit G: Box Cox Plot