Using Poverty to Predict Cancer Mortality

Introduction

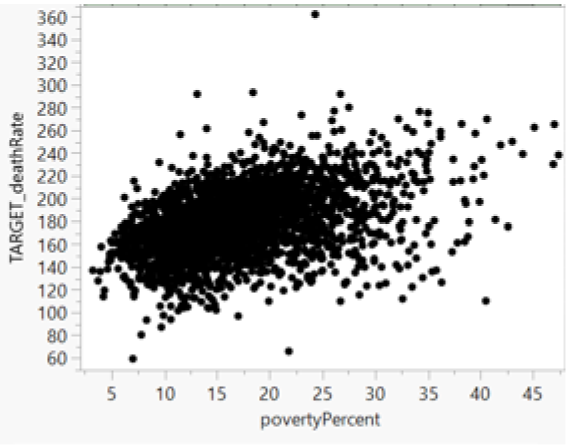
Health care is an ongoing topic in America with a major emphasis on affordability. Many people cannot afford health insurance, nor have a job that provides health insurance. Hospital visits can put those people in a lifetime of debt, or they may choose not to check in on their health at all. I have personal experience with this struggle as a close family member once had colon cancer. While I am not living in poverty, I wanted to investigate the relationship between poverty and deaths caused by cancer. I believe that poverty rate is a significant predictor of cancer mortality rates.

The data I used to examine this relationship is a combination of information from cancer.gov, clinicaltrials.gov, and the American Community Survey gathered by Noah Rippner (rippner, 2017). It contains 3047 observations from counties across America from years 2010 through 2016. The variables I used in my analysis were: *TARGET_deathRate*, which is the per capita (100,000) cancer mortalities from 2010 through 2016, and *povertyPercent*, which is the percent of populace in poverty based on the 2013 Census estimates. *Table 1* includes the means and variances of the mortalities and poverty rates, and *Graph 1* shows the plot of cancer mortalities and percentage of poverty. The mean per capita cancer mortalities seem to be fairly normally distributed while the population poverty percentage is more skewed to the right (*Appendix 1*). These two variables have a positive relationship [covariance = 76.37, correlation coefficient = 0.43].

Table 1. Marginal Distributions

Variable	**Mean**	**Variance**
DeathRate	178.66	770.17
PovertyPercent	16.88	41.08

Graph 1. Cancer Mortalities vs. Poverty Percentage

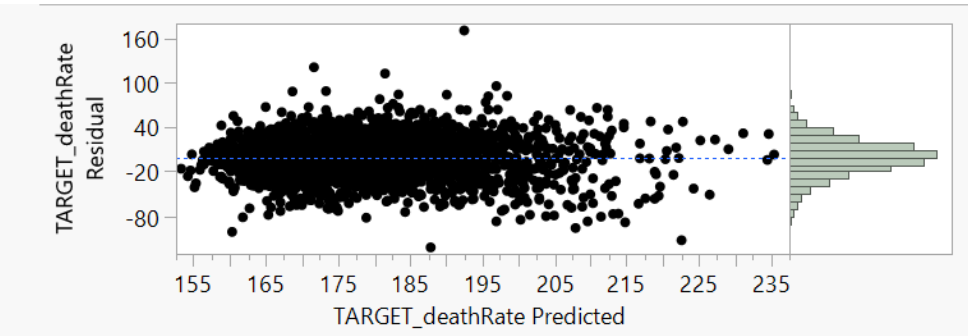


Regression Analysis

The results of a simple linear regression analysis indicated that the percentage of a populace in poverty (in a given US county) is a significant predictor of mean per capita cancer mortality [F(1,3045) = 688.33, p < 0.0001, R^2 = 0.18] (*Appendix 2&3)*. For each additional one percent increase in poverty, the mortality increased by 1.86 [95% CI: (1.72, 2.00)] (*Appendix 4*). The regression equation used to predict mortality is *(mean per capita cancer mortality) = 147.28 + 1.86 * (percent of population in poverty).*

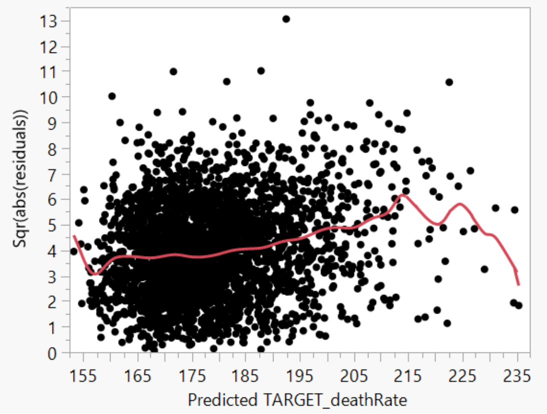
A closer look at the simple linear regression assumptions strengthens the validity of these results. In respect to linearity, using a graph of the death rate residuals vs. the death rates predicted by my model (*Graph 2*) we can see that observations are evenly spread above and below the residual = 0 line, so we can assume this assumption is met.

Graph 2. Residuals vs. Predicted Cancer Mortalities



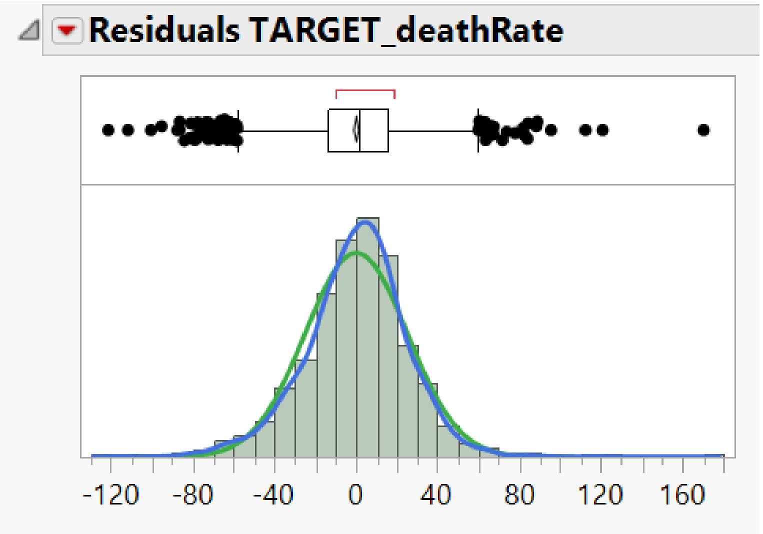
The constant variance assumption is less convincing as a look at a *Graph 3*; the absolute value of residuals squared and the predicted cancer mortalities does not show a consistently flat line. What we see is that as the predicted rates increase so does the variability. This is easier seen on the residuals and predicted cancer mortalities plot (*Graph 2*). While we simply have less data for those predictions, this may still be a little concerning.

Graph 3. Absolute Value of Residuals Squared vs. Predicted Cancer Mortalities

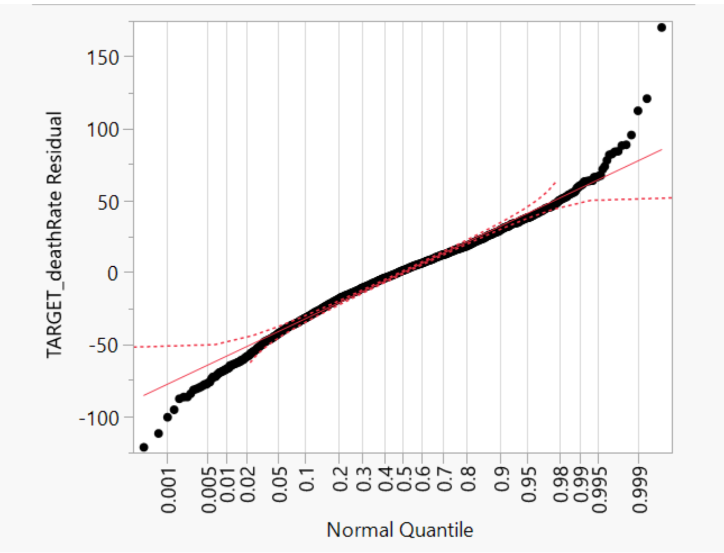


For the assumption of normality, I think that the residuals are fairly normally distributed; they follow, pretty well, a bell-shaped curve (*Graph 4*). It may be skewed slightly to the right which can be seen a little more clearly by the tails in the normal quantile plot (*Graph 5*), but mostly I think this assumption is met.

Graph 4. Residual Distribution with Normality Curve



Graph 5. Normal Quantile Plot for Residuals



Lastly, these data were not collected in any sort of order, so we can assume that the independence assumption is met.

When looking for outliers with respect to poverty percentage, four data points especially caught my attention due to their leverage values being larger than most of the other observations (*Appendix 5*). Using Bonferroni's correction for analysis of outliers with respect to cancer mortality resulted in five observations being deemed outliers (*Appendix 6*). I investigated these outliers further to determine whether they are influential or not. I computed the Cook's distance for each observation. I decided to deem observations influential if they exceed the 50th percentile of the F-distribution [F(2,3045) = 0.69]. Using this, no observation qualifies; therefore, none are influential.

Discussion

According to my model, I am 95% confident that for each additional one percent increase of the population in poverty, the mean cancer mortality increases by between 1.72 and 2.00. For the conditional mean, I am 95% confident that the mean cancer mortality when the percentage of the population in poverty is 20 is somewhere between 183.48 and 185.46. And to predict for a specific individual county with a poverty percentage of 20, I am 95% confident that the mean cancer mortality is between 135.31 and 233.63. (*Appendix 7*)

Conclusion

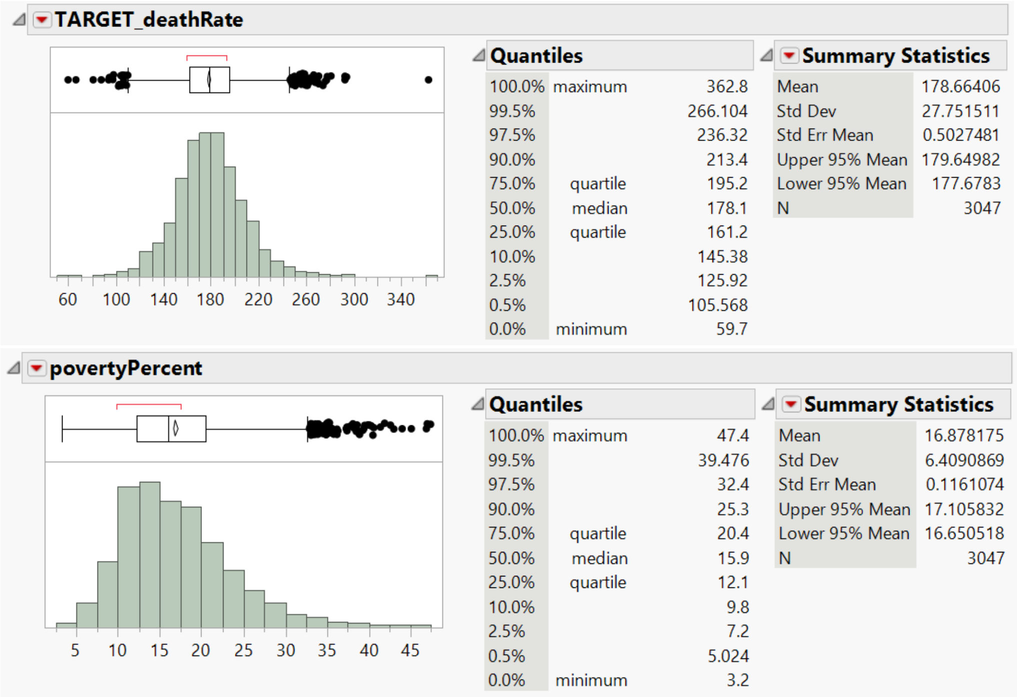
The results of my analysis are statistically significant and state that poverty percentage explains for about 18% of the variability we see in the mean cancer mortalities in counties across America. A limitation of this analysis is that it is only generalizable to the United States, as it only included data from American counties and there may be any number of different factors elsewhere. For example, being poorer in some other countries may not dissuade those individuals from seeking medical advice on account of possibly more forgiving health care payment systems. Another note is that this study does not imply causation as it was an observational study. Researchers cannot assign individuals to poverty nor choose who has cancer, and therefore these findings are only correlational. I would like to research further into this to see what more goes into this result. Are sick people more likely to be impoverished? Are there significant lifestyle differences between those in poverty and those who are not that would lead them to be more likely to have or die from cancer? Are impoverished people more likely to not seek medical help for sickness whether because of lack of resources, funds, etc.?

References:

Rippner, Noah. “OLS Regression Challenge - Dataset by Nrippner.” *Data.world*, 4 Jan. 2017, https://data.world/nrippner/ols-regression-challenge.

Appendix:

Appendix 1. Marginal Distributions



TARGET_deathRate

Quantiles

100.0%	maximum	362.8
99.5%		266.104
97.5%		236.32
90.0%		213.4
75.0%	quartile	195.2
50.0%	median	178.1
25.0%	quartile	161.2
10.0%		145.38
2.5%		125.92
0.5%		105.568
0.0%	minimum	59.7

Summary Statistics

Mean	178.66406
Std Dev	27.751511
Std Err Mean	0.5027481
Upper 95% Mean	179.64982
Lower 95% Mean	177.6783
N	3047

povertyPercent

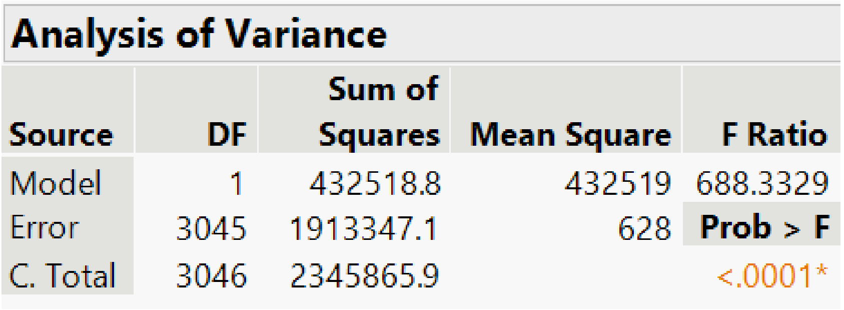
Quantiles

100.0%	maximum	47.4
99.5%		39.476
97.5%		32.4
90.0%		25.3
75.0%	quartile	20.4
50.0%	median	15.9
25.0%	quartile	12.1
10.0%		9.8
2.5%		7.2
0.5%		5.024
0.0%	minimum	3.2

Summary Statistics

Mean	16.878175
Std Dev	6.4090869
Std Err Mean	0.1161074
Upper 95% Mean	17.105832
Lower 95% Mean	16.650518
N	3047

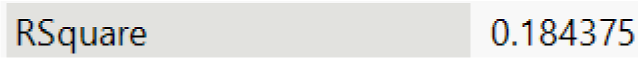
Appendix 2. ANOVA Table



Analysis of Variance

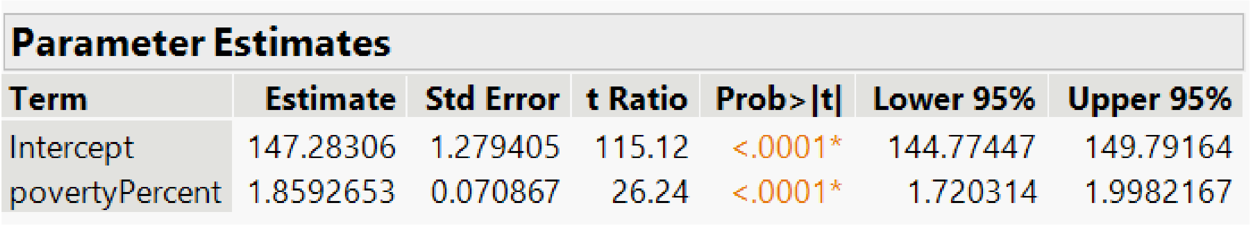
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	432518.8	432519	688.3329
Error	3045	1913347.1	628	Prob > F
C. Total	3046	2345865.9		<.0001*

Appendix 3. R^2



RSquare	0.184375

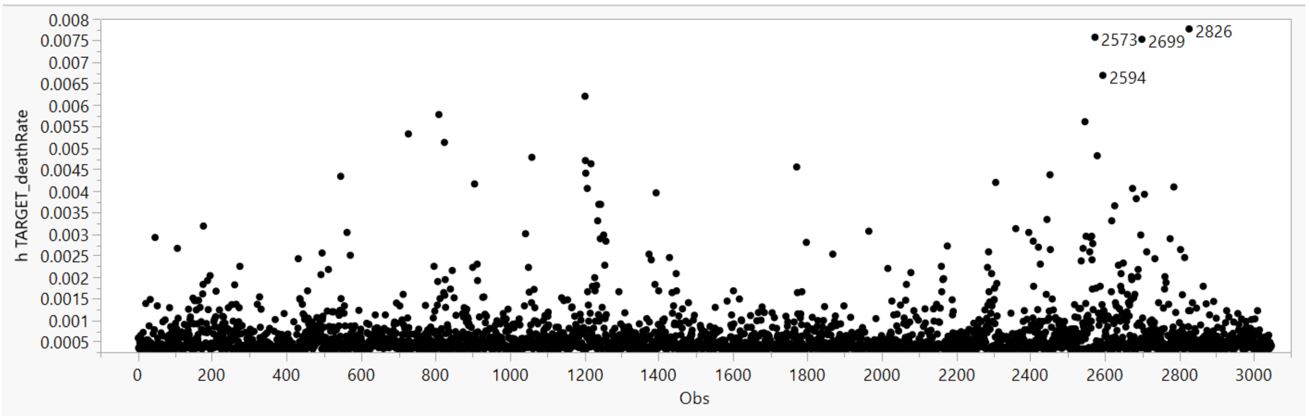
Appendix 4. Parameter Estimates



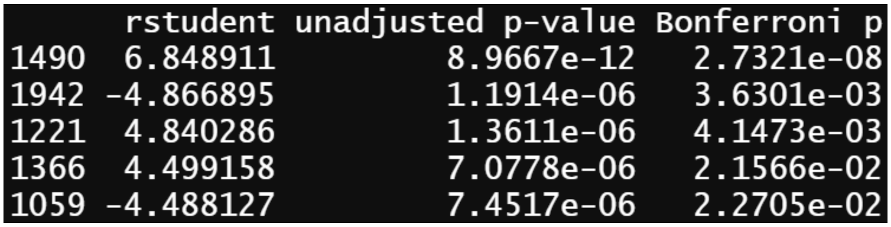
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	147.28306	1.279405	115.12	<.0001*	144.77447	149.79164
povertyPercent	1.8592653	0.070867	26.24	<.0001*	1.720314	1.9982167

Appendix 5. Leverage Values of Poverty Percentages

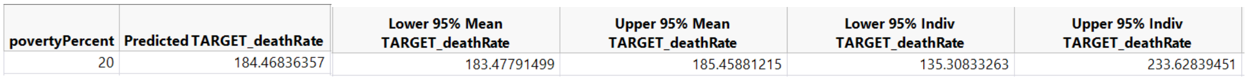


Appendix 6. Bonferroni's Correction for Cancer Mortality Outliers



```
      rstudent unadjusted p-value Bonferroni p
1490  6.848911          8.9667e-12   2.7321e-08
1942 -4.866895          1.1914e-06   3.6301e-03
1221  4.840286          1.3611e-06   4.1473e-03
1366  4.499158          7.0778e-06   2.1566e-02
1059 -4.488127          7.4517e-06   2.2705e-02
```

Appendix 7. 95% Confidence Intervals for Mean Response and Individual Value



povertyPercent	Predicted TARGET_deathRate	Lower 95% Mean TARGET_deathRate	Upper 95% Mean TARGET_deathRate	Lower 95% Indiv TARGET_deathRate	Upper 95% Indiv TARGET_deathRate
20	184.46836357	183.47791499	185.45881215	135.30833263	233.62839451