Which factors affect the average life expectancy the most in a state

STATS 512

Spring 2018

Group 13

Justin Boudreau

Benya Chongolnee

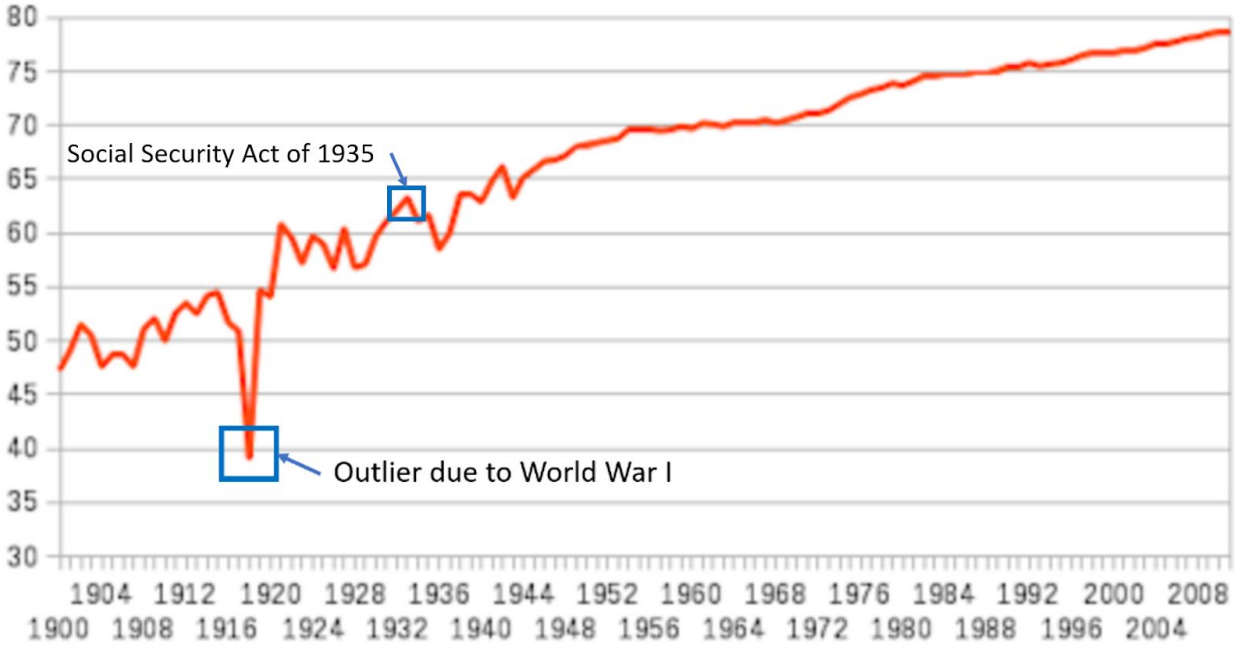
Taylor Duncan

Nidhi Sakhala

Xiaoyu Yu

Introduction

The average life expectancy in the United States in 1900 was approximately 45 years of age. Major policies are determined on the basis of the average life expectancy such as the age when a person may receive Social Security benefits. As demonstrated in the picture below, since 1900 the average life expectancy for the common person rose to approximately 78-79 years old [2]. This has placed enormous stress on the Social Security budget as more people are gaining longer access to benefits that one would only receive after they lived beyond the average life expectancy at the time Social Security was established, which was 65 years old. In addition to major federal policies, there is a continued growing interest in how a typical person can increase the number of years of life through the advancements and increased availability of modern medicine.



Life Expectancy in the US (1900-2011)

Because the average life expectancy (ALE) of a population heavily influences that population's policies, agendas, and economics, to help improve the amount of years a person can expect to live as a factor of life choices, it is worthwhile identifying common factors that have a positive effect on ALE as well as identifying common factors that have a negative effect. Common factors that arguably increase a person's life expectancy represents a multi-billion dollar industry with revenue coming from gym memberships, home fitness equipment, and helpful dieting tools. However, common factors that arguably decrease a person's life expectancy also represent a multi-billion dollar industry with fast food, alcohol, and cigarettes. Further compounding the issue are various information campaigns to help inform the population for better life choices with regards to working out and eating healthy. Other factors that have a net effect on a person's life expectancy are non-physical as the aforementioned factors are, such as maintaining gainful employment at a large company to retain health insurance for annual preventative treatments.

It is important to understand what factors can increase or decrease life expectancy in order to live longer. For this report, we will be using statistics that are publicly available from Centers for Disease Control and Prevention [1]. We will be analyzing the following variables: colon cancer, heart disease, lung cancer, motor vehicle injuries, stroke, injury, uninsured, disabled medicare, major depression, unemployment, population size, primary care physician rate, and average life expectancy. It seems reasonable that these effects are more than likely to either contribute to or detract from the average life expectancy. However, the strength of the correlation is desired to be observed. We would want perform multiple linear regression to find the parameter estimates of each variables, coefficient of determination, standard errors and more.

The purpose of this study is to target key factors that have a relationship with average life expectancy to inform the general population, and to inform decision makers specifically at the federal and state levels using 50 counties in Missouri (MO) and validate the model using 65 other counties in Missouri.

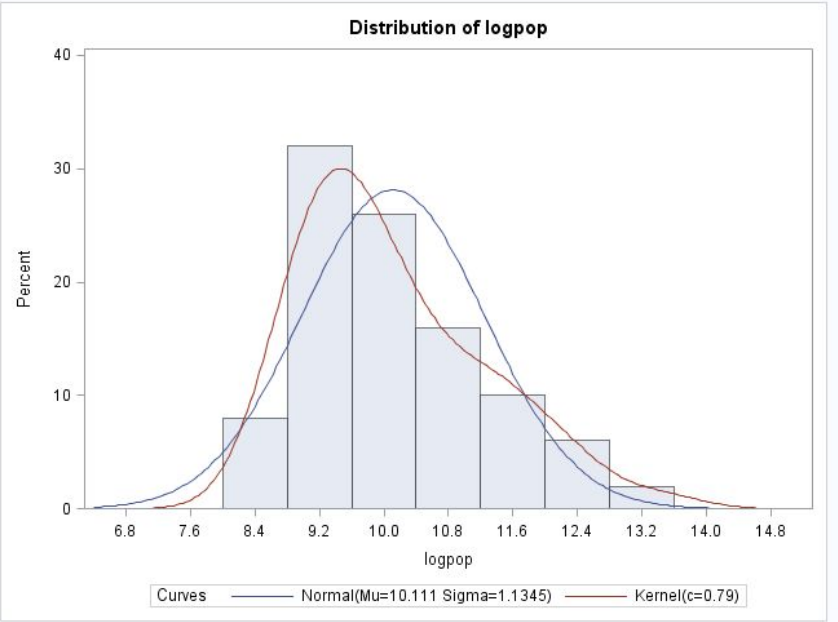
Methods

Data Details:

As stated in the introduction, the data used for this project is publicly available from Centers for Disease Control and Prevention [1]. This dataset contains health indicators for communities all over America such as obesity, heart disease, cancer, average life expectancy. We will be reducing this big dataset to only use the desired attributes and variables and will only want to focus on the state of Missouri because the state has relatively normal factors that distinguished itself from other states. Other states such as Colorado may have outliers since the state is at a high altitude or Arizona may also have outliers since it is mostly hotter than other states. The sample size that we are using is 50 randomly selected counties in Missouri. The variables that we will be testing are colon cancer (Col_Cancer), heart disease (CHD), lung cancer (Lung_Cancer), motor vehicle injuries (MVA), stroke (Stroke), injury (Injury), uninsured (Uninsured_percent), disabled medicare (Disabled_Medicare_percent), major depression (Major_Depression_percent), unemployment (Unemployed_percent), population size (Population_size), and primary care physician rate (Prim_Care_Phys_Rate). These variables will produce an effective model that can predict the average life expectancy (ALE) with a good degree of precision.

Preliminary Exploratory Analyses:

Originally, the data set that we acquired had multiple different scales or units so our first step was transforming it all into the same unit. We realized some of the variables were in different scales and units. By plotting the data, we found outliers from Jackson county due to its population size. We then studied the histograms for each variable and the bivariate scatterplots for each pair of variables in the scatterplot matrix observed in Fig 21. We transformed the data so each variable was set as a ratio out of the population of the county. After this transformation, we examined the distribution of each variables to check for any more outliers or skewness in the data. The histogram distributions (Fig. 1 - Fig. 13) showed that uninsured, primary care physician rate, injury, stroke, motor vehicle accidents, lung cancer, heart disease, colon cancer, average life expectancy were normally distributed while depression, unemployment, and disabled medicare were slightly left skewed and population size was very left skewed. From this we chose to use the log transformation of population size in order to get a much more normal distribution to use in our model. Though performing the log transformation reduce the correlation by a little, it drastically changed the distribution of population size as seen below compare to Fig. 5.



Next, we examined the correlation matrix for all of the variables before the log transformation. From the matrix in Fig. 18, we found only two variables showed a high level of correlation of greater than 0.7: disabled medicare and uninsured. This shows that there could potentially be an interaction term between the two variables. From this conclusion, we decided to test how the results will change after adding the interaction term.

Model Building Process:

Before doing any changes to the dataset, we tried model selection using the best subsets method to identify our potential best model. We decided to choose the best model based on lowest Cp and got the best model to have 5 parameters - Lung Cancer, Prim_Care_Phys_Rate, Population_Size, Major_Depression_percent and Unemployed_percent. As seen in Fig. 20, the R^2 for this model is 0.6531 but the individual T-tests of these parameters are worrying. There are 2 parameters - Prim_Care_Phys_Rate and Population_Size which are insignificant to the model.

After log-transforming the Population_Size because of its non-normality, we iterated the model selection process and we saw that Population was not a part of our best model anymore.

Next, we added the interaction term to the dataset and proceeded with best subsets selection again. We chose best 5 models from among all the models generated, as explained below.

Inferential Methods:

The table below shows the top five models chosen based on lowest SBC. Among the five models, the AIC and SBC values are pretty close to each other. The Cp values of all the 5 models are less than the number of parameters; so all of them are valid. We preferred a model with a lower number of parameters. So considering all these criterias, the two models we selected are Model 1 and Model 2 as shown in the table below.

Model number	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Variable names	Lung Cancer Unemployed Major Depression Interaction term	Lung Cancer Major Depression Unemployed Uninsured	Lung Cancer Uninsured Unemployed Disabled Medicare Major Depression	Lung Cancer Disabled Medicare Unemployed Major Depression	Lung Cancer Major Depression Unemployed Uninsured Interaction term
Number of variables	4	4	5	4	5
SBC	-36.01092	-34.65267	-33.29742	-33.29742	-33.20765
R^2	0.7572	0.7505	0.769	0.7437	0.7625
R^2_{adj}	0.7356	0.7284	0.7427	0.7209	0.7356
Cp	0.2342	1.3421	0.2887	2.478	1.3518
AIC	-45.571	-44.2128	-46.0492	-42.8575	-44.6798

After this step, we then used the rest of the Missouri counties that were not used in the model building process to test the Model 1 and Model 2 as will be further discussed in the result section.

Results

After looking at the analysis of the top 5 best models, we concluded that two models stand out the most. We found out that the two models found had the best SBC and AIC. The two models that stand out were Model 1 and Model 2. Model 1 included lung cancer, depression, unemployed, and medicare-uninsured interaction. Model 2 included lung cancer, uninsured, unemployed, and depression. Below is the breakdown of the Model 1 and Model 2.

	Model 1	**Model 2**
R^2	0.7572	0.7505
R^2_{adj}	0.7356	0.7284
Cp	0.2342	1.3421
AIC	-45.571	-44.2128
SBC	-36.01092	-34.65267

	Intercept	**Lung Cancer**	**Unemployed**	**Depression**	**Uninsured**	**Medicare* Uninsured**
Model 1	71.76666	-0.03617	-1.25721	1.61220	-	-0.01984
Model 2	73.74058	-0.04345	-1.33887	1.53035	-0.12171	-

Model 1:

$$\widehat{ALE} = 71.76 - 0.03617(Lung\ Cancer) - 1.25721(Unemployed) + 1.61220(Depression) - 0.01984(Medicare)(Uninsured)$$

Model 2:

$$\widehat{ALE} = 73.74 - 0.04345(Lung\ Cancer) - 0.12171(Uninsured) - 1.33887(Unemployed) + 1.53035(Depression)$$

The breakdown shows all analyses for Model 1 and Model 2 are similar, except that Cp for model 1 is relatively low to model 2. On top of this, AIC and SBC for model 1 is also lower than model 2 AIC and SBC. To confirm that model 1 is a better model, we want to test the two

models with 65 other counties in Missouri. We do this by checking its sum of squares error and mean square error given in the chart below. We found that the model with the interaction term had a lower sum of squares error (68.48499) compared to the sum of squares error for the model without the interaction term (75.5210) in addition to the mean square error of 1.0536 and 1.1618 for the models with and without the interaction term respectively. This supports selecting the model 1 in addition to the Cp, AIC, and SBC values that were found when comparing all different types of models.

	Model 1	**Model 2**
SSE	68.48499024	75.52099643
MSE	1.053615234	1.161861484

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	51.32165	12.83041	35.09	<.0001
Error	45	16.45455	0.36566		
Corrected Total	49	67.77620			

Root MSE	0.60470	R-Square	0.7572
Dependent Mean	76.12600	Adj R-Sq	0.7356
Coeff Var	0.79434		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.76565	3.67849	19.51	<.0001
Lung_Cancer	1	-0.03617	0.00919	-3.94	0.0003
Unemployed_percent	1	-1.25721	0.30429	-4.13	0.0002
Major_Depression_percent	1	1.61220	0.53121	3.03	0.0040
int_term	1	-0.01984	0.00419	-4.73	<.0001

The above analysis of variance and parameter estimates show the breakdown of our chosen best model. It can be seen by the t-tests that that all variables are significant with R^2 of 0.7572. Significant t-tests means that each variables are statistically significant in predicting the average life expectancy. There is also no multicollinearity that can be seen from the table. From the ANOVA table above, the F-test shows that the model itself is statically significant with a p-value of <.0001. This means that the model accurately represents the parameter interaction with average life expectancy. Fig. 14 and Fig. 15 shows that all the assumptions for multiple linear regression are met. Fig. 15 shows independent residuals, non-constant variance, and approximately normal residuals.

The parameter estimates seen in Fig. 16 also shows the 95% confidence limits of each variables. This reveals that lung cancer, unemployment, and the interaction term all reduce the average life expectancy while depression increases the average life expectancy. The parameter estimates shows an increase in one unit of the population affected by lung cancer means that the average life expectancy decrease by 0.03617 years. An increase in one percent of the population who are unemployed means that the average life expectancy decrease by 1.25721 years. An increase in one unit of the population affected by depression means that the average life expectancy increase by 1.6122 years. Lastly, an increase in one unit of the population who are uninsured and the population that have disabled medicare means that the average life expectancy decrease by 0.01984 years.

Discussion

Surprisingly, only one major factor initially considered was part of the final model with the other terms being related to non-physical characteristics of a person's lifestyle. It is also noteworthy to point out that the depression variable had a positive correlation, whereas the other variables had a negative correlation, which might imply that people that identify with depression find treatment continue on to live longer lives. It is also possible that some that were not identified with depression in the population lived a short life expectancy, possibly significantly shorter to bring down the ALE of the non-Depressed population at large.

Among the deterministic variables for ALE, unemployment accounted for the most significant reduction in years. From this data, we can conclude that in 2010 being unemployed (or having a significant amount of time unemployed) could have a severe negative impact on one's life expectancy. This can be caused by many things. Being unemployed could suggest that the person do not have money to find shelter or meals. Because of this, they could become sick. With no money, they are unable to get the help they need; therefore, they could die from this.

The original data includes colon cancer, heart disease, lung cancer, motor vehicle injuries, stroke, injury, uninsured, disabled medicare, major depression, unemployment, population size, and primary care physician rate. It is surprising how the best model for average life expectancy only includes lung cancer, unemployment, depression, uninsured and disabled medicare. The other variables such as stroke, motor vehicle injuries and heart disease were not included in the model. This suggests that the variables are not statistically significant; therefore, it does not help predict the average life expectancy. This could be caused by many different reasonings such that there are new medical technology that help injuries and stroke so that there are lower chance of people dying from it.

References

[1] Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer. (2012, May 30). Retrieved April 17, 2018, from https://www.healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer

[2] Disabled World. (2017, November 21). U.S. Life Expectancy Statistics Chart by States. Retrieved April 17, 2018, from https://www.disabled-world.com/calculators-charts/states.php

[3] Dr. Lichti's Statistics 512 Notes. (2018). Purdue Department of Statistics, 250 N. University St, West Lafayette, IN 47907.

[4] Kutner, Michael H., Nachtsheim, Christopher J., Neter, John, and Li, William. (2013). *Applied Linear Statistical Model Fifth Edition.* McGraw-Hill Education P.O. Box 182605, Columbus, OH 43218.

Appendix A: Code

```
*Read the data-
data state;
infile 'W:\stat\final_dataset.csv' delimiter='2c'x firstobs=2;
input County_Code County_Name$ Col_Cancer CHD Lung_Cancer MVA Stroke Injury Prim_Care_Phys_Rate
Population_Size Uninsured_percent Disabled_Medicare_percent Unemployed_percent Major_Depression_percent
ALE;
run;

*Preliminary analysis-
*For histograms of individual predictors-
  proc univariate data=state noprint;
  histogram ALE/ normal kernel (L=2);
  histogram Col_Cancer/ normal kernel (L=2);
  histogram CHD/ normal kernel (L=2);
  histogram Lung_Cancer/ normal kernel (L=2);
  histogram MVA/ normal kernel (L=2);
  histogram Stroke/ normal kernel (L=2);
  histogram Injury/ normal kernel (L=2);
  histogram Prim_Care_Phys_Rate/ normal kernel (L=2);
  histogram Population_Size/ normal kernel (L=2);
  histogram Uninsured_percent/ normal kernel (L=2);
  histogram Disabled_Medicare_percent/ normal kernel (L=2);
  histogram Unemployed_percent/ normal kernel (L=2);
  histogram Major_Depression_percent/ normal kernel (L=2);
  run;

*Correlation matrix-
proc corr data=state noprob;
  run;

*Scatterplot matrix
proc sgscatter data=state;
  matrix ALE Col_Cancer CHD Lung_Cancer MVA Stroke Injury Prim_Care_Phys_Rate Population_Size
Uninsured_percent Disabled_Medicare_percent Unemployed_percent
Major_Depression_percent/diagonal=(histogram);
  run;

*Best subsets selection for best model -
proc reg data=state;
model ALE= Col_Cancer CHD Lung_Cancer MVA Stroke Injury Prim_Care_Phys_Rate Uninsured_percent
Disabled_Medicare_percent Unemployed_percent Major_Depression_percent Population_Size / selection= rsquare
cp adjrsq aic sbc press b best=8;
run;
```

```
*Analysis of best model-
proc reg data=inter;
model ALE= Lung_Cancer Unemployed_percent Major_Depression_percent Prim_Care_Phys_Rate
Population_Size;
run;


*Tto log transform population
data transformed;
set state;
logpop = log(Population_Size);
proc print data=transformed;
run;

*Histogram of population after log transformation
proc univariate data=transformed noprint;
histogram logpop/ normal kernel (L=2);
run;


*Correlation among variables -
  proc corr data=transformed noprob;
  run;


*Best subsets selection for best model -
proc reg data=transformed;
model ALE= Col_Cancer CHD Lung_Cancer MVA Stroke Injury Prim_Care_Phys_Rate Uninsured_percent
Disabled_Medicare_percent Unemployed_percent Major_Depression_percent logpop / selection= rsquare cp adjrsq
aic sbc press b best=8;
run;

*Fitting the best model obtained from model selection
proc reg data=transformed;
model ALE= Lung_Cancer Uninsured_percent Disabled_Medicare_percent Unemployed_percent
Major_Depression_percent/ cli p r;
run;

*Adding the interaction term -
data inter;
set transformed;
int_term = Uninsured_percent*Disabled_Medicare_percent;
proc print data=inter;
Run;

*Best subsets selection for best model -
proc reg data=inter;
```

```
model ALE= Col_Cancer CHD Lung_Cancer MVA Stroke Injury Prim_Care_Phys_Rate Uninsured_percent
Disabled_Medicare_percent Unemployed_percent Major_Depression_percent logpop int_term/ selection= rsquare
cp adjrsq aic sbc press b best=8;
run;

*Confidence limits for final model -
proc reg data=inter;
model ALE= Lung_Cancer Unemployed_percent Major_Depression_percent int_term/ clb clm p r;
run;

*Sums of squares for final model -
proc reg data=inter;
model ALE= Lung_Cancer Unemployed_percent Major_Depression_percent int_term/ ss1 ss2;
run;
```

Appendix B: SAS Output

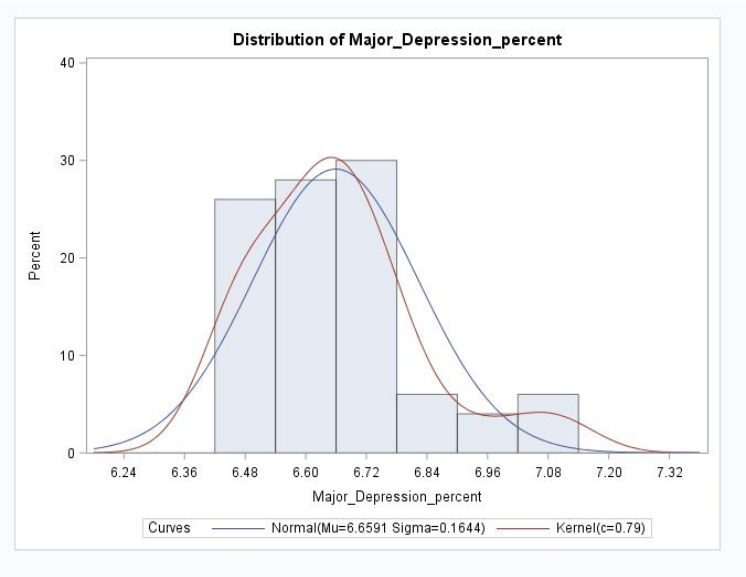


Fig. 1: Distribution of depression before log transformation

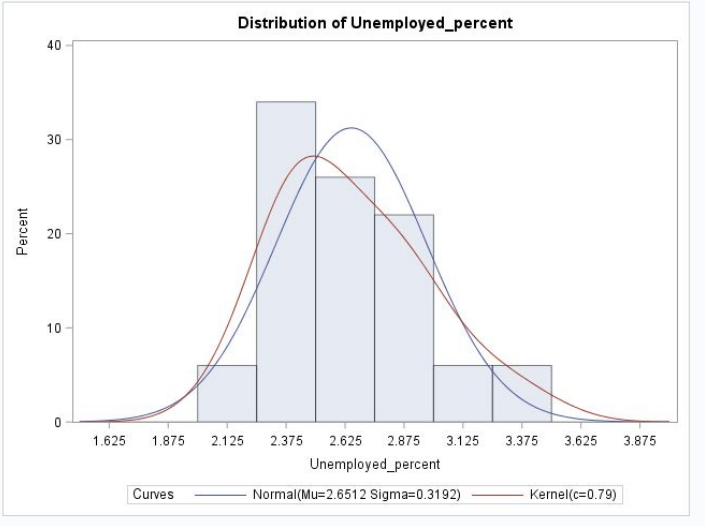


Fig. 2: Distribution of Unemployed before log transformation

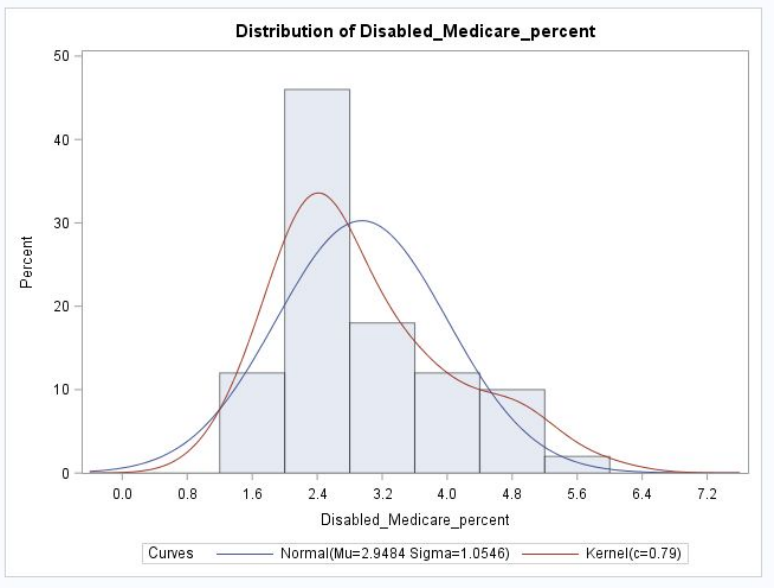


Fig. 3: Distribution of Disabled_Medicare before log transformation

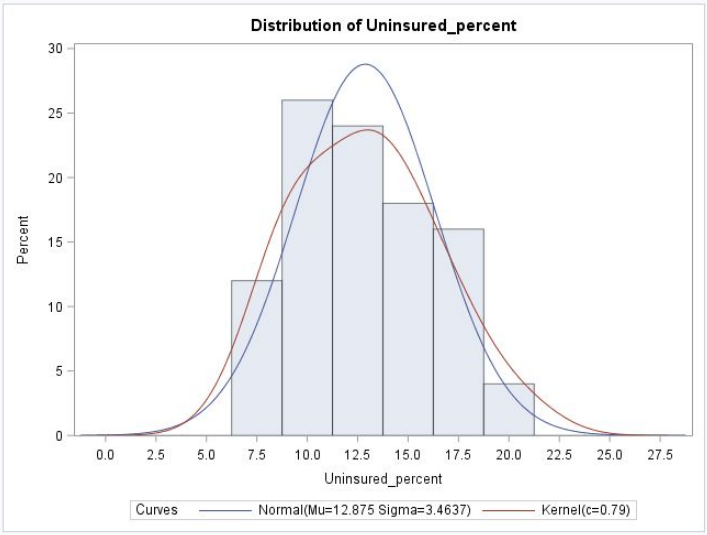


Fig. 4: Distribution of Uninsured before log transformation

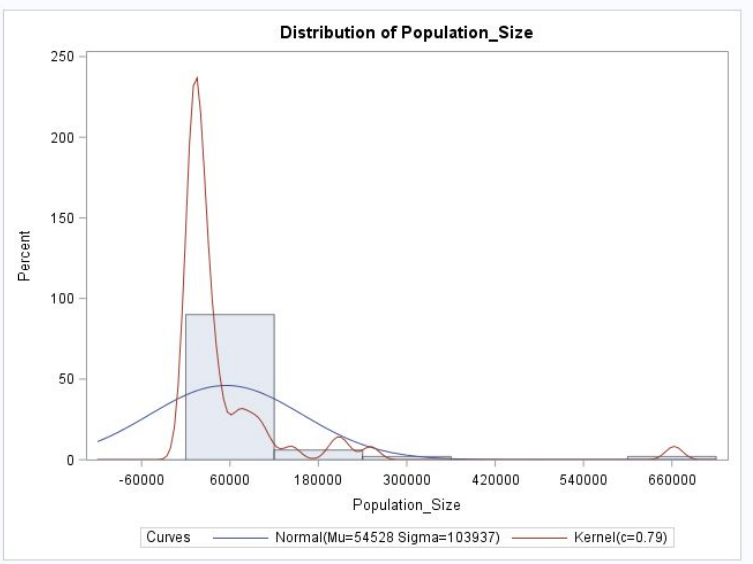


Fig. 5: Distribution of the population size before log transformation

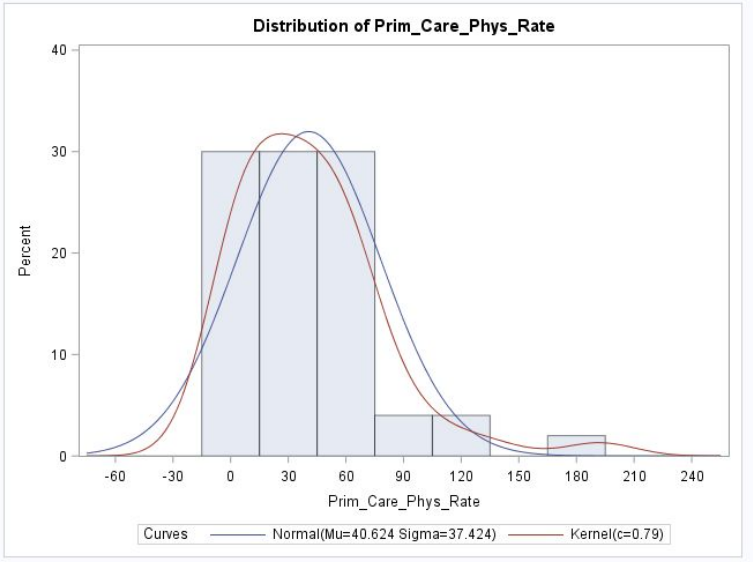


Fig. 6: Distribution of primary care physician rate before log transformation

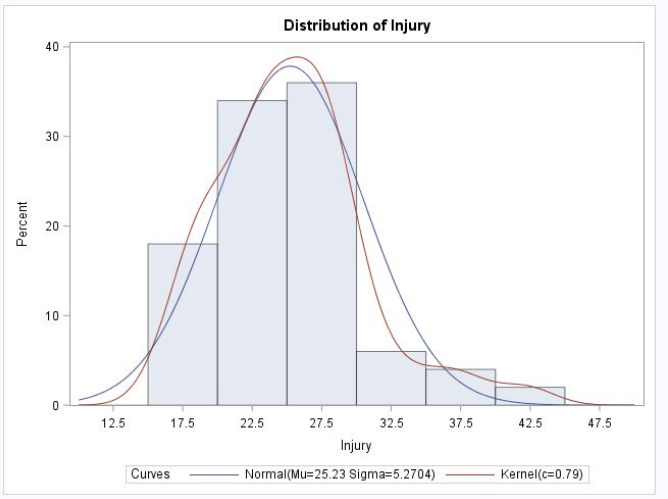


Fig. 7: Distribution of injury before log transformation

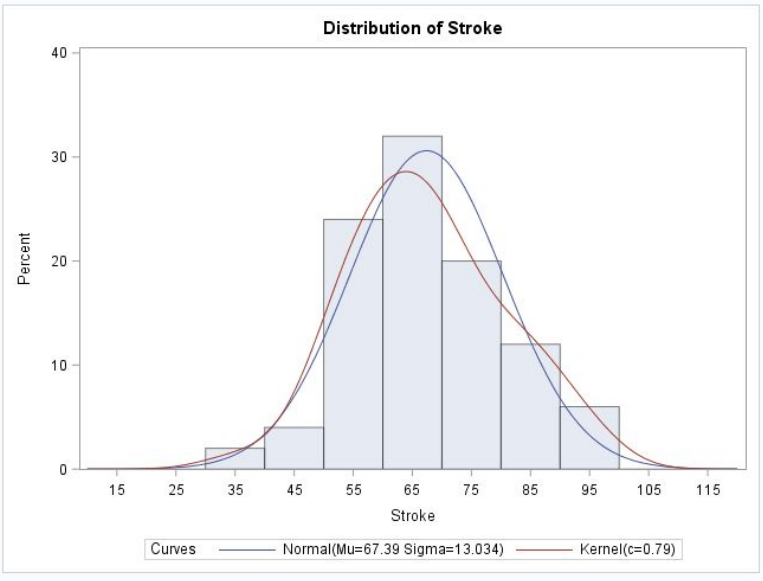


Fig. 8: Distribution of stroke before log transformation

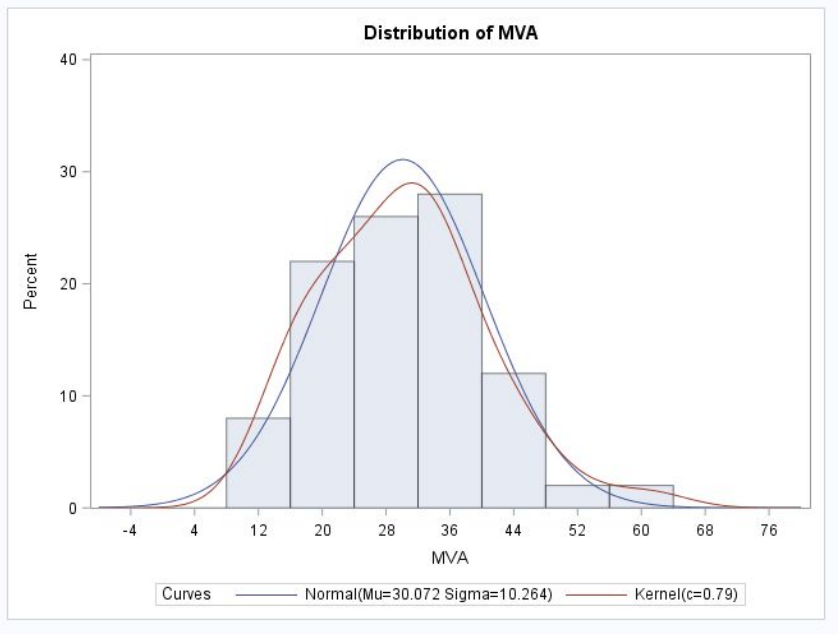


Fig. 9: Distribution of motor vehicle accident before log transformation

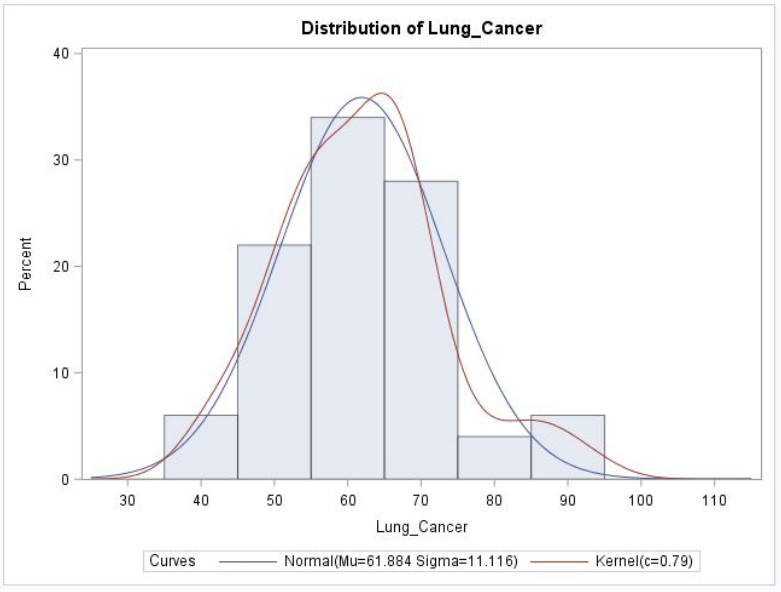


Fig. 10: Distribution of lung cancer before log transformation

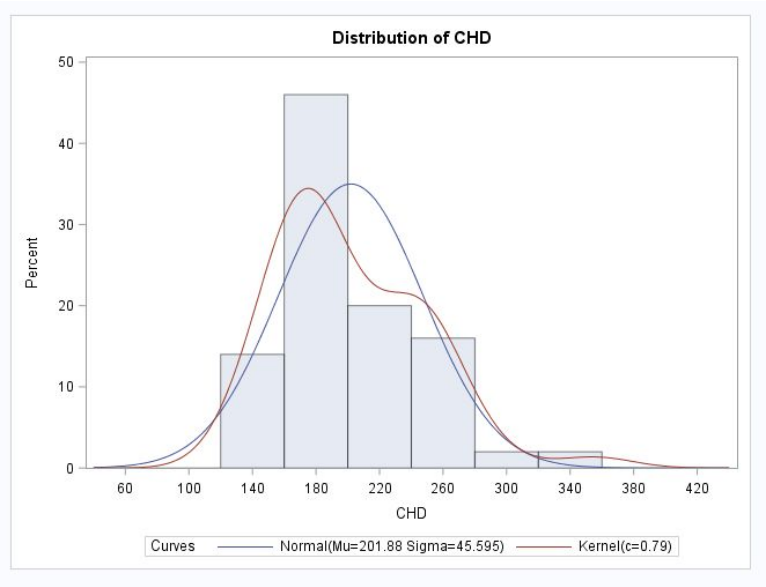


Fig. 11: Distribution of heart disease before log transformation

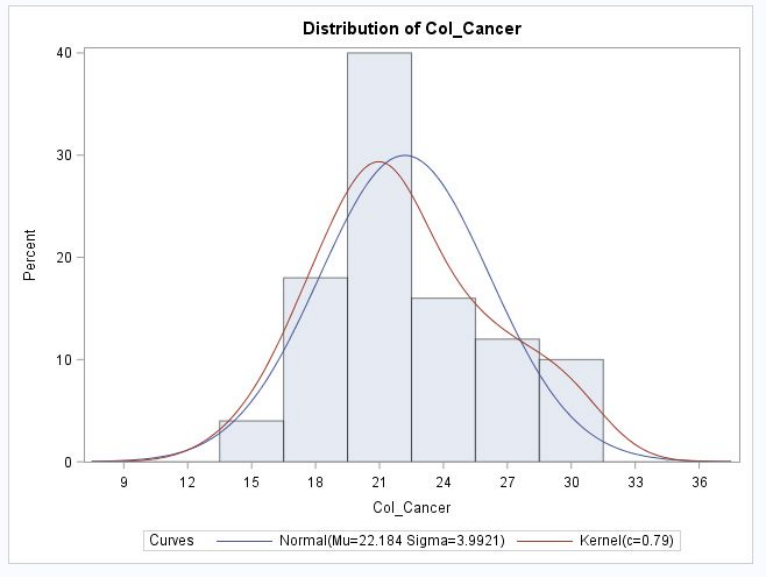


Fig. 12: Distribution of colon cancer before log transformation

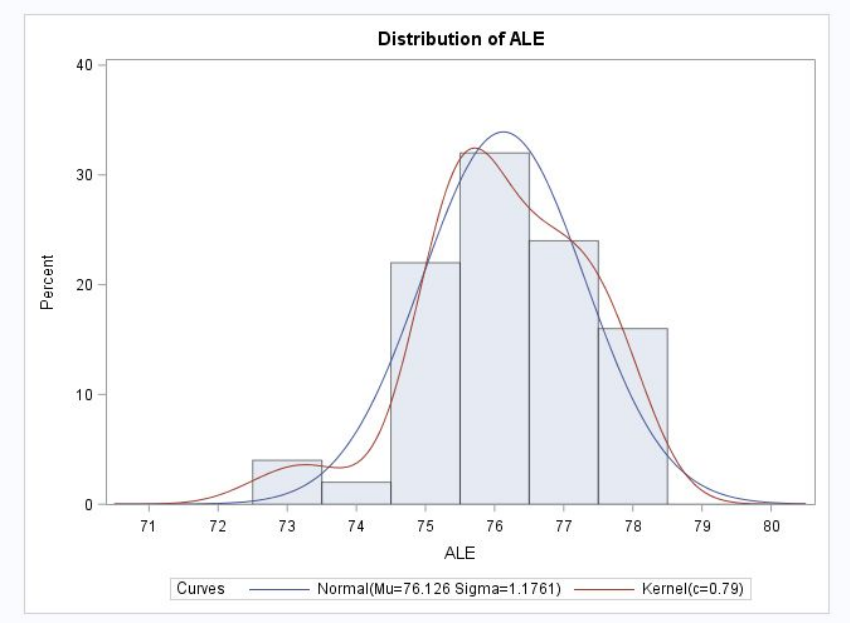


Fig. 13: Distribution of average life expectancy before log transformation

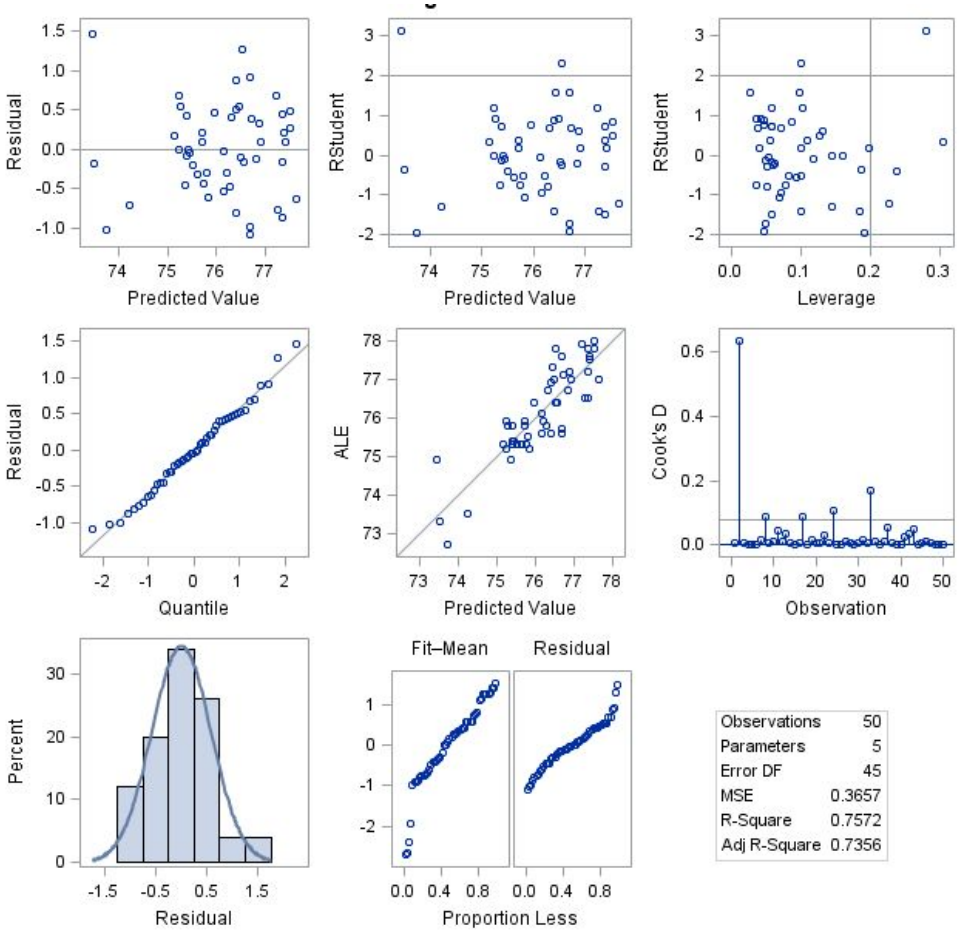


Fig. 14: Diagnostic for the final model (after log transformation and with interaction term)

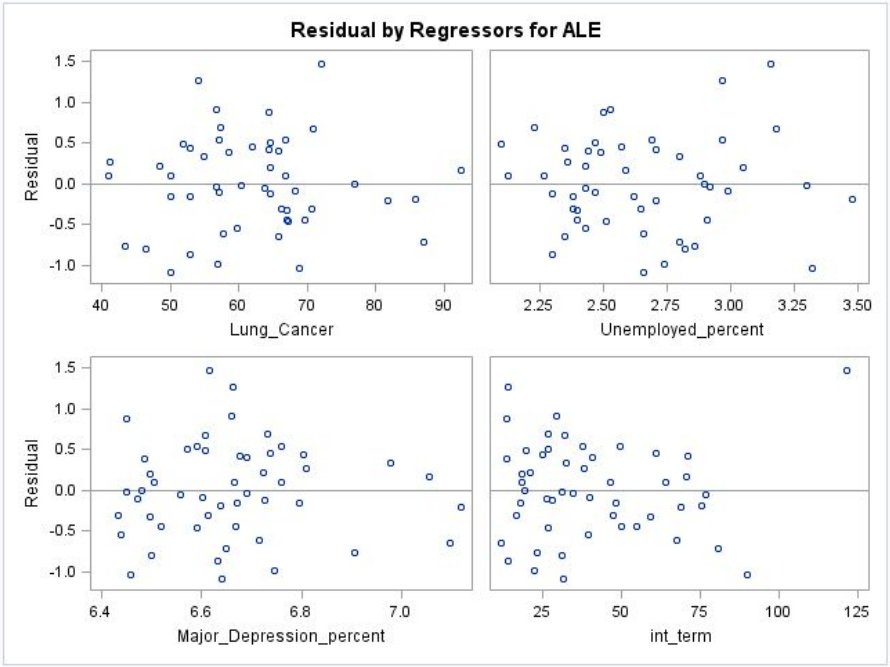


Fig. 15: Residual plots for the final model (after log transformation and with interaction term)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	71.76565	3.67849	19.51	<.0001	64.35679	79.17450
Lung_Cancer	1	-0.03617	0.00919	-3.94	0.0003	-0.05468	-0.01767
Unemployed_percent	1	-1.25721	0.30429	-4.13	0.0002	-1.87008	-0.64434
Major_Depression_percent	1	1.61220	0.53121	3.03	0.0040	0.54230	2.68210
int_term	1	-0.01984	0.00419	-4.73	<.0001	-0.02828	-0.01139

Fig. 16: Parameter estimates with the 95% confidence limits of the final model (after log transformation and with interaction term)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	71.76565	3.67849	19.51	<.0001	289758	139.17728
Lung_Cancer	1	-0.03617	0.00919	-3.94	0.0003	28.89272	5.66949
Unemployed_percent	1	-1.25721	0.30429	-4.13	0.0002	11.46695	6.24183
Major_Depression_percent	1	1.61220	0.53121	3.03	0.0040	2.77913	3.36810
int_term	1	-0.01984	0.00419	-4.73	<.0001	8.18284	8.18284

Fig. 17: Parameter estimates with Type I and Type II SS for the final model

Pearson Correlation Coefficients, N = 50														
	County_FIPS_Code	Col_Cancer	CHD	Lung_Cancer	MVA	Stroke	Injury	Prim_Care_Phys_Rate	Population_Size	Uninsured_percent	Disabled_Medicare_percent	Unemployed_percent	Major_Depression_percent	ALE
County_FIPS_Code	1.00000	0.00659	0.24140	0.06169	0.17915	0.21034	-0.05731	-0.00647	0.26654	0.14076	0.19999	0.13486	-0.02493	-0.21111
Col_Cancer	0.00659	1.00000	0.19557	0.03871	0.19177	0.00430	0.03271	-0.33186	-0.32764	0.16626	0.22080	0.17814	0.28277	-0.14553
CHD	0.24140	0.19557	1.00000	0.37246	0.40627	0.27406	0.43054	-0.11933	-0.17709	0.39347	0.50762	0.41410	-0.14002	-0.47019
Lung_Cancer	0.06169	0.03871	0.37246	1.00000	0.19783	0.11198	0.23527	0.01631	0.03561	0.30096	0.49415	0.40171	0.05420	-0.65291
MVA	0.17915	0.19177	0.40627	0.19783	1.00000	0.07361	0.32480	-0.46235	-0.43506	0.64239	0.51420	0.22775	0.21259	-0.35043
Stroke	0.21034	0.00430	0.27406	0.11198	0.07361	1.00000	-0.00581	0.01706	-0.13179	0.16707	0.27656	0.09050	-0.13602	-0.23868
Injury	-0.05731	0.03271	0.43054	0.23527	0.32480	-0.00581	1.00000	0.15290	0.04350	0.24629	0.39421	0.34882	-0.15626	-0.41080
Prim_Care_Phys_Rate	-0.00647	-0.33186	-0.11933	0.01631	-0.46235	0.01706	0.15290	1.00000	0.39955	-0.20566	-0.03573	-0.17342	-0.19384	0.06325
Population_Size	0.26654	-0.32764	-0.17709	0.03561	-0.43506	-0.13179	0.04350	0.39955	1.00000	-0.22854	-0.21137	0.15923	-0.32073	-0.03005
Uninsured_percent	0.14076	0.16626	0.39347	0.30096	0.64239	0.16707	0.24629	-0.20566	-0.22854	1.00000	0.71270	0.25385	0.02457	-0.56903
Disabled_Medicare_percent	0.19999	0.22080	0.50762	0.49415	0.51420	0.27656	0.39421	-0.03573	-0.21137	0.71270	1.00000	0.35099	0.08393	-0.65342
Unemployed_percent	0.13486	0.17814	0.41410	0.40171	0.22775	0.09050	0.34882	-0.17342	0.15923	0.25385	0.35099	1.00000	-0.09161	-0.63896
Major_Depression_percent	-0.02493	0.28277	-0.14002	0.05420	0.21259	-0.13602	-0.15626	-0.19384	-0.32073	0.02457	0.08393	-0.09161	1.00000	0.21617
ALE	-0.21111	-0.14553	-0.47019	-0.65291	-0.35043	-0.23868	-0.41080	0.06325	-0.03005	-0.56903	-0.65342	-0.63896	0.21617	1.00000

Fig. 18: Pearson Correlation Coefficients before the log transformation

Pearson Correlation Coefficients, N = 50																	
	County_Code	Col_Cancer	CHD	Lung_Cancer	MVA	Stroke	Injury	Prim_Care_Phys_Rate	Population_Size	Uninsured_percent	Disabled_Medicare_percent	Unemployed_percent	Major_Depression_percent	ALE	logpop	int_term
County_Code	1.00000	0.00659	0.24140	0.06169	0.17915	0.21034	-0.05731	-0.00647	0.26654	0.14076	0.19999	0.13486	-0.02493	-0.21111	0.09836	0.16956
Col_Cancer	0.00659	1.00000	0.19557	0.03871	0.19177	0.00430	0.03271	-0.33186	-0.32764	0.16626	0.22080	0.17814	0.28277	-0.14553	-0.41696	0.17628
CHD	0.24140	0.19557	1.00000	0.37246	0.40627	0.27406	0.43054	-0.11933	-0.17709	0.39347	0.50762	0.41410	-0.14002	-0.47019	-0.18643	0.52046
Lung_Cancer	0.06169	0.03871	0.37246	1.00000	0.19783	0.11198	0.23527	0.01631	0.03561	0.30096	0.49415	0.40171	0.05420	-0.65291	0.15637	0.46410
MVA	0.17915	0.19177	0.40627	0.19783	1.00000	0.07361	0.32480	-0.46235	-0.43506	0.64239	0.51420	0.22775	0.21259	-0.35043	-0.64979	0.60756
Stroke	0.21034	0.00430	0.27406	0.11198	0.07361	1.00000	-0.00581	0.01706	-0.13179	0.16707	0.27656	0.09050	-0.13602	-0.23868	-0.10839	0.22706
Injury	-0.05731	0.03271	0.43054	0.23527	0.32480	-0.00581	1.00000	0.15290	0.04350	0.24629	0.39421	0.34882	-0.15626	-0.41080	-0.00753	0.39094
Prim_Care_Phys_Rate	-0.00647	-0.33186	-0.11933	0.01631	-0.46235	0.01706	0.15290	1.00000	0.39955	-0.20566	-0.03573	-0.17342	-0.19384	0.06325	0.58778	-0.10709
Population_Size	0.26654	-0.32764	-0.17709	0.03561	-0.43506	-0.13179	0.04350	0.39955	1.00000	-0.22854	-0.21137	0.15923	-0.32073	-0.03005	0.78593	-0.22687
Uninsured_percent	0.14076	0.16626	0.39347	0.30096	0.64239	0.16707	0.24629	-0.20566	-0.22854	1.00000	0.71270	0.25385	0.02457	-0.56903	-0.42116	0.87479
Disabled_Medicare_percent	0.19999	0.22080	0.50762	0.49415	0.51420	0.27656	0.39421	-0.03573	-0.21137	0.71270	1.00000	0.35099	0.08393	-0.65342	-0.26927	0.94653
Unemployed_percent	0.13486	0.17814	0.41410	0.40171	0.22775	0.09050	0.34882	-0.17342	0.15923	0.25385	0.35099	1.00000	-0.09161	-0.63896	-0.01536	0.34837
Major_Depression_percent	-0.02493	0.28277	-0.14002	0.05420	0.21259	-0.13602	-0.15626	-0.19384	-0.32073	0.02457	0.08393	-0.09161	1.00000	0.21617	-0.42333	0.05471
ALE	-0.21111	-0.14553	-0.47019	-0.65291	-0.35043	-0.23868	-0.41080	0.06325	-0.03005	-0.56903	-0.65342	-0.63896	0.21617	1.00000	0.00177	-0.66628
logpop	0.09836	-0.41696	-0.18643	0.15637	-0.64979	-0.10839	-0.00753	0.58778	0.78593	-0.42116	-0.26927	-0.01536	-0.42333	0.00177	1.00000	-0.33037
int_term	0.16956	0.17628	0.52046	0.46410	0.60756	0.22706	0.39094	-0.10709	-0.22687	0.87479	0.94653	0.34837	0.05471	-0.66628	-0.33037	1.00000

Fig. 19: Pearson Correlation Coefficients after the log transformation

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44.26408	8.85282	16.57	<.0001
Error	44	23.51212	0.53437		
Corrected Total	49	67.77620			

Root MSE	0.73100	R-Square	0.6531
Dependent Mean	76.12600	Adj R-Sq	0.6137
Coeff Var	0.96026		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.94658	4.67917	15.38	<.0001
Lung_Cancer	1	-0.05212	0.01040	-5.01	<.0001
Unemployed_percent	1	-1.63581	0.37975	-4.31	<.0001
Major_Depression_percent	1	1.75299	0.67783	2.59	0.0131
Prim_Care_Phys_Rate	1	-0.00048096	0.00319	-0.15	0.8810
Population_Size	1	0.00000162	0.00000117	1.38	0.1743

Fig. 20: Proc Reg results for the first model (on original data)

Fig. 21: Scatter Plot Matrix