Exploring Factors related to Cancer Incidence in the US

A STATISTICAL APPROACH TO FACTOR RELEVANCE

AUTHOR:
MUHAMMAD KHIZAR HAYAT TAHIR

Table of Contents

Research objective:

To explore factors related to cancer incidence and deaths in the United States and create a model that identifies key factors from the rest and explains how they impact the incidence rate. With this, we expect to enable the American Cancer Society to identify key regions and an intervention plan.

Executive summary:

In this project, we will use a dataset containing information about factors leading to cancer occurrence and also deaths in the US, arranged according to zip codes.

We want to know how different factors affect the Probability of occurrence of cancer in a population in a specified period of time, more formally known as the Incidence Rate of cancer. Therefore, the Incidence rate of cancer serves as our Response (Dependent) variable of interest.

From the dataset, we identify the following Independent variables (Factors):

1. *Study count:* the number of cancer clinical trials held for all types of cancer by Zip Code.
2. *Poverty Percentage*: percentage of county population below the poverty line.
3. *Median Income:* the median household income by county.
4. *Population Estimate:* the number of people in a county.
5. *Average annual count:* mean incidences per county.
6. *State:* States in the United States.
7. *Recent Trend:* recent trend of incidence of cancer.

Note:

Study Count is by Zip Code, other 3 factors are by County. Incidence Rate is also by County.

For full list of terms used, see appendix.

Problem to solve:

Knowing that there are several factors related to cancer incidence in the US, we want to use statistical procedures to develop a model that explains which factors are related to our response variable IncidenceRate and how; and which of them do not add any value.

Design of Analysis:

We will do our analysis to create our model in 5 broad steps:

1. We will first identify the regions of the US where the incidence rate of cancer is highest. This will give us an overall idea of the spread and prevalence of the disease with respect to regions.
2. We will conduct hypothesis tests individually for two of our factors, recent trend of incidence rate and study count in the area, to understand if they lead to **more** incidence rate of cancer. Each hypothesis test will be conducted separately so this step will have two substeps.
3. In this step, we will conduct hypothesis testing on another factor: Median Income. We will break this factor into 4 groups (populations). We will use the statistical procedure of Analysis of Variance (ANOVA) to determine if the mean incidence rate for each of the populations created by breaking this variable is the same or not.
4. We will then do a correlation analysis to understand which of our factors are most highly correlated to our response variable; and also, to understand their correlation with each other.
5. Finally, we will create an actionable regression model for our factors and response variable. Our goal will be to explain as much variability in incidence rate as possible from our factors. We will only identify key factors. Less impacting factors may be dropped.

Data cleaning:

To conduct our analyses, we will perform the following data cleaning procedures in their mentioned steps:

- For step 2(a), we need to subset our dataset to keep only two values for the recentTrend variable: Stable and Falling. It originally has other values too which are irrelevant to us and might contaminate our results. Also, we will only keep columns that are the same length.
- All further procedure will be carried on this subset data.
- For step 2(b), we need to create two categories for our studyCount variable: High and Low. All Zip Codes with studyCount less than the mean study count will be tagged Low and vice versa. We will actually recode to numeric 0,1.
- For step 3, we will break our median income variable into 4 categories: Very low, Low, High, Very high and recode these categories as numeric to 0,1,2,3.

- For step 4, we will ensure we only deal with continuous variables for determining the correlations.
- For step 5, we would ideally like to avoid multicollinearity problems that would be evident from step 4 if it exists by carefully selecting continuous predictors that avoid it. We will also recode categorical factors that are further explained under step 5 - method.

Note**:** Details of exactly how this cleaning is achieved is under Methods section.

Methods & Results:

Step 1 - Method:

- In order to determine which regions of the country (States) are most prone to cancer, meaning highest mean incidence rate, we will run Minitab's 'Descriptive Statistics' on our dataset. We will use IncidenceRate as our variable, and State as our categorical by variable and ensure the Mean is marked as the primary statistic to calculate. We will store the results in the worksheet.
- We will then use the 'Custom Sort' tool to store newly added columns (State name and its Mean incidence rate) into a new worksheet in decreasing order of mean incidence rates. Then, the starting rows tell us the regions most prone to cancer as they will have highest mean incidence rates.

Step 1 - Results:

- **Top 10 States w/ highest incidence rates:**

  - Kentucky: Avg. 517.33
  - Delaware: Avg. 498.19
  - New York: Avg. 497.52
  - New Jersey: Avg. 494.01
  - Louisiana: Avg. 486.81
  - New Hampshire: Avg. 485.15

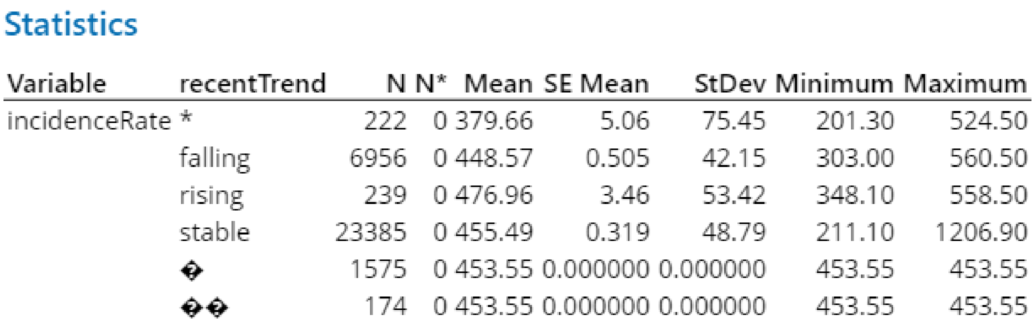
- DC: Avg. 483.70
- Illinois: Avg. 482.91
- Connecticut: Avg. 482.50
- Pennsylvania: Avg. 481.62

The 26 states East of the Mississippi River
MAINE
VERMONT
NEW HAMPSHIRE
MASSACHUSETTS
NEW YORK
RHODE ISLAND
CONNECTICUT
MICHIGAN
WISCONSIN
MICHIGAN
PENNSYLVANIA
NEW JERSEY
DELAWARE
MARYLAND
OHIO
INDIANA
ILLINOIS
WEST VIRGINIA
VIRGINIA
KENTUCKY
NORTH CAROLINA
TENNESSEE
SOUTH CAROLINA
MISSISSIPPI
ALABAMA
GEORGIA
FLORIDA

Step 2(a) - Method:

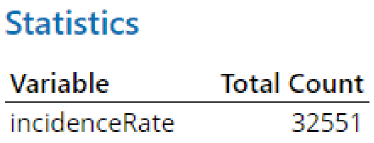
- We will first subset our data to keep only those rows that have either Stable or Falling under recentTrend. This will make it a dichotomous variable.

  The *Pre-cleaning* view shows 6 different inputs for recentTrend and a total number of observations, N, as 32,551.



Statistics

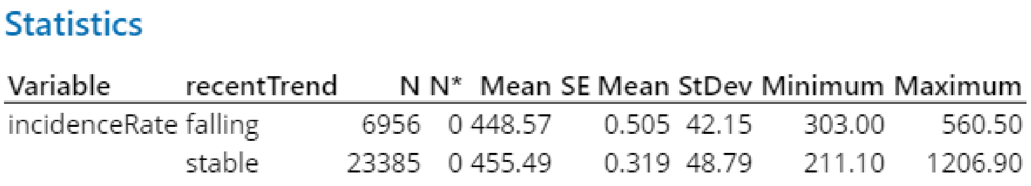
Variable	recentTrend	N	N*	Mean	SE Mean	StDev	Minimum	Maximum
incidenceRate	*	222	0	379.66	5.06	75.45	201.30	524.50
	falling	6956	0	448.57	0.505	42.15	303.00	560.50
	rising	239	0	476.96	3.46	53.42	348.10	558.50
	stable	23385	0	455.49	0.319	48.79	211.10	1206.90
	�	1575	0	453.55	0.000000	0.000000	453.55	453.55
	��	174	0	453.55	0.000000	0.000000	453.55	453.55



Statistics

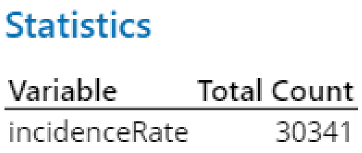
Variable	Total Count
incidenceRate	32551

Post-cleaning view shows only 2 categories for recentTrend and a total number of observations N as 30341.



Statistics

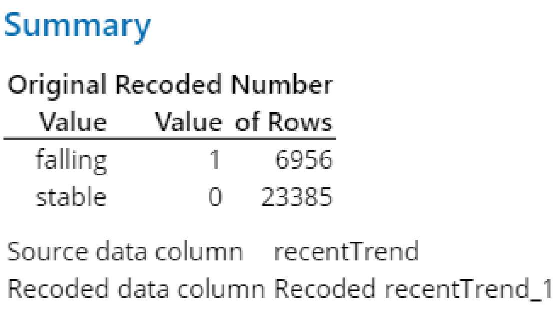
Variable	recentTrend	N	N*	Mean	SE Mean	StDev	Minimum	Maximum
incidenceRate	falling	6956	0	448.57	0.505	42.15	303.00	560.50
	stable	23385	0	455.49	0.319	48.79	211.10	1206.90



Statistics

Variable	Total Count
incidenceRate	30341

- We will Recode these two text values to numeric - **0 for Stable, 1 for Falling** from Minitab: Data -> recode -> to numeric. We want to showcase Stable vs Falling and Minitab does mean of population 0 - mean of population 1, hence this recoding.



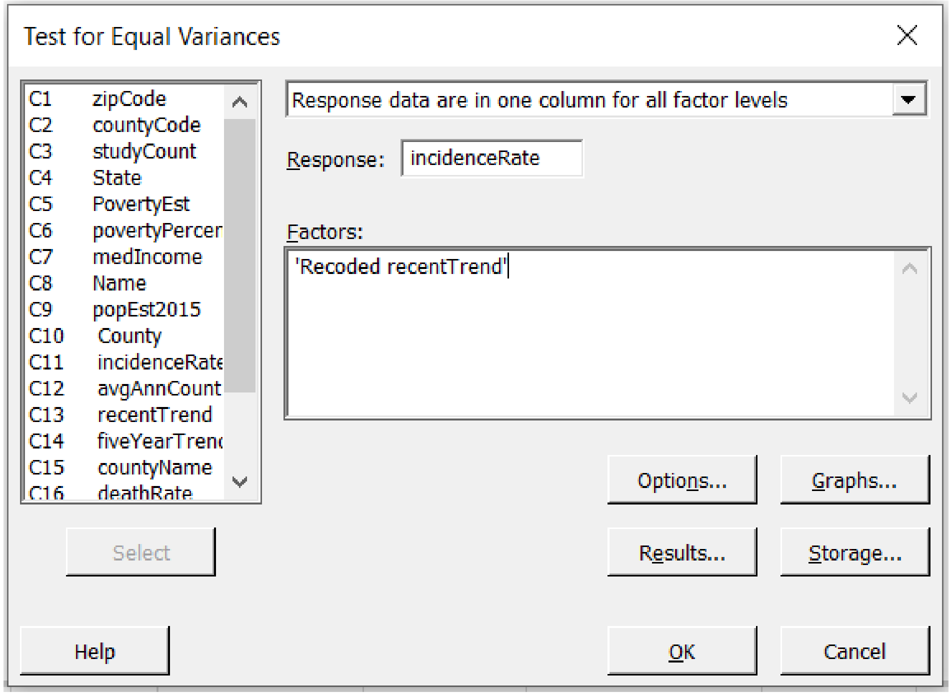
Summary

Original	Recoded	Number
Value	Value	of Rows
falling	1	6956
stable	0	23385

Source data column recentTrend
Recoded data column Recoded recentTrend_1

- Then, we will essentially have 2 populations: one of all those counties with recent trend as Stable (Population 1), the second population of all those counties with recent trend as Falling (Population 2).

- Stable vs Falling: stable trend means that there has not been much variation in incidence rate recently. Falling means that there has been a decline in incidence rate recently.
- We will do a hypothesis test to determine if the recent trend leads to **more** incidence rate. Let 'd' = Mean incidence rate of population 1 (Stable) - Mean incidence rate of population 2 (Falling).
- Our Null hypothesis will be - H(0): d <= 0.
- Our Alternative hypothesis will be - H(a): d > 0. In other words, our Alternative hypothesis states that the mean incidence rate of the 'Stable' population is higher than the other's.
- Because the two populations are Independent, we will use a **2 sample T-test** and its resulting p-value to determine the results of the hypothesis test.
- Before we run the test, we will do a Test for Equal Variance between the two populations with regards to our response variable to know if we can assume equal variances or not.



Test for Equal Variances

Response data are in one column for all factor levels

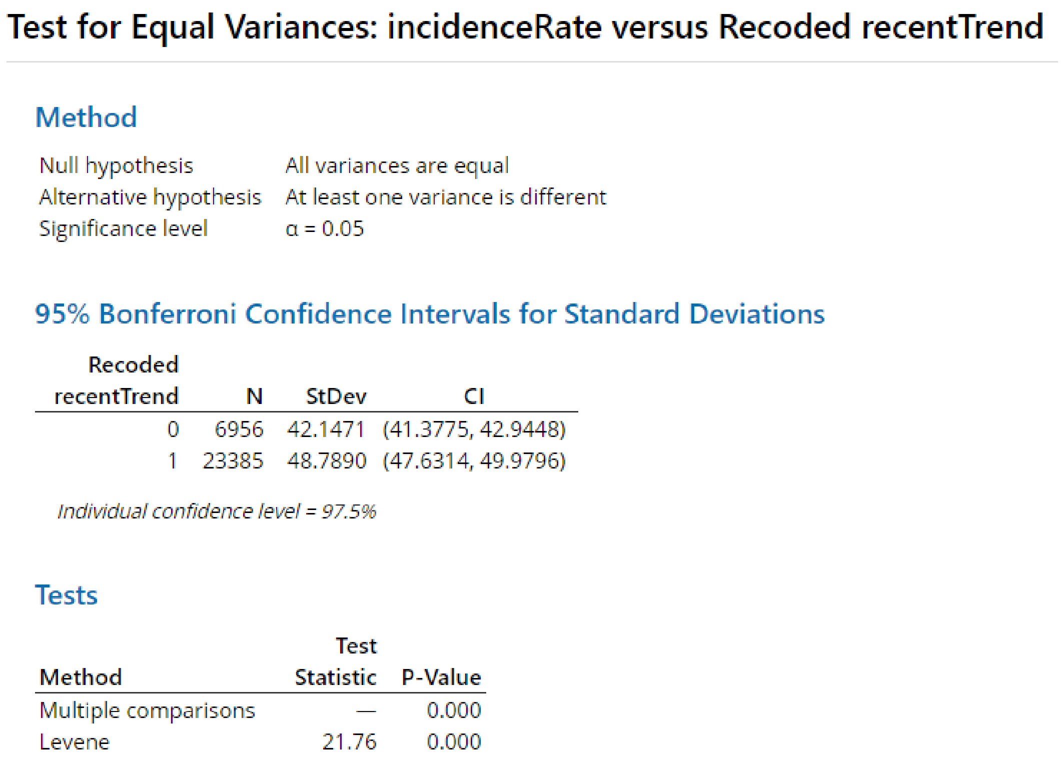
Response: incidenceRate

Factors: 'Recoded recentTrend'

- Then, we will run the T-test accordingly. If the resulting p-value is less than or equal to our alpha, we will reject the NULL hypothesis.

Step 2(a) - Results:

- Test for Equal Variances resulted in a Levene method's p-value of 0.000. Since this p-value is < alpha, we reject the Null hypothesis stating that 'All variances are equal'. We cannot assume equal variances. At least one variance is different at alpha of 0.05. Plot of simultaneous intervals also do not overlap meaning standard deviations are significantly different. Results below:



Test for Equal Variances: incidenceRate versus Recoded recentTrend

Method

Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	α = 0.05

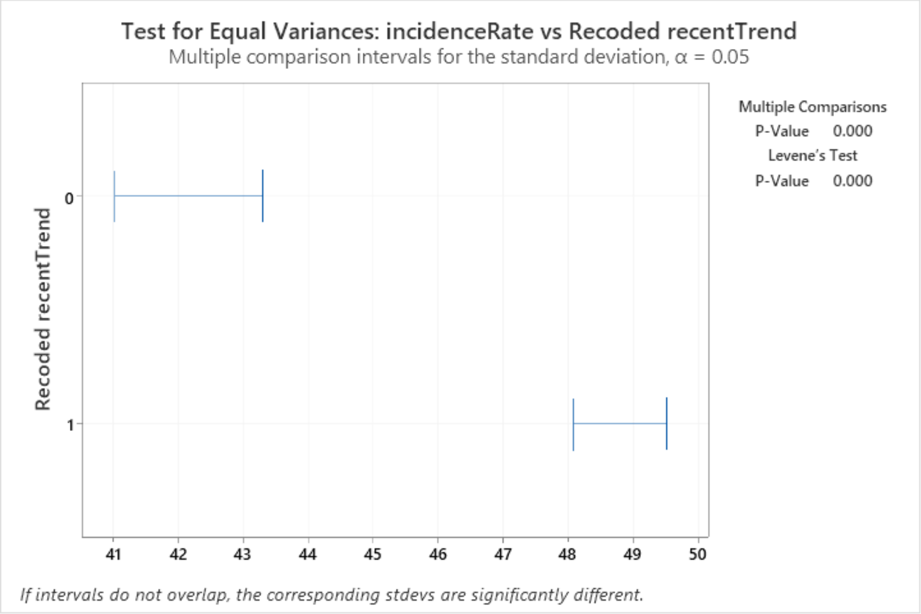
95% Bonferroni Confidence Intervals for Standard Deviations

Recoded recentTrend	N	StDev	CI
0	6956	42.1471	(41.3775, 42.9448)
1	23385	48.7890	(47.6314, 49.9796)

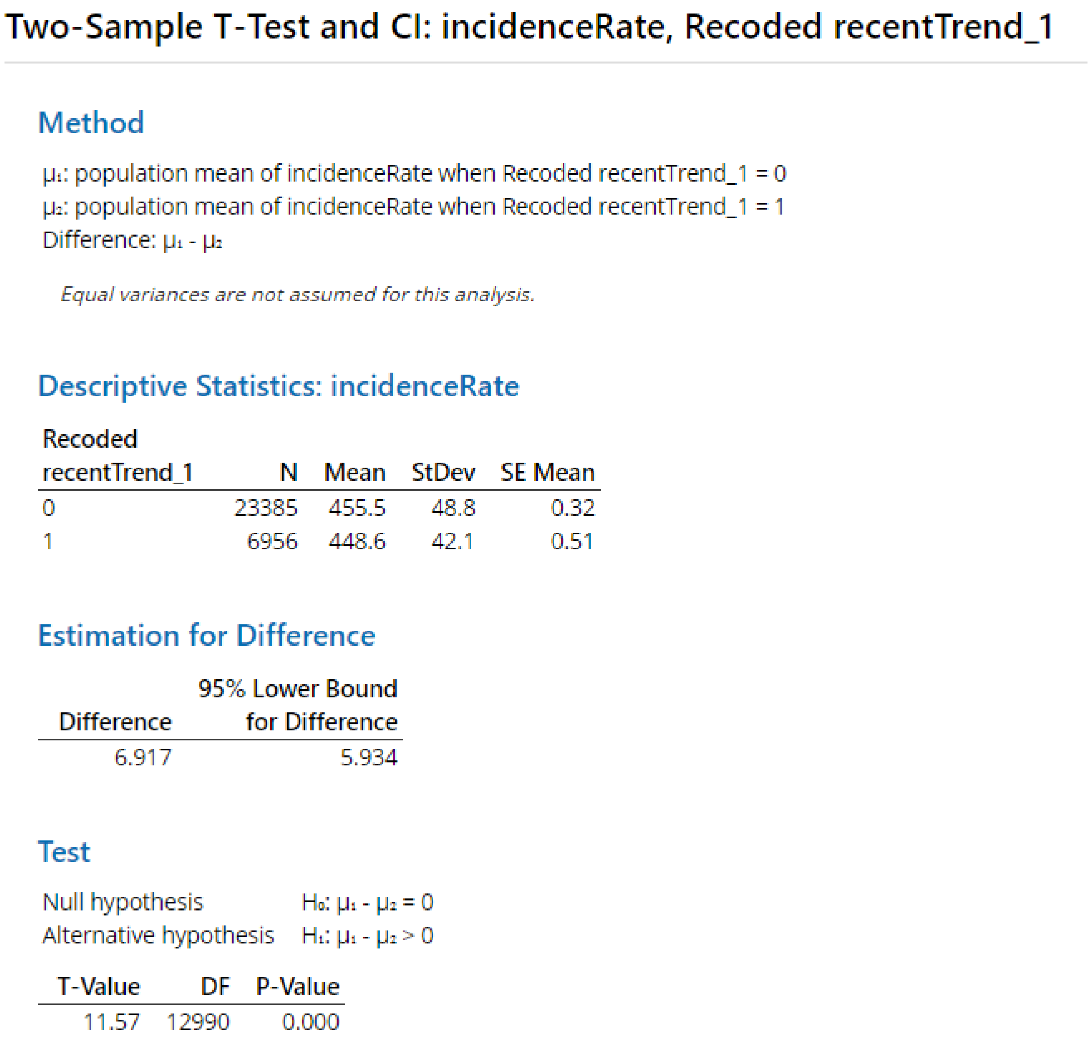
Individual confidence level = 97.5%

Tests

Method	Test Statistic	P-Value
Multiple comparisons	—	0.000
Levene	21.76	0.000



- From the hypothesis test (Without assuming equal variance) results below, we see that the p-value (0.000) corresponding to the test statistic T (11.57) is less than our Alpha (0.05).



Two-Sample T-Test and CI: incidenceRate, Recoded recentTrend_1

Method

μ_1: population mean of incidenceRate when Recoded recentTrend_1 = 0
μ_2: population mean of incidenceRate when Recoded recentTrend_1 = 1
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics: incidenceRate

Recoded recentTrend_1	N	Mean	StDev	SE Mean
0	23385	455.5	48.8	0.32
1	6956	448.6	42.1	0.51

Estimation for Difference

Difference	95% Lower Bound for Difference
6.917	5.934

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

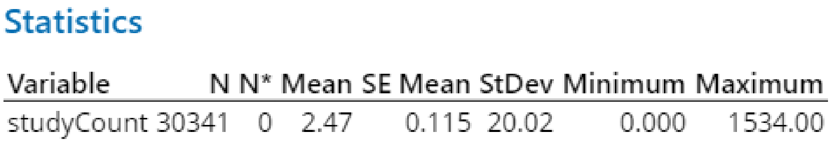
T-Value	DF	P-Value
11.57	12990	0.000

- **Conclusion:** This provides enough evidence for us to reject the Null hypothesis that the two population means are equal and that the difference is 0. Implicitly, we conclude that Stable recentTrend has a mean incidence rate of Cancer greater than Falling.

Step 2(b) - Method:

- We will first run descriptive statistics on the study count column, mainly to know two important values:
    - mean study count across the US. This will be used to break the population of study count into 2 groups.

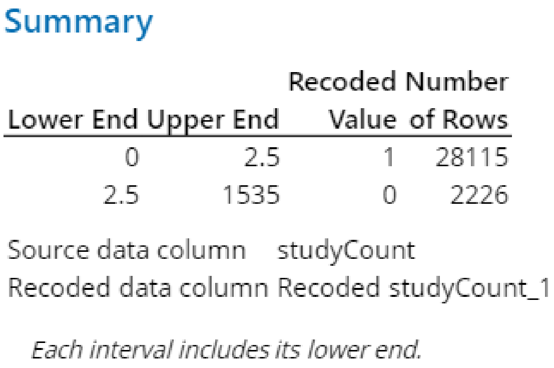
- ○ Maximum and Minimum values of the study count across the US to know the upper and lower limits.



Statistics

```
Variable       N N* Mean SE Mean StDev Minimum Maximum
studyCount 30341  0 2.47   0.115 20.02   0.000 1534.00
```

- Then we will recode the data using ‘recode ranges of values’. Study count of 0 till mean (ceiled to next integer) will be re-coded to 1 (Low). Study count of mean till Max (ceiled to next integer) count will be re-coded to 0 (High). We want to do High vs Low and Minitab does population mean 0 - population mean 1 hence this recoding. This will create 2 populations like in 2(a).



Summary

```
                     Recoded Number
Lower End Upper End    Value of Rows
        0       2.5        1   28115
      2.5      1535        0    2226

Source data column   studyCount
Recoded data column  Recoded studyCount_1

  Each interval includes its lower end.
```

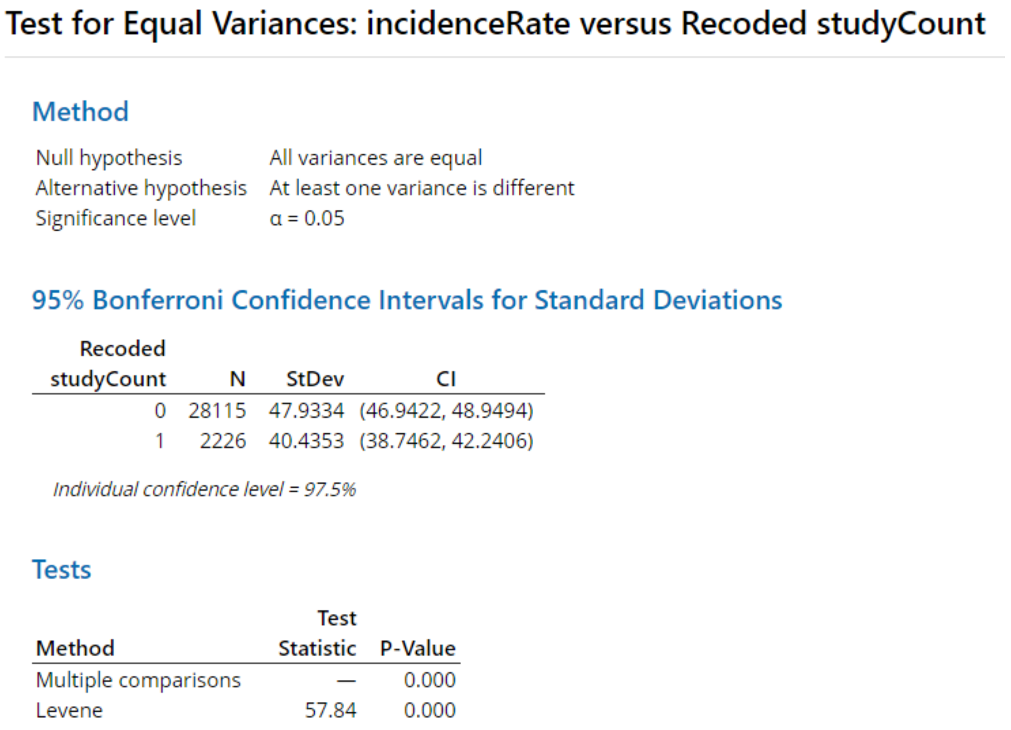
- We want to test for High vs Low. let ‘d’ be the difference between the mean incidence rate of Population (High) and mean incidence rate of population (Low). Our NULL hypothesis will be - H(0): d <= 0. Alternative hypothesis will be - H(a): d > 0. In other words, our Alternative hypothesis states that the mean incidence rate of the ‘High’ studyCount population is higher than the other’s.
- Because the two populations are Independent, we will use a **2 sample T-test** and its resulting p-value to determine the results of the hypothesis test.
- Before we run the test, we will do a Test for Equal Variance between the two populations with regards to our response variable to know if we can assume equal variances or not.

  This will be similar to the test in 2(a) but with Recoded studyCount.

- We can then do the 2 sample T-test like in 2(a) and get results.

Step 2(b) - Results:

- Test for Equal Variances resulted in a Levene method's p-value of 0.000. Since this p-value is < alpha, we reject the Null hypothesis stating that 'All variances are equal'. We cannot assume equal variances. At least one variance is different at alpha of 0.05. Plot of simultaneous intervals also do not overlap meaning standard deviations are significantly different. Results below:



Test for Equal Variances: incidenceRate versus Recoded studyCount

Method

Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	α = 0.05

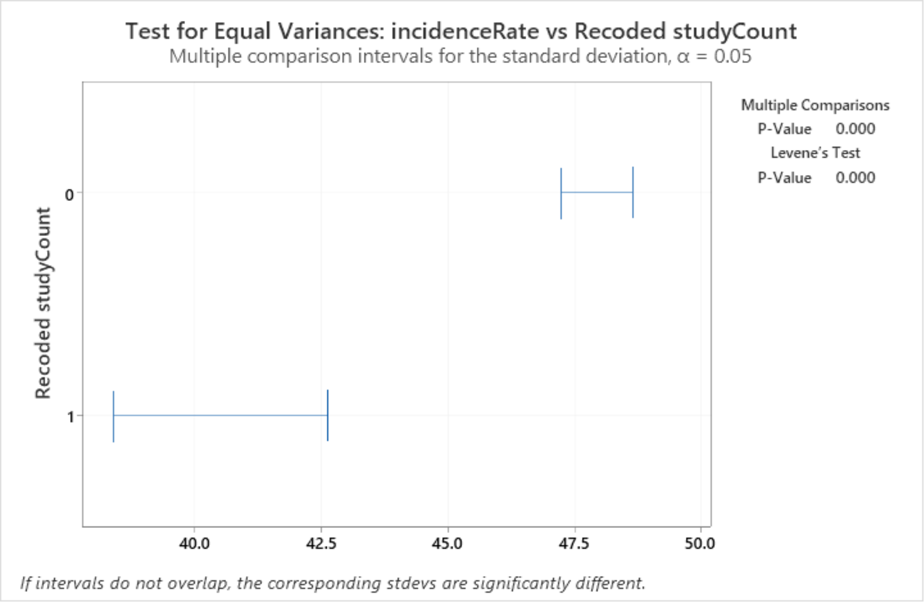
95% Bonferroni Confidence Intervals for Standard Deviations

Recoded studyCount	N	StDev	CI
0	28115	47.9334	(46.9422, 48.9494)
1	2226	40.4353	(38.7462, 42.2406)

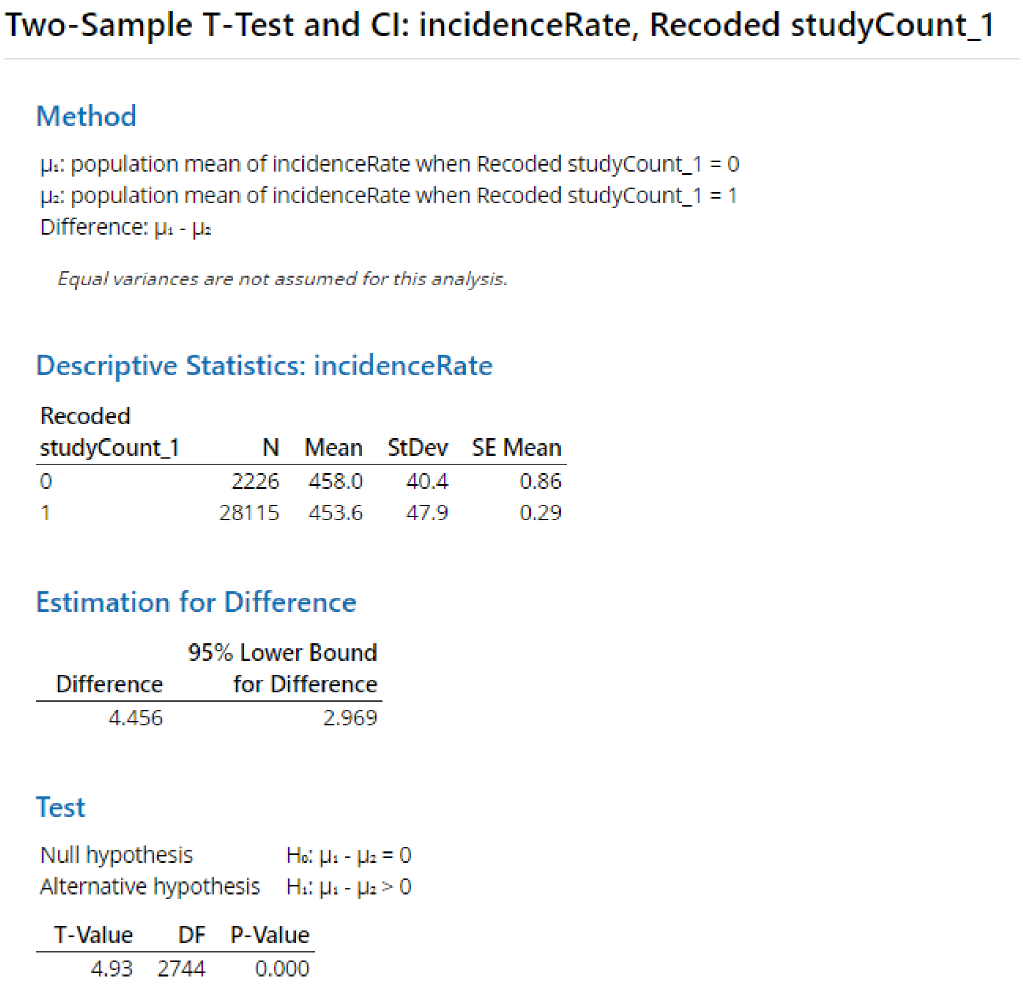
Individual confidence level = 97.5%

Tests

Method	Test Statistic	P-Value
Multiple comparisons	—	0.000
Levene	57.84	0.000



- From the hypothesis tests results below, we see that the p-value (0.000) corresponding to the test statistic T (4.93) is less than our Alpha (0.05).



Two-Sample T-Test and CI: incidenceRate, Recoded studyCount_1

Method

μ_1: population mean of incidenceRate when Recoded studyCount_1 = 0
μ_2: population mean of incidenceRate when Recoded studyCount_1 = 1
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics: incidenceRate

Recoded studyCount_1	N	Mean	StDev	SE Mean
0	2226	458.0	40.4	0.86
1	28115	453.6	47.9	0.29

Estimation for Difference

Difference	95% Lower Bound for Difference
4.456	2.969

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$

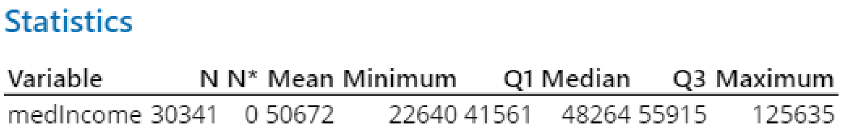
T-Value	DF	P-Value
4.93	2744	0.000

- **Conclusion:** This provides enough evidence for us to reject the Null hypothesis that the two population means are equal and that the difference between them is 0. Implicitly, we conclude that a higher studycount has a higher mean incidenceRate than lower.

Step 3 - Method:

- For this part, we first need to create 4 groups for the median income. This will be Very low, Low, High, Very high recoded as 0,1,2,3 respectively.

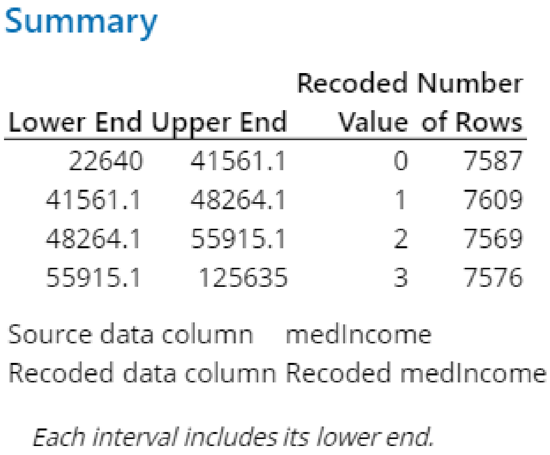
- We will run descriptive statistics on the median income variable to get: Q1, Median of the variable across US, Q3.



Statistics

Variable	N	N*	Mean	Minimum	Q1	Median	Q3	Maximum
medIncome	30341	0	50672	22640	41561	48264	55915	125635

- With the three markers, we will divide the median income into 4 groups (0-3 codes) using recode data. We add 0.1 to all values except the lowest limit so that quartile values fall into the correct range.
- We will call group 0 as population 1, group 1 as population 2 and so on.



Summary

		Recoded	Number
Lower End	Upper End	Value	of Rows
22640	41561.1	0	7587
41561.1	48264.1	1	7609
48264.1	55915.1	2	7569
55915.1	125635	3	7576

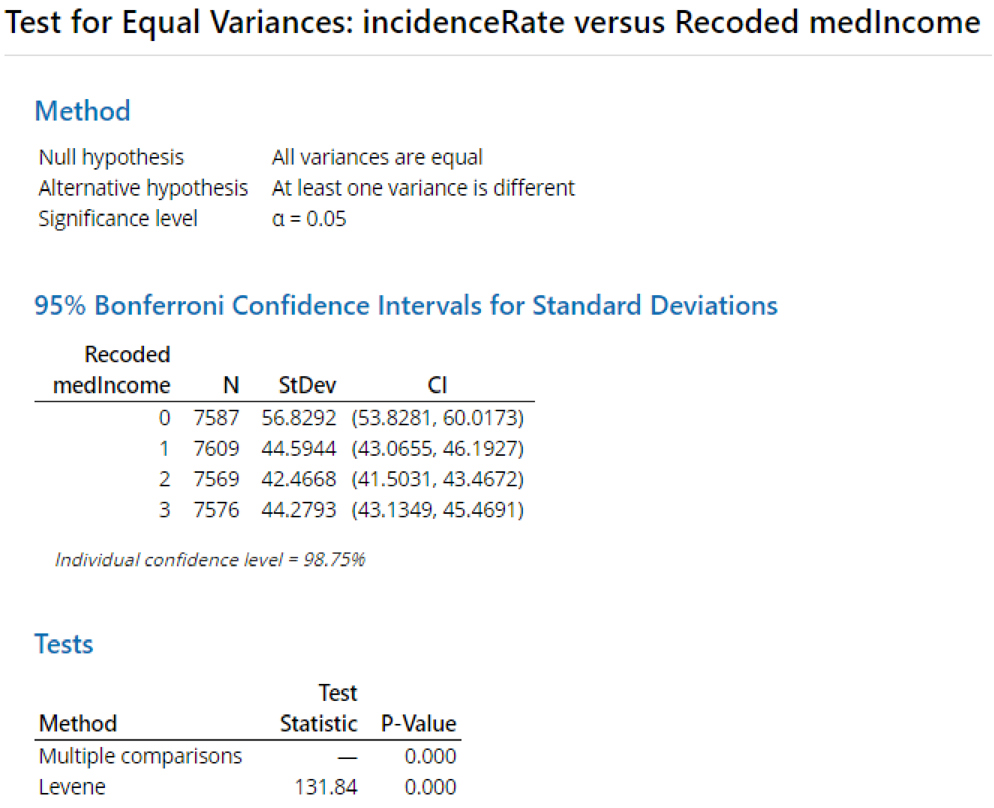
Source data column medIncome
Recoded data column Recoded medIncome

Each interval includes its lower end.

- NULL hypothesis - H(0): mean incidence rate of population 1 = mean incidence rate of population 2 = mean incidence rate of population 3 = mean incidence rate of population 4.
- Alternate hypothesis - H(a): Not all population mean incidence rates are equal.
- We will first **Test for equal variances** between the 4 populations to know if we can assume equal variance or not.
- Then we will run a **One-way ANOVA test with incidence rate as the Response variable and Recoded median income as Factor.**
- If the p-value from the test statistic (F) is less than our alpha, we will reject the null hypothesis.

Step 3 - Results:

- Test for Equal Variances resulted in a Levene method's p-value of 0.000. Since this p-value is < alpha, we reject the Null hypothesis stating that 'All variances are equal'. We cannot assume equal variances. At least one variance is different at alpha of 0.05. Plot of simultaneous intervals also do not overlap fully meaning standard deviation(s) are significantly different. Results below:



Test for Equal Variances: incidenceRate versus Recoded medIncome

Method

Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	α = 0.05

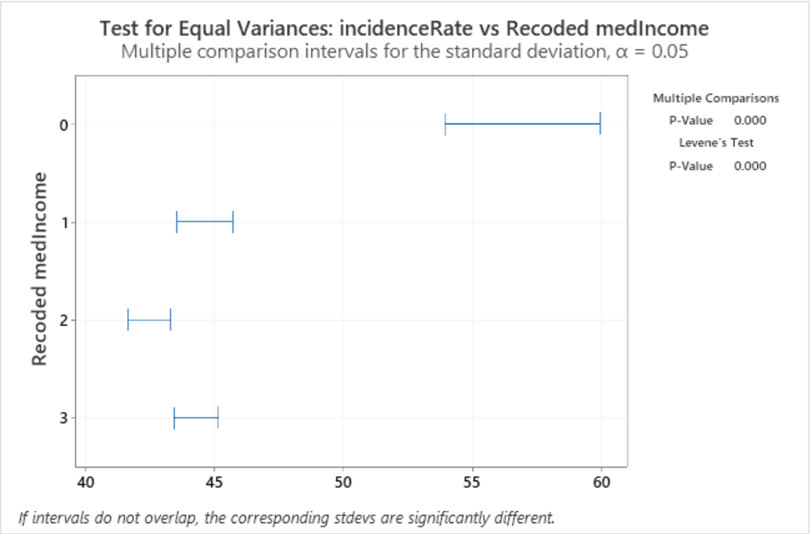
95% Bonferroni Confidence Intervals for Standard Deviations

Recoded medIncome	N	StDev	CI
0	7587	56.8292	(53.8281, 60.0173)
1	7609	44.5944	(43.0655, 46.1927)
2	7569	42.4668	(41.5031, 43.4672)
3	7576	44.2793	(43.1349, 45.4691)

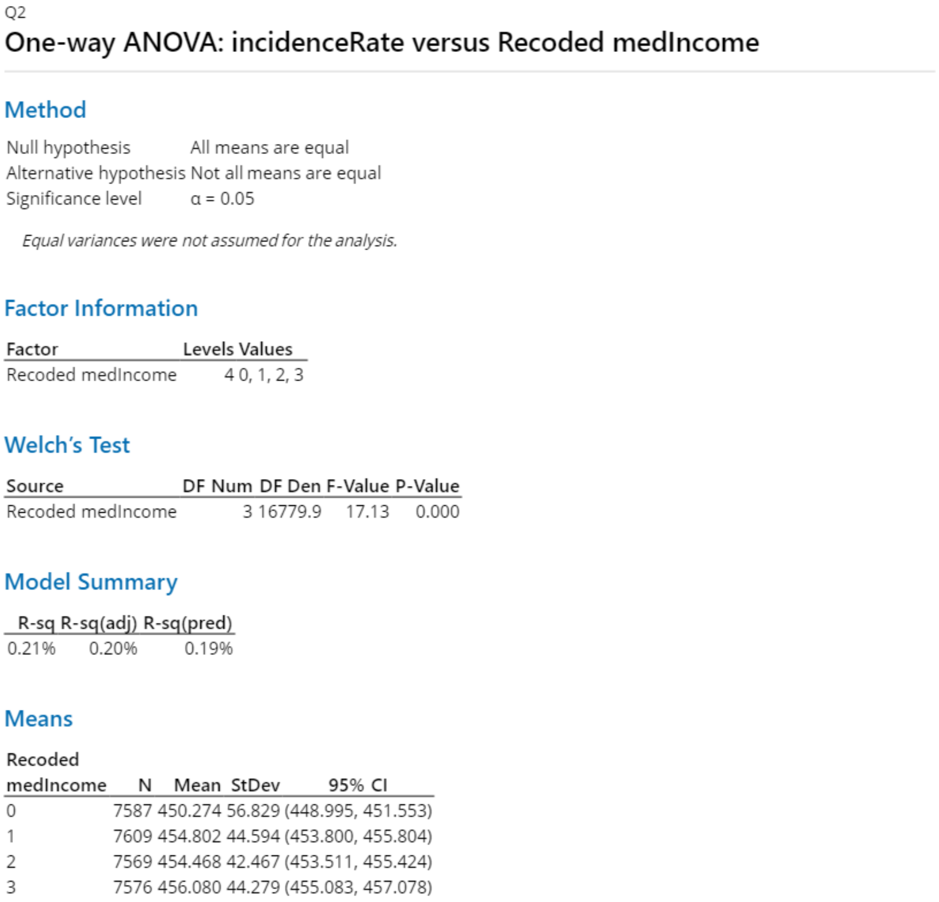
Individual confidence level = 98.75%

Tests

Method	Test Statistic	P-Value
Multiple comparisons	—	0.000
Levene	131.84	0.000



- One-Way ANOVA (without assuming equal variances):



Q2

One-way ANOVA: incidenceRate versus Recoded medIncome

Method

```
Null hypothesis         All means are equal
Alternative hypothesis  Not all means are equal
Significance level      α = 0.05
```

Equal variances were not assumed for the analysis.

Factor Information

```
Factor             Levels Values
Recoded medIncome       4 0, 1, 2, 3
```

Welch's Test

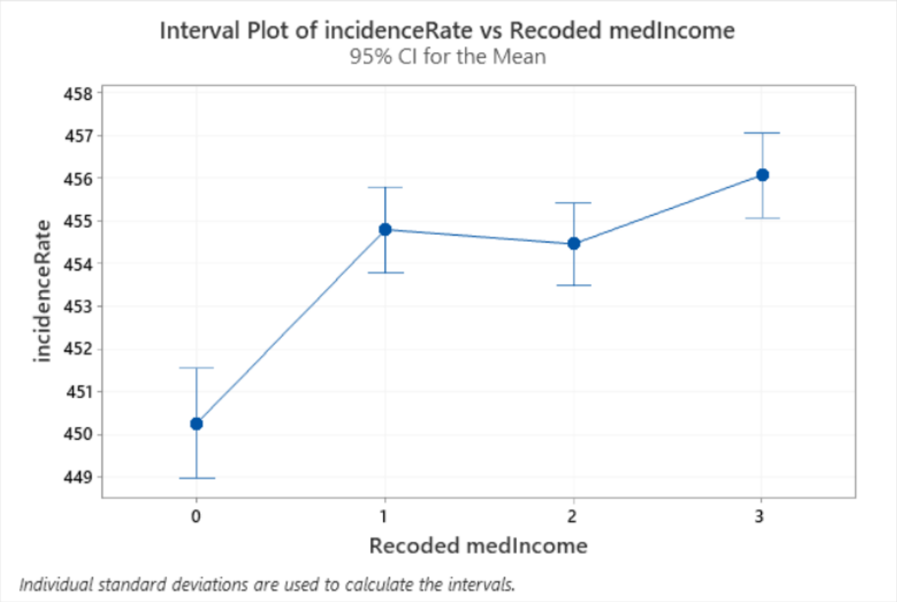
```
Source             DF Num  DF Den F-Value P-Value
Recoded medIncome       3 16779.9   17.13   0.000
```

Model Summary

```
 R-sq R-sq(adj) R-sq(pred)
0.21%     0.20%      0.19%
```

Means

```
Recoded
medIncome     N    Mean  StDev       95% CI
0          7587 450.274 56.829 (448.995, 451.553)
1          7609 454.802 44.594 (453.800, 455.804)
2          7569 454.468 42.467 (453.511, 455.424)
3          7576 456.080 44.279 (455.083, 457.078)
```



Individual standard deviations are used to calculate the intervals.

- ANOVA used Welch’s test as the variances were not equal.

- **Conclusion:** Since the p-value (0.000) corresponding to the F-statistic (17.13) of ANOVA is less than our alpha (0.05), we will reject the Null hypothesis that all population means are equal. Implicitly, we conclude that at least one population mean is different. So, incidence rate is not the same for all groups of income.

Step 4 - Method:

- For this step, we identified variables that could potentially impact cancer incidence rate, i.e., our aim was to find the correlation between different factors and cancer incidence rate. Since some of the variables were qualitative and could not be re-coded into indicator variables, they were dropped from consideration. We only used **continuous variables.**

- We used the correlation function on Minitab to calculate the Pearson Correlation coefficients for each of the selected continuous variables with the response variable (incidenceRate) **AND** the correlation coefficient of independent variables with other independent variables. This is to identify possible **multicollinearity** which is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. It can lead to skewed results so we need to carefully check for it.

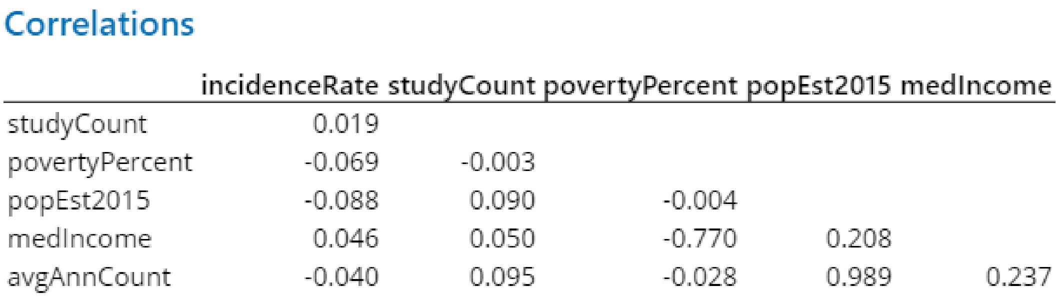
- Based on our result, we will identify which independent variables are highly correlated with our response variable and which ones cause high multicollinearity. We will iterate to keep only those that have a high correlation and cause minimum multicollinearity. Those variables will be used in regression analysis later.

Step 4 - Results:

- For our **first iteration** of correlation analysis, we selected all relevant continuous variables: studyCount, povertyPercent, medIncome, avgAnnCount and popEst2015. Using the ‘Correlation’ function on Minitab, we measured the Pearson Correlation coefficients for the different variables

- From the Correlations table below, we can look at all rows and column 1 to see correlations of each independent variable with Incidence rate. **We see that population**

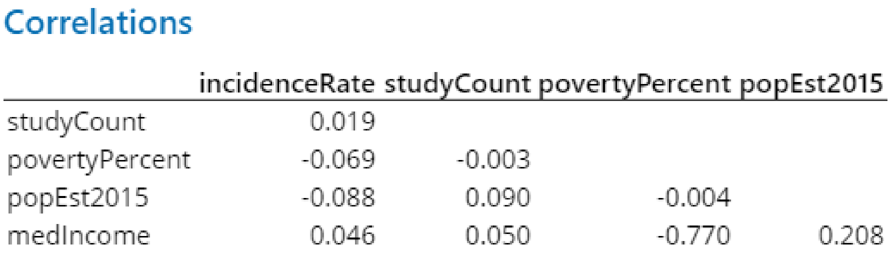
estimate has the highest absolute correlation with incidenceRate, followed by povertyPercent, median income, avgAnnCount, and studycount in that decreasing order.



Correlations

	incidenceRate	studyCount	povertyPercent	popEst2015	medIncome
studyCount	0.019				
povertyPercent	-0.069	-0.003			
popEst2015	-0.088	0.090	-0.004		
medIncome	0.046	0.050	-0.770	0.208	
avgAnnCount	-0.040	0.095	-0.028	0.989	0.237

- We also notice that the correlation coefficient between avgAnnCount and popEst2015 is quite higher (0.989) than our multicollinearity threshold of 0.7. This is unacceptable and we need to remove one of them.
- For our **second iteration**, we observe that popEst2015 has a greater absolute correlation with IncidenceRate than avgAnnCount, so we drop avgAnnCount and redo correlation analysis.



Correlations

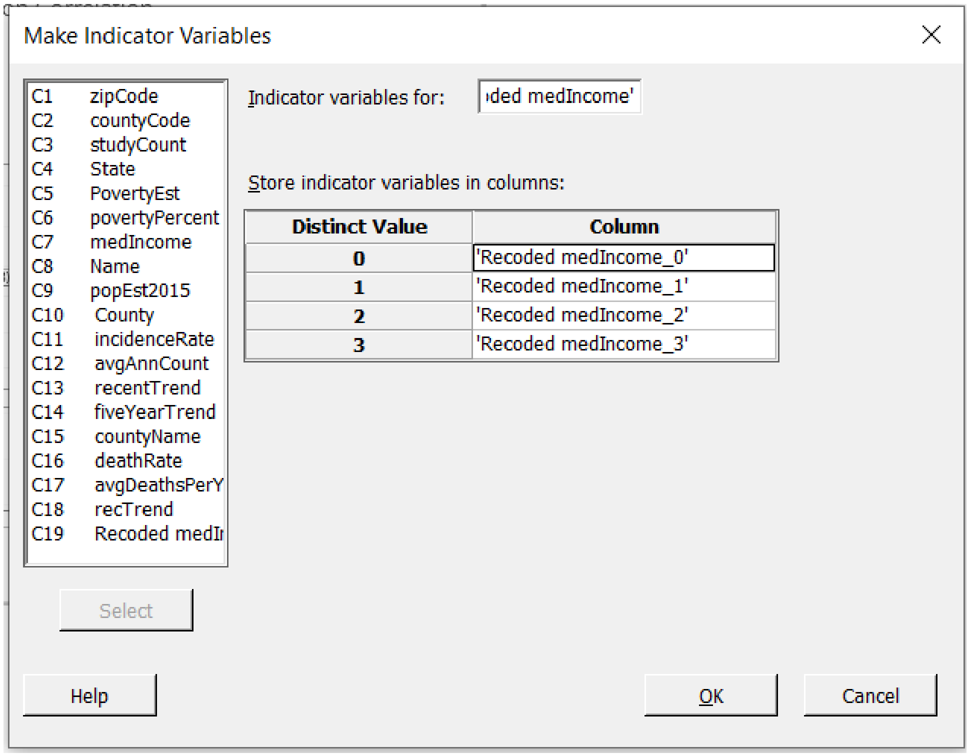
	incidenceRate	studyCount	povertyPercent	popEst2015
studyCount	0.019			
povertyPercent	-0.069	-0.003		
popEst2015	-0.088	0.090	-0.004	
medIncome	0.046	0.050	-0.770	0.208

- We now notice that medIncome and povertyPercent also have multicollinearity with correlation coefficient of -0.770. We may need to remove one of them after we see regression results. We are not removing just now since both have high correlation with incidence rate.

Step 5 - Method:

- As a preliminary step, we want to make sure that most of our dependent variables for the regression analysis are continuous. If any of them is categorical, it's better to recode them using *Dummy* or *Indicator* variables.

- Median income had 4 levels as recoded in step 3. We need k - 1 = 4 - 1 = 3 dummy variables to represent them: Recoded_medIncome_1, Recoded_medIncome_2, Recoded_medIncome_3.



Make Indicator Variables

Indicator variables for: ıded medIncome'

Store indicator variables in columns:

Distinct Value	Column
0	'Recoded medIncome_0'
1	'Recoded medIncome_1'
2	'Recoded medIncome_2'
3	'Recoded medIncome_3'

- Then we will run the **Regression using Incidence rate as our response variable, and the following as our continuous predictors**: studyCount, povertyPercent, Recoded_medIncome_1, Recoded_medIncome_2, Recoded_medIncome_3, and popEst2015.

- In addition to continuous predictors, we will include our **two important categorical variables:** State and recentTrend**.**

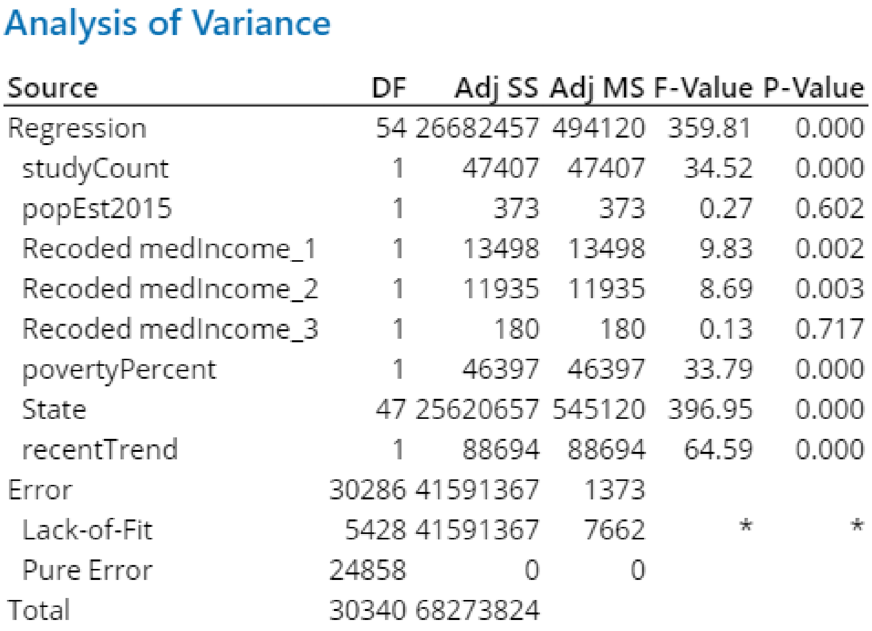
- In this regression, we will have results of the ANOVA for the hypothesis test with Null hypothesis - H(0): B1 = B2 = ….Bi = 0; and Alternative hypothesis - H(a): not all parameters are equal to 0.

- We will perform the following tests to check if the model assumptions are valid:
  - Normal probability plot of residuals to verify the assumption that the residuals are normally distributed.
  - Residuals versus fits plot to verify the assumption that the residuals are randomly distributed and have a constant variance.
  - Histogram of residuals to determine whether the data is skewed or if outliers exist in our data.

- ○ Residuals versus order plot to verify the assumption that the residuals are uncorrelated with each other.

Step 5 - Results:

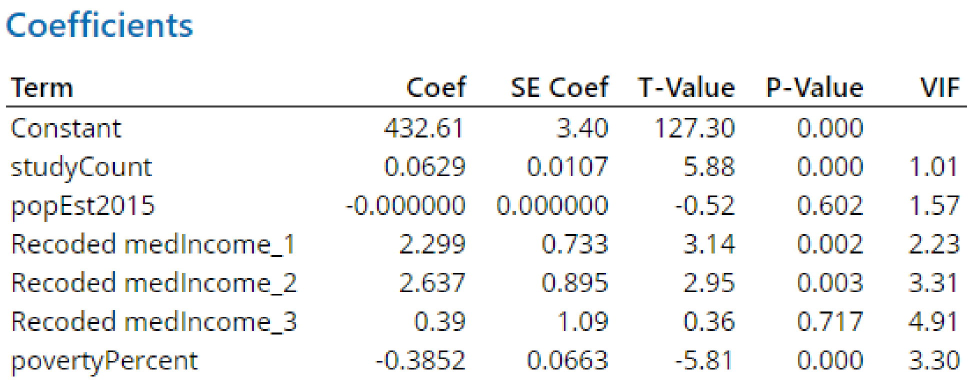
- The **first iteration** ANOVA table from our regression below shows that the regression p-value is 0.000 which means the Null hypothesis can be rejected. Not all parameters are equal to 0. This p-value based on the F test serves as the *overall significance* test of parameters.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	54	26682457	494120	359.81	0.000
studyCount	1	47407	47407	34.52	0.000
popEst2015	1	373	373	0.27	0.602
Recoded medIncome_1	1	13498	13498	9.83	0.002
Recoded medIncome_2	1	11935	11935	8.69	0.003
Recoded medIncome_3	1	180	180	0.13	0.717
povertyPercent	1	46397	46397	33.79	0.000
State	47	25620657	545120	396.95	0.000
recentTrend	1	88694	88694	64.59	0.000
Error	30286	41591367	1373		
Lack-of-Fit	5428	41591367	7662	*	*
Pure Error	24858	0	0		
Total	30340	68273824			

- A look at the **continuous predictor part** of the Coefficient table below shows how significant each independent variable is in determining our response variable. We can notice that the p-value for Recoded_median_income_3 is very high suggesting it is not significant. But we know this is a result of the multicollinearity with povertyPercent. Population estimate also has a high p-value which we will tackle if removing the first multicollinearity does not solve the problem.

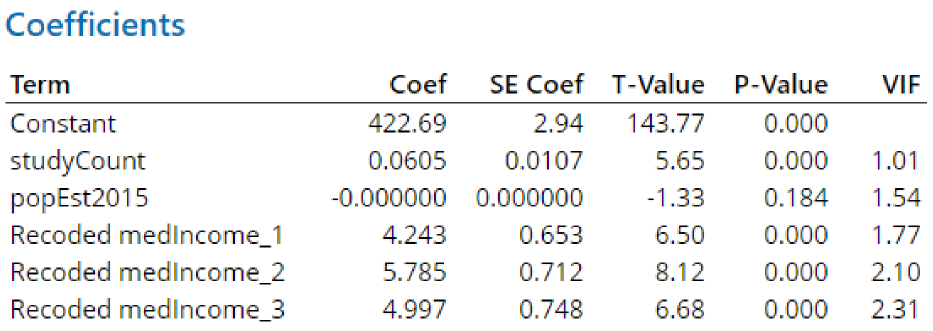


Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	432.61	3.40	127.30	0.000	
studyCount	0.0629	0.0107	5.88	0.000	1.01
popEst2015	-0.000000	0.000000	-0.52	0.602	1.57
Recoded medIncome_1	2.299	0.733	3.14	0.002	2.23
Recoded medIncome_2	2.637	0.895	2.95	0.003	3.31
Recoded medIncome_3	0.39	1.09	0.36	0.717	4.91
povertyPercent	-0.3852	0.0663	-5.81	0.000	3.30

- We can avoid multicollinearity by dropping povertyPercent from our predictors. We do not want to drop medianIncome as only the third variable has multicollinearity and that cannot be the basis for removal of all variables of medianIncome.

- Below are the results of the **second iteration** showing the continuous predictor part of coefficients table after dropping povertyPercent.



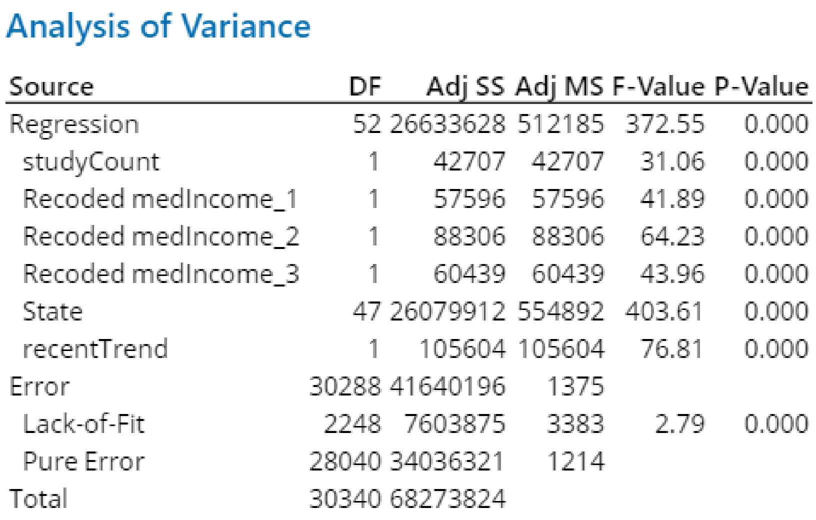
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	422.69	2.94	143.77	0.000	
studyCount	0.0605	0.0107	5.65	0.000	1.01
popEst2015	-0.000000	0.000000	-1.33	0.184	1.54
Recoded medIncome_1	4.243	0.653	6.50	0.000	1.77
Recoded medIncome_2	5.785	0.712	8.12	0.000	2.10
Recoded medIncome_3	4.997	0.748	6.68	0.000	2.31

- The p-value of popEst2015 is still greater than our alpha meaning it is insignificant. The coefficients are also zeros so in the next iteration, we drop popEst2015. Although it has a decent correlation, it adds nothing to our regression model.

- Adjusted R-squared at the end of iteration 2 is 38.91%.

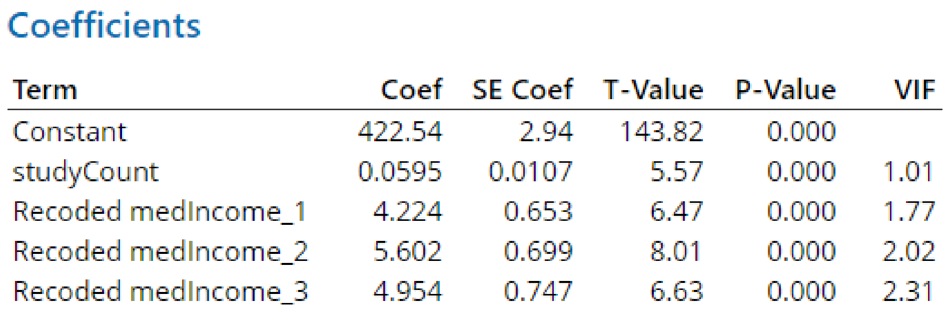
- For our **third and last iteration**, we dropped popEst2015 and the results are below:



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	52	26633628	512185	372.55	0.000
studyCount	1	42707	42707	31.06	0.000
Recoded medIncome_1	1	57596	57596	41.89	0.000
Recoded medIncome_2	1	88306	88306	64.23	0.000
Recoded medIncome_3	1	60439	60439	43.96	0.000
State	47	26079912	554892	403.61	0.000
recentTrend	1	105604	105604	76.81	0.000
Error	30288	41640196	1375		
Lack-of-Fit	2248	7603875	3383	2.79	0.000
Pure Error	28040	34036321	1214		
Total	30340	68273824			

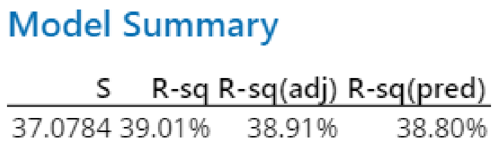
- The p-value of regression is 0.000 so the model is significant.



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	422.54	2.94	143.82	0.000	
studyCount	0.0595	0.0107	5.57	0.000	1.01
Recoded medIncome_1	4.224	0.653	6.47	0.000	1.77
Recoded medIncome_2	5.602	0.699	8.01	0.000	2.02
Recoded medIncome_3	4.954	0.747	6.63	0.000	2.31

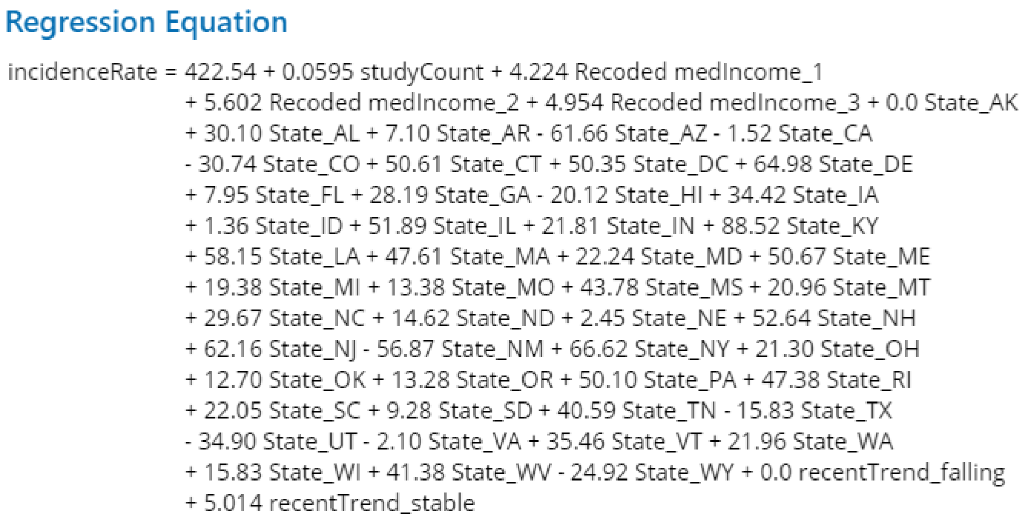
- Continuous predictor part of the coefficient table now shows that all have 0.000 p-value - all are significant. Full coefficients table can be found in appendix.
- The model summary below shows that the Adjusted R-squared value has remained the same while getting rid of an unnecessary variable.



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
37.0784	39.01%	38.91%	38.80%

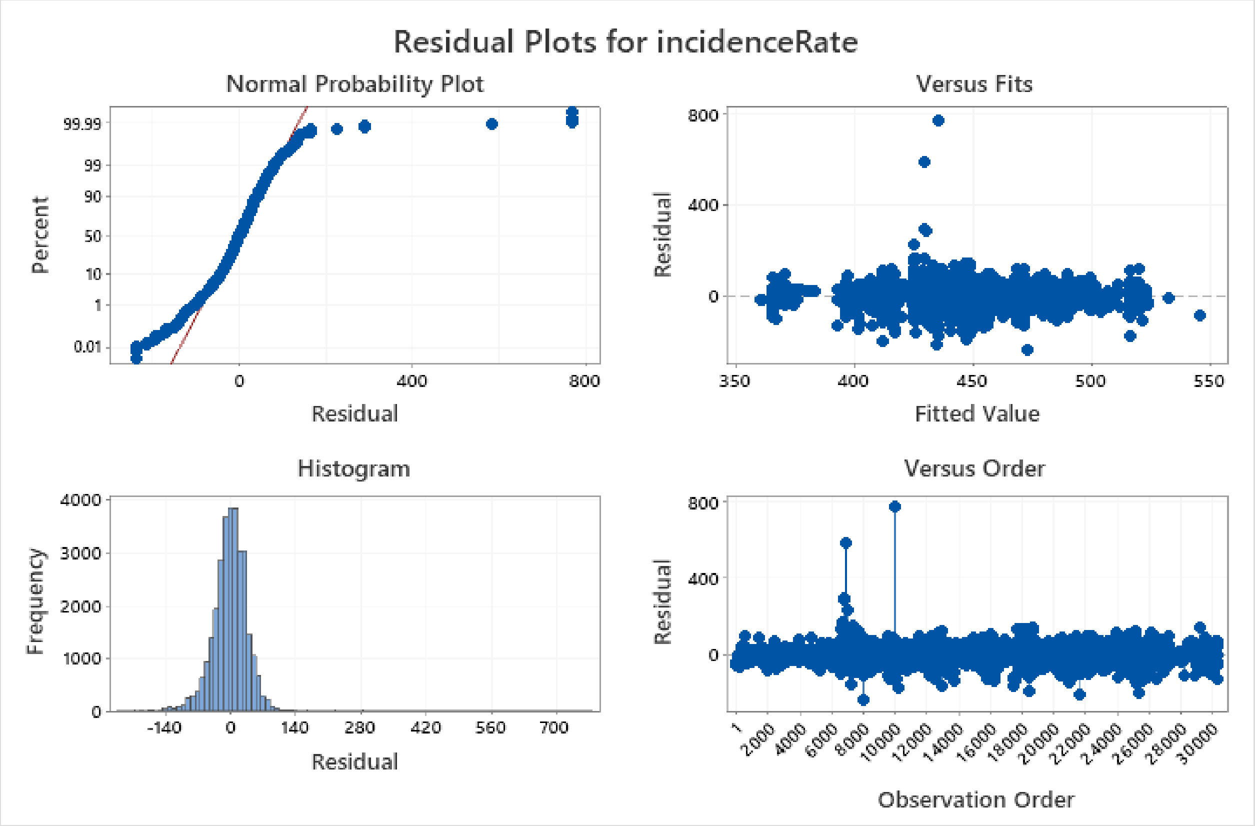
- **Conclusion:** We conclude that 38.91% of the total sum of squares can be explained by our model. In other words, our model explains 38.91% of variability of incidence Rates of cancer.
- Below is the regression equation which is complex due to categorical variables:



Regression Equation

```
incidenceRate = 422.54 + 0.0595 studyCount + 4.224 Recoded medIncome_1
                + 5.602 Recoded medIncome_2 + 4.954 Recoded medIncome_3 + 0.0 State_AK
                + 30.10 State_AL + 7.10 State_AR - 61.66 State_AZ - 1.52 State_CA
                - 30.74 State_CO + 50.61 State_CT + 50.35 State_DC + 64.98 State_DE
                + 7.95 State_FL + 28.19 State_GA - 20.12 State_HI + 34.42 State_IA
                + 1.36 State_ID + 51.89 State_IL + 21.81 State_IN + 88.52 State_KY
                + 58.15 State_LA + 47.61 State_MA + 22.24 State_MD + 50.67 State_ME
                + 19.38 State_MI + 13.38 State_MO + 43.78 State_MS + 20.96 State_MT
                + 29.67 State_NC + 14.62 State_ND + 2.45 State_NE + 52.64 State_NH
                + 62.16 State_NJ - 56.87 State_NM + 66.62 State_NY + 21.30 State_OH
                + 12.70 State_OK + 13.28 State_OR + 50.10 State_PA + 47.38 State_RI
                + 22.05 State_SC + 9.28 State_SD + 40.59 State_TN - 15.83 State_TX
                - 34.90 State_UT - 2.10 State_VA + 35.46 State_VT + 21.96 State_WA
                + 15.83 State_WI + 41.38 State_WV - 24.92 State_WY + 0.0 recentTrend_falling
                + 5.014 recentTrend_stable
```

- We observe from the equation that studyCount, medianIncome, and Stable trend all are **positively** impacting incidence rate – higher these predictors, higher the value of incidence rate.
- Following results were obtained from the tests for model assumptions:

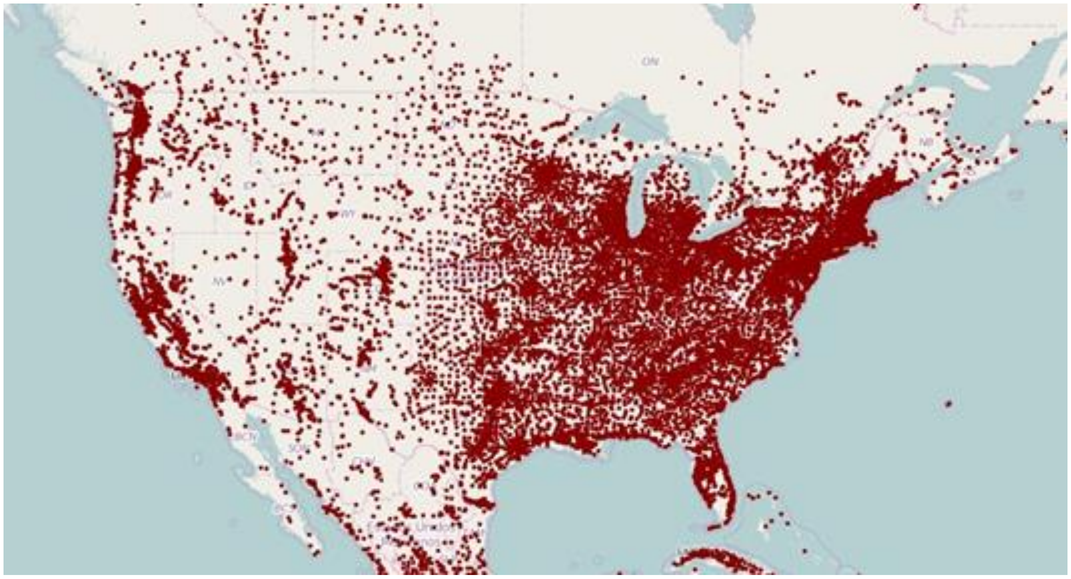


- We make the following observations:
  - Although there are values not on the line, our Normal probability plot *approximately* follows a straight line.
  - Residual versus fits plot has random distribution about 0 with no recognizable pattern.
  - Histogram of residuals is not skewed and does not have outliers.
  - Residual versus order shows no trends or patterns.
- With the above 4 results, we conclude that our model meets the assumptions.

Project Findings:

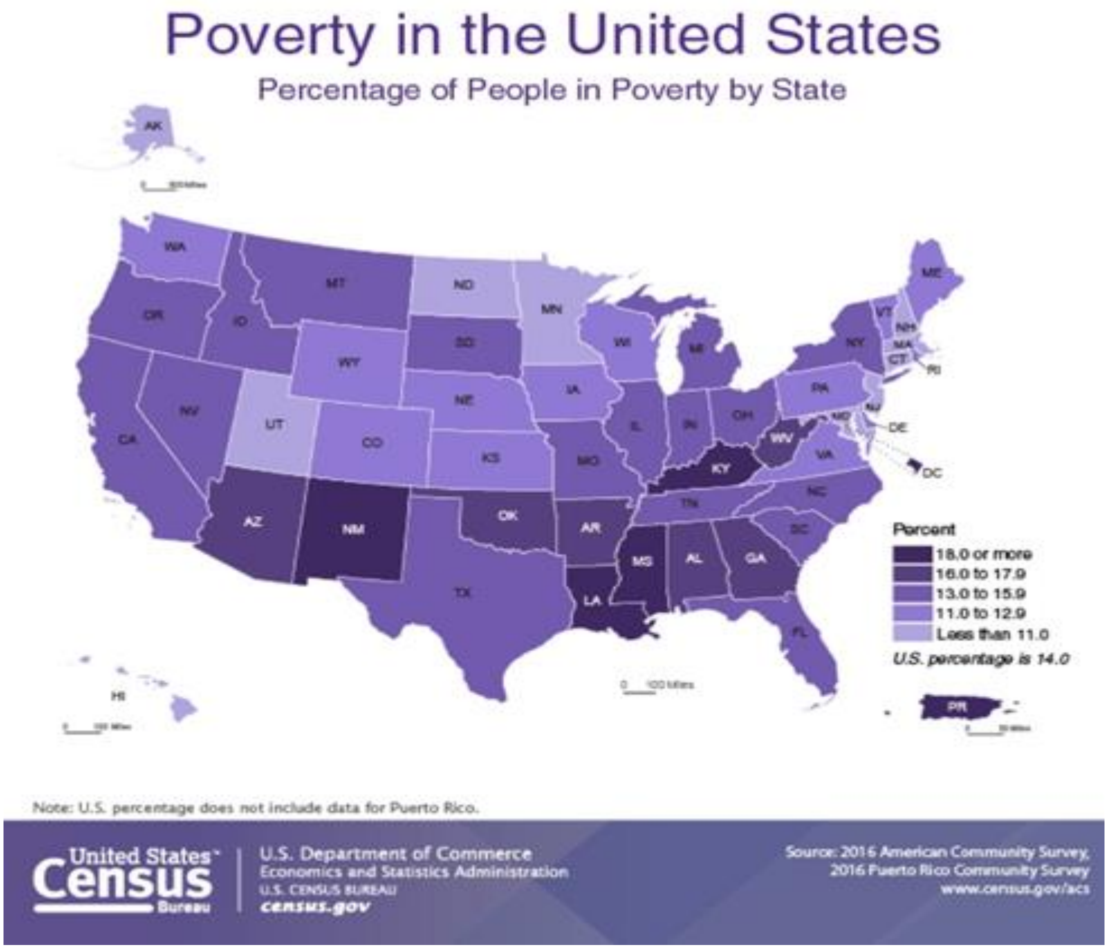
- According to the statistics in step 1, Kentucky is the most prone to cancer with an incidence rate of 517.33 people per 100,000. The next highest incidence rates are Delaware, New York, New Jersey, and Louisiana with 498.19, 497.52, 494.01, and 486.81 incidents per 100,000 respectively. On top of the top 5 mentioned, the top 10 have all been mentioned in Step 1 - Results to showcase which states/regions have the biggest incidence rates across the nation.

  - The top 3 reasons for Kentucky having the highest incidence rate based on other researches is because of smoking, obesity, and poverty. Additionally, we believe there is a positive linear relationship between population density and incidence rate, as the highest 10 incidence rates are located on the East Coast, in regions with high population density.



- Based on our results in 2(a), we can conclude that the recent trend is associated with the incidence rate of cancer. When looking at the “Stable” recent incidence trend, it looks like it leads to a higher cancer incidence rate, compared to a falling trend. Incidence Rate is defined as the ratio of the number of cases to the total time the population is at risk of disease [1]. Given that time plays a role in measuring incidence rate, it is safe to conclude that the statistical results are in-line with reality, as a steady number of cases for cancer in a given period of time is bound to result in an increased cancer incidence rate.
- We also conclude that the study count is associated with the incidence rate. Counties with a higher study count had higher incidence rates. This seems to be attributed to the fact that counties who have a higher population tend to have the opportunity for more cases.
- The results from our One-Way ANOVA test tell us that belonging to certain income groups can have an impact on cancer incidence rate. We believe that factors such as quality healthcare, access to higher quality ingredients are not available to lower income groups, and the added stress of financial problems coupled with a lack of education regarding unhealthy habits contribute to higher incidence rates at lower income levels. Cancer found in those with a lower income is often found at a later stage, making it more

difficult to cure. This result will help us transition into our Regression Analysis since we are able to identify which groups of Median income we can use to analyze their relationship with other factors affecting Cancer Incidence Rate.


Poverty in the United States
Percentage of People in Poverty by State
AK
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
KY
TN
MS
AL
GA
OH
WV
VA
NC
SC
FL
PA
NY
VT
NH
ME
MA
CT
RI
NJ
DE
MD
DC
HI
PR
Percent
18.0 or more
16.0 to 17.9
13.0 to 15.9
11.0 to 12.9
Less than 11.0
U.S. percentage is 14.0
Note: U.S. percentage does not include data for Puerto Rico.
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov
Source: 2016 American Community Survey, 2016 Puerto Rico Community Survey
www.census.gov/acs

- Our correlation analysis showed us that population size estimate has the highest association with incidence rate, followed by poverty, median income, and studycount in that decreasing order. We also observed that median income and poverty percentage in a county are interlinked as variables.
- Lastly, our regression analysis showed that each of the factors under consideration are indeed associated with the incidence rate. However, population size adds little to no value to our model hence got dropped. Poverty percent is interlinked with median income so it added lesser value and was removed.
- Our regression results could explain about 39% of the variation in incidence rate. Meaning of all the changes that take place in cancer incidence rate, our model can explain 39% of them which is valuable.

Discussion:

Because our model could explain only 39% of variation is evidence that there are other factors that are more directly related to incidence of cancer like exposure to UV radiation, smoking, obesity, etc. As an extension of this project, those other factors can also be studied and statistical procedures applied to determine their impact on incidence rate to better fit a model.

Appendix:

Definitions:

Incidence rate - the ratio of the number of cases to the total time the population is at risk of disease. A measure of the frequency with which some event occurs over a specified period of time. In our case, incidence rates are in the context of Cancer.

Average Annual Count - 2009-2013 mean incidences per county.

Recent Trend - a change or development towards something new or different. In our context, it is the recent trend in incidences of cancer.

Recent Trend - Falling Rate - a decrease in incidences to a lower level over a period of time.

Stable Rate - little to no change in incidence rate of cancer over a period of time.

Mean - the average value of a dataset.

Median Income - the point that divides a population, by income, into two groups.

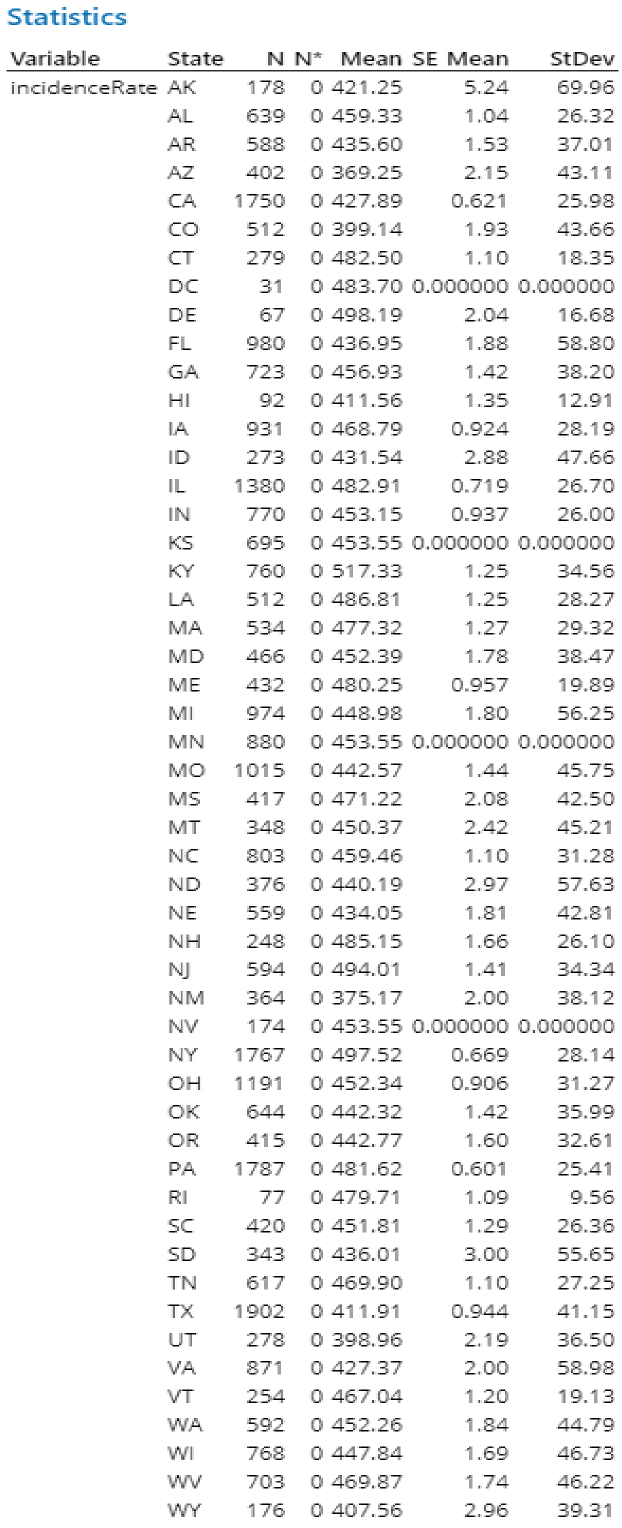
Population Estimate - calculation of the size of a population for a year between census periods, or for the current year. In our context, year 2015.

Poverty Percentage (or poverty rate) - the ratio of the number of people whose income falls below the poverty line. The poverty line is the estimated minimum level of income needed to secure the necessities to live.

State - a territory considered an organized political community under one government.

Study Count - the number of cancer clinical trials conducted.

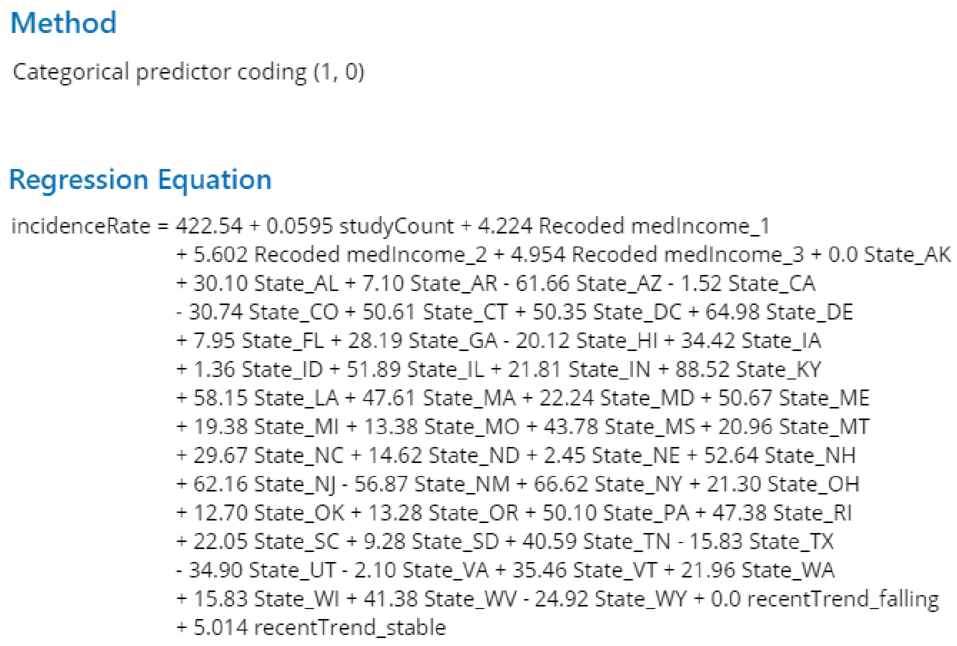
Part 1 – full list of States in descending order of incidence:



Statistics

Variable	State	N	N*	Mean	SE Mean	StDev
incidenceRate	AK	178	0	421.25	5.24	69.96
	AL	639	0	459.33	1.04	26.32
	AR	588	0	435.60	1.53	37.01
	AZ	402	0	369.25	2.15	43.11
	CA	1750	0	427.89	0.621	25.98
	CO	512	0	399.14	1.93	43.66
	CT	279	0	482.50	1.10	18.35
	DC	31	0	483.70	0.000000	0.000000
	DE	67	0	498.19	2.04	16.68
	FL	980	0	436.95	1.88	58.80
	GA	723	0	456.93	1.42	38.20
	HI	92	0	411.56	1.35	12.91
	IA	931	0	468.79	0.924	28.19
	ID	273	0	431.54	2.88	47.66
	IL	1380	0	482.91	0.719	26.70
	IN	770	0	453.15	0.937	26.00
	KS	695	0	453.55	0.000000	0.000000
	KY	760	0	517.33	1.25	34.56
	LA	512	0	486.81	1.25	28.27
	MA	534	0	477.32	1.27	29.32
	MD	466	0	452.39	1.78	38.47
	ME	432	0	480.25	0.957	19.89
	MI	974	0	448.98	1.80	56.25
	MN	880	0	453.55	0.000000	0.000000
	MO	1015	0	442.57	1.44	45.75
	MS	417	0	471.22	2.08	42.50
	MT	348	0	450.37	2.42	45.21
	NC	803	0	459.46	1.10	31.28
	ND	376	0	440.19	2.97	57.63
	NE	559	0	434.05	1.81	42.81
	NH	248	0	485.15	1.66	26.10
	NJ	594	0	494.01	1.41	34.34
	NM	364	0	375.17	2.00	38.12
	NV	174	0	453.55	0.000000	0.000000
	NY	1767	0	497.52	0.669	28.14
	OH	1191	0	452.34	0.906	31.27
	OK	644	0	442.32	1.42	35.99
	OR	415	0	442.77	1.60	32.61
	PA	1787	0	481.62	0.601	25.41
	RI	77	0	479.71	1.09	9.56
	SC	420	0	451.81	1.29	26.36
	SD	343	0	436.01	3.00	55.65
	TN	617	0	469.90	1.10	27.25
	TX	1902	0	411.91	0.944	41.15
	UT	278	0	398.96	2.19	36.50
	VA	871	0	427.37	2.00	58.98
	VT	254	0	467.04	1.20	19.13
	WA	592	0	452.26	1.84	44.79
	WI	768	0	447.84	1.69	46.73
	WV	703	0	469.87	1.74	46.22
	WY	176	0	407.56	2.96	39.31

Part 5: final regression analysis results (Full):

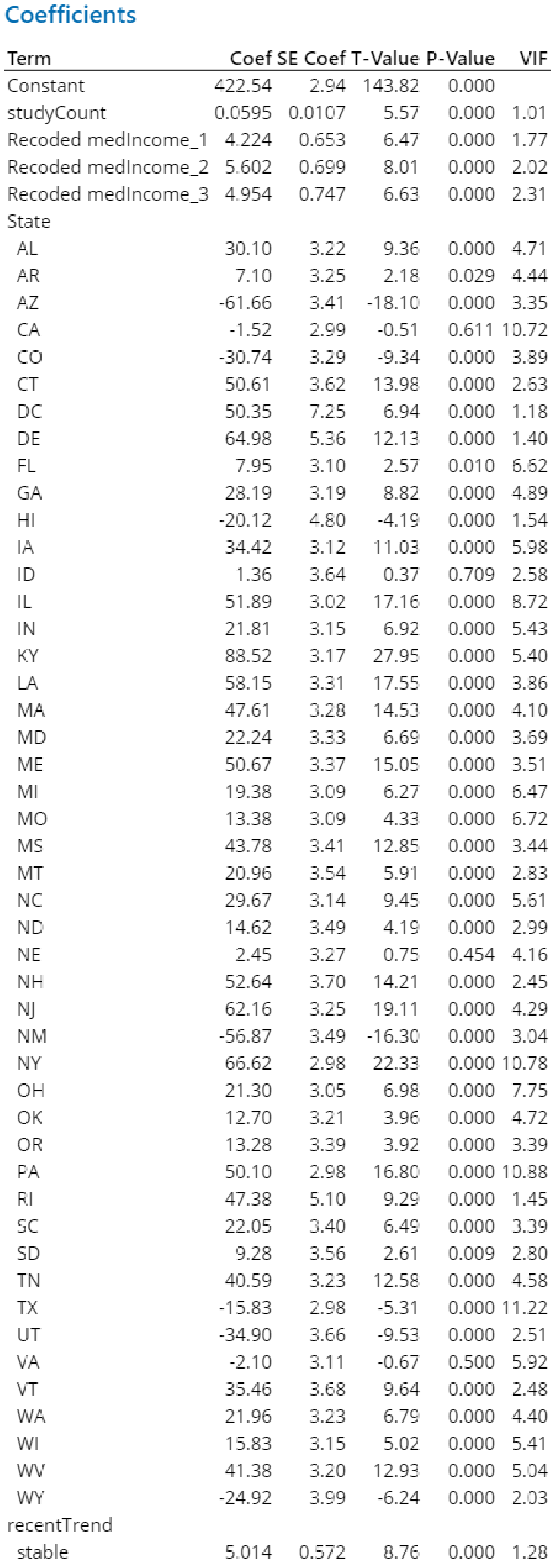


Method

Categorical predictor coding (1, 0)

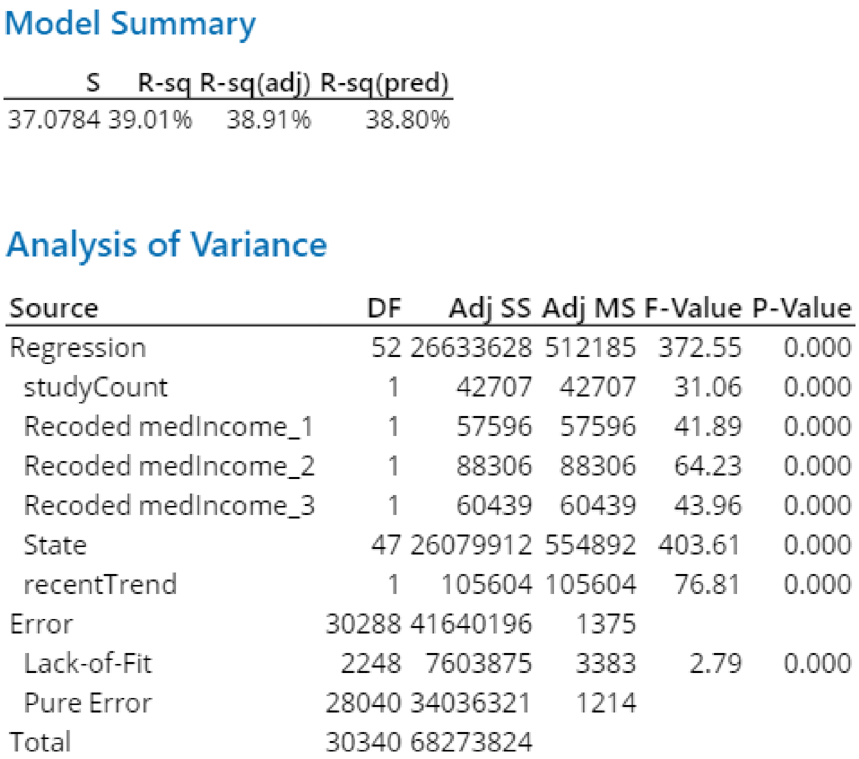
Regression Equation

```
incidenceRate = 422.54 + 0.0595 studyCount + 4.224 Recoded medIncome_1
                + 5.602 Recoded medIncome_2 + 4.954 Recoded medIncome_3 + 0.0 State_AK
                + 30.10 State_AL + 7.10 State_AR - 61.66 State_AZ - 1.52 State_CA
                - 30.74 State_CO + 50.61 State_CT + 50.35 State_DC + 64.98 State_DE
                + 7.95 State_FL + 28.19 State_GA - 20.12 State_HI + 34.42 State_IA
                + 1.36 State_ID + 51.89 State_IL + 21.81 State_IN + 88.52 State_KY
                + 58.15 State_LA + 47.61 State_MA + 22.24 State_MD + 50.67 State_ME
                + 19.38 State_MI + 13.38 State_MO + 43.78 State_MS + 20.96 State_MT
                + 29.67 State_NC + 14.62 State_ND + 2.45 State_NE + 52.64 State_NH
                + 62.16 State_NJ - 56.87 State_NM + 66.62 State_NY + 21.30 State_OH
                + 12.70 State_OK + 13.28 State_OR + 50.10 State_PA + 47.38 State_RI
                + 22.05 State_SC + 9.28 State_SD + 40.59 State_TN - 15.83 State_TX
                - 34.90 State_UT - 2.10 State_VA + 35.46 State_VT + 21.96 State_WA
                + 15.83 State_WI + 41.38 State_WV - 24.92 State_WY + 0.0 recentTrend_falling
                + 5.014 recentTrend_stable
```



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	422.54	2.94	143.82	0.000	
studyCount	0.0595	0.0107	5.57	0.000	1.01
Recoded medIncome_1	4.224	0.653	6.47	0.000	1.77
Recoded medIncome_2	5.602	0.699	8.01	0.000	2.02
Recoded medIncome_3	4.954	0.747	6.63	0.000	2.31
State					
AL	30.10	3.22	9.36	0.000	4.71
AR	7.10	3.25	2.18	0.029	4.44
AZ	-61.66	3.41	-18.10	0.000	3.35
CA	-1.52	2.99	-0.51	0.611	10.72
CO	-30.74	3.29	-9.34	0.000	3.89
CT	50.61	3.62	13.98	0.000	2.63
DC	50.35	7.25	6.94	0.000	1.18
DE	64.98	5.36	12.13	0.000	1.40
FL	7.95	3.10	2.57	0.010	6.62
GA	28.19	3.19	8.82	0.000	4.89
HI	-20.12	4.80	-4.19	0.000	1.54
IA	34.42	3.12	11.03	0.000	5.98
ID	1.36	3.64	0.37	0.709	2.58
IL	51.89	3.02	17.16	0.000	8.72
IN	21.81	3.15	6.92	0.000	5.43
KY	88.52	3.17	27.95	0.000	5.40
LA	58.15	3.31	17.55	0.000	3.86
MA	47.61	3.28	14.53	0.000	4.10
MD	22.24	3.33	6.69	0.000	3.69
ME	50.67	3.37	15.05	0.000	3.51
MI	19.38	3.09	6.27	0.000	6.47
MO	13.38	3.09	4.33	0.000	6.72
MS	43.78	3.41	12.85	0.000	3.44
MT	20.96	3.54	5.91	0.000	2.83
NC	29.67	3.14	9.45	0.000	5.61
ND	14.62	3.49	4.19	0.000	2.99
NE	2.45	3.27	0.75	0.454	4.16
NH	52.64	3.70	14.21	0.000	2.45
NJ	62.16	3.25	19.11	0.000	4.29
NM	-56.87	3.49	-16.30	0.000	3.04
NY	66.62	2.98	22.33	0.000	10.78
OH	21.30	3.05	6.98	0.000	7.75
OK	12.70	3.21	3.96	0.000	4.72
OR	13.28	3.39	3.92	0.000	3.39
PA	50.10	2.98	16.80	0.000	10.88
RI	47.38	5.10	9.29	0.000	1.45
SC	22.05	3.40	6.49	0.000	3.39
SD	9.28	3.56	2.61	0.009	2.80
TN	40.59	3.23	12.58	0.000	4.58
TX	-15.83	2.98	-5.31	0.000	11.22
UT	-34.90	3.66	-9.53	0.000	2.51
VA	-2.10	3.11	-0.67	0.500	5.92
VT	35.46	3.68	9.64	0.000	2.48
WA	21.96	3.23	6.79	0.000	4.40
WI	15.83	3.15	5.02	0.000	5.41
WV	41.38	3.20	12.93	0.000	5.04
WY	-24.92	3.99	-6.24	0.000	2.03
recentTrend					
stable	5.014	0.572	8.76	0.000	1.28

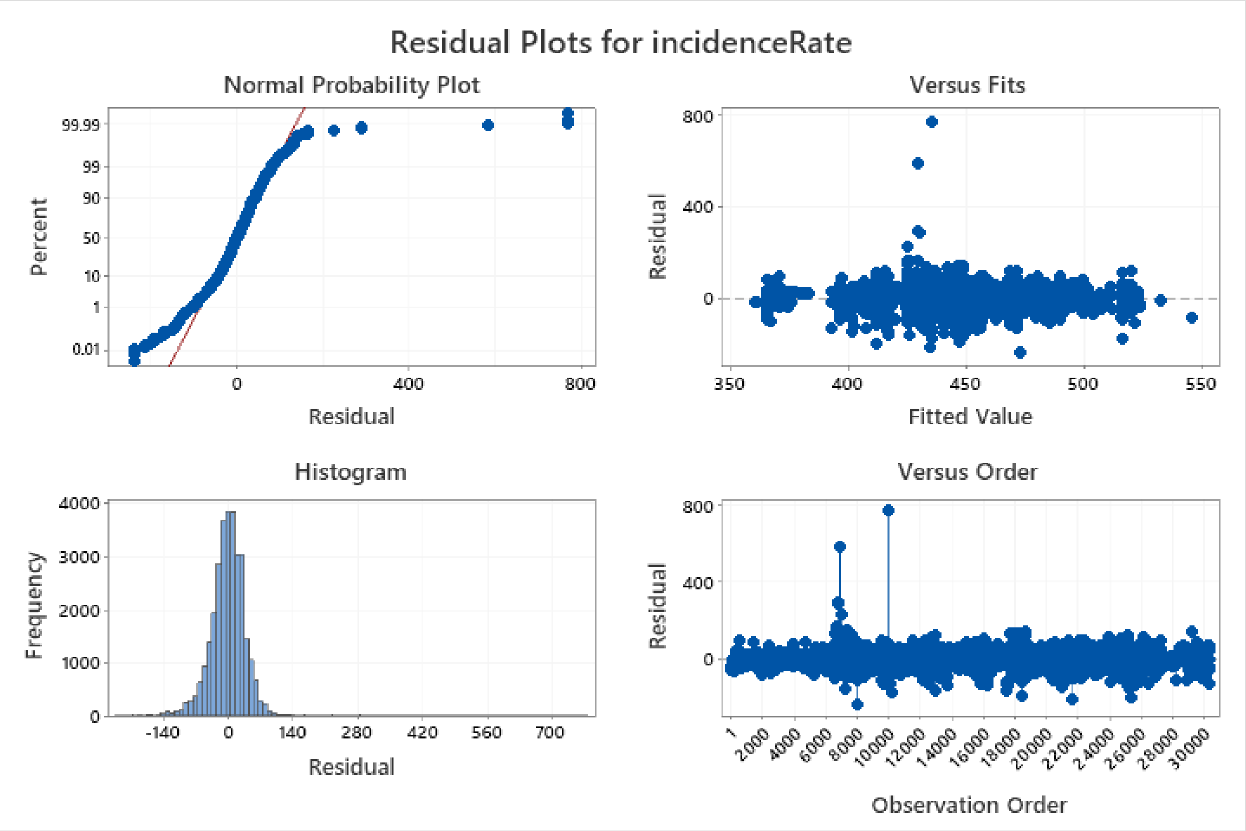


Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
37.0784	39.01%	38.91%	38.80%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	52	26633628	512185	372.55	0.000
studyCount	1	42707	42707	31.06	0.000
Recoded medIncome_1	1	57596	57596	41.89	0.000
Recoded medIncome_2	1	88306	88306	64.23	0.000
Recoded medIncome_3	1	60439	60439	43.96	0.000
State	47	26079912	554892	403.61	0.000
recentTrend	1	105604	105604	76.81	0.000
Error	30288	41640196	1375		
Lack-of-Fit	2248	7603875	3383	2.79	0.000
Pure Error	28040	34036321	1214		
Total	30340	68273824			



Works cited:

www.cancerstatisticscenter.org

www.health.ny.gov

www.cdc.gov

www.cancer.org