OCTOBER 10, 2021

Multi linear regression analysis of Cancer Incidence Rate

Group assignment 1

1. Usama Tariq
2. Jordan Miller
3. Leon Parfenov
4. Rex Onwubuche
5. Qudus Abiola Lawal

Contents

Introduction:

Cancer is one of the major public health problems plaguing people not just in the US but worldwide. Although it has decreased in recent times due to new and improved prevention, detection and treating measures it is still the second leading cause of mortality in the US after Heart disease and a major proportion of cancers could still be prevented. Besides genetics, newly diagnosed cancers in the US are caused by a combination of excess body weight, physical inactivity, excess alcohol consumption, and poor nutrition and are potentially avoidable, this includes all the cancers that are caused by smoking. With access to the right information and programs regarding cancer detection and prevention the incidence rate of cancer can be reduced.

In this report we aim to highlight the regions of the country with the highest incidence rates. Moreover, we want to determine what factors contribute to the incidence rate of cancer and using them how can we predict it in order to identify regions and associated partners for cancer interventions across the US.

Methodology:

Data:

We took publicly available State and National level government data which was aggregated county wise. The county codes, poverty percentage and estimate, median income and population data was taken from the US Census Bureau. While data about the incidence rate, average annual count of cancers, recent and five-year trend of incidence rate, cancer death rate, average deaths per year and recent trend of deaths by cancer was obtained from the State Cancer Profiles government website.

Analysis:

We aggregated the data county wise and cleaned it. After which we analyzed the incidence rate to identify which regions of the country where most prone to cancer. We identified the regions at both the county and state level. We also compared the incidence rate with the median income, by dividing the median income in four categories of Very Low, Low, High and Very High. In addition to this we also identified the counties which had the highest and lowest median income and incidence rate in each of the four categories. Lastly, to predict incidence rate we identified the factors which would most likely effect it and developed a statistical regression model using them.

Regions most prone to cancer:

Cancer is caused by genetic changes which leads to uncontrollable tumor formation. However, a minority of cancers are due to genetic or hereditary genetic mutations. Most cancers are caused by

external factors like environmental influences, lifestyle choices, neglect of health, viruses, bacteria, and radiation. All these factors are deeply influenced by the region where a person lives. As certain regions on top of having shown a greater inclination towards an unhealthy lifestyle, also have a long history of obesity, diabetes and other conditions which lead to cancers.

To determine which region was most prone to cancer we took incidence rate data and filtered it from region wise to state wise and finally county wise. Calculating the average of the incidence rate we found that top ten counties where cancer was most prevalent were all located in the **Southern States**. The highest being in the **Union County of Florida** with an incidence rate of **1206.9** (age adjusted incidences of all cancers per 100,000).

Region	**State**	**County Name**	**Incidence Rate**
South	Florida	Union County	1206.9
South	Virginia	Williamsburg City	1014.2
South	Virginia	Charlottesville City	718.9
South	Virginia	Petersburg City	651.3
South	Kentucky	Bracken County	639.7
South	Kentucky	Powell County	630.4
South	Virginia	Waynesboro City	596.9
South	Virginia	Harrisonburg City	591.1
South	West Virginia	Clay County	591

Table 1 Top Ten Counties with the highest Incidence Rate

In terms of States, we discovered **Kentucky** to have the highest average incidence rate of **513** among all. The breakdown of incidence rate State and Region wise is depicted in these graphs.

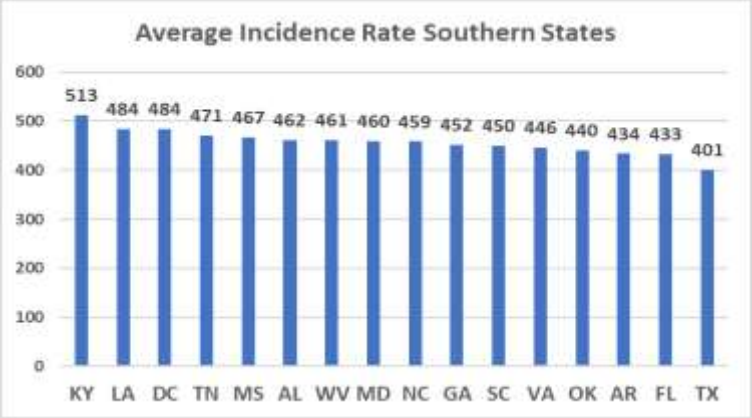

Figure 1 Incidence Rate of Southern States

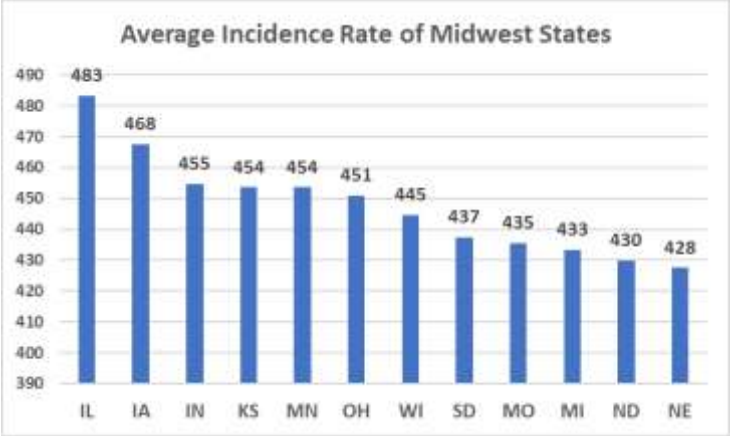

Figure 2 Incidence Rate of Midwestern States

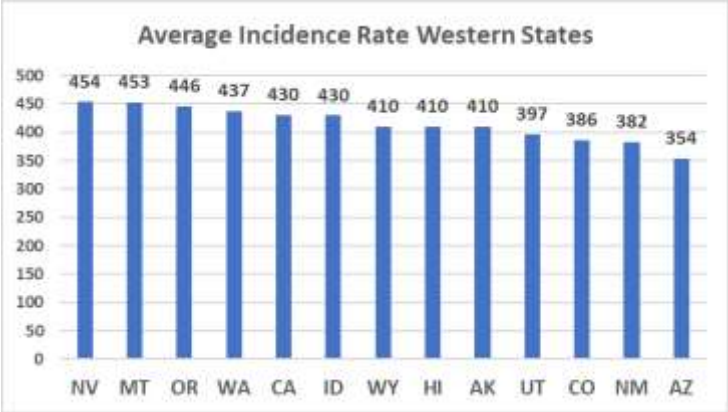

Figure 3 Incidence Rate of Western States

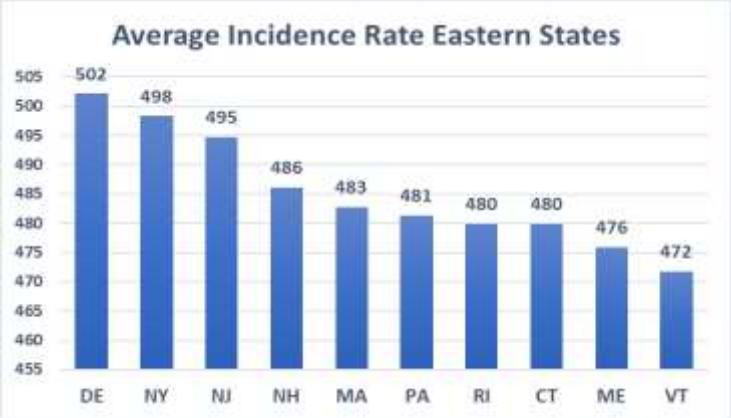

Figure 4 Incidence Rate of Eastern States

Interpretation:

As per our results the Southern States and their counties are most prone to cancer in the country. These results make sense when looked in the context that Southern States have some of the highest rates of diabetes, obesity, smoking and other such factors which cause cancer.

Comparison of Median Income to Incidence Rate:

The median income variable was separated into 4 different indicator levels in order to demonstrate the variability of incident rates between clustered segments of Income. The LOW segment is an annual median income of less than $30,000. The MEDIUM is an annual median income between $30,000 and $60,000. The HIGH median income is an annual income between $60,000 and $90,000. Finally, those with an annual median income greater than $90,000 were placed in the EXTREME group. After the data was clustered together, the average incident rate was taken in each segment to compare the relationship between income levels and cancer incident rates.

Median Income	Average of Incidence Rate
LOW	460.0537879
MEDIUM	453.2954693
HIGH	456.4101677
EXTREME	436.504717

Table 2 Median Income Categories and their associated Incidence Rates

Interpretation:

It is immediately clear that the EXTREME segment has disproportionate rate of cancer incidence, where LOW, MEDIUM, and HIGH segments have relatively equal averaged values (see Table 2). Thus, the highest income group corresponds with the lowest Incident rate. This can be due to underlying financial factors such as: living in a cleaner area with less harmful pollutants, access to more products and services that reduce the risk of cancer, and generally healthier lifestyles (from food consumption to UV exposure protection).

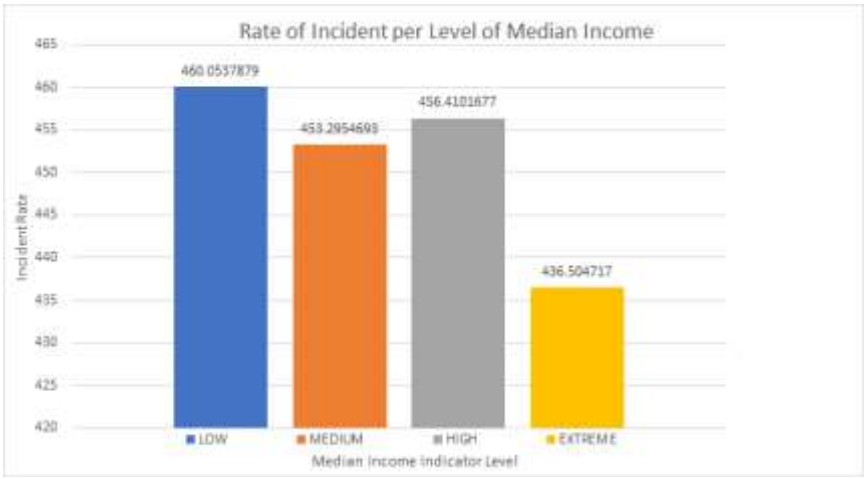

Figure 5 Bar Chart Average Incidence Rate of each Median Income Category

The MEDIUM average of incidence rate is lower than the HIGH rate, and while it is just slightly lower, it is enough to diminish the linearity of the model. However, even with the slight bump in the graph, a negative linear trend line was implemented to better highlight the general downward slope which proves an inverse relation between average incidence rate and the median income group. (See Figure 6). There is a distinct downward trend.

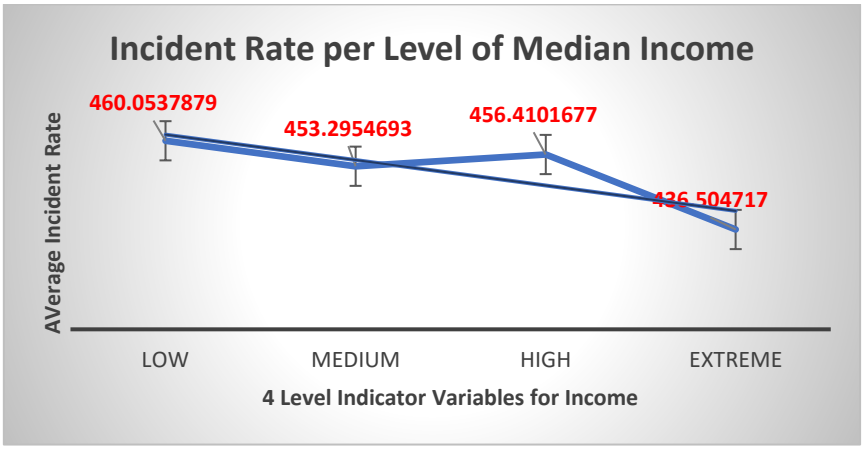

Figure 6 Line Chart Average Incidence Rate Vs Median Income Categories

Incidence Rates with Focus on Median Income:

In addition to viewing the categories, we also identified and highlighted the highest and lowest Median Income and their associated Incidence Rate in each category

Low Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Holmes County, Mississippi	44	Very Low	22640	554.1	18340	239.7
Highest	Rooks County, Kansas	12	Very Low	48383	453.5	5174	164.7

Medium Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Gates County, North Carolina	15.6	Low	48413	370.9	11431	171.5
Highest	Mercer County, North Dakota	7.3	Low	74047	482	8853	151.9

High Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	St. Charles County, Missouri	6.8	High	74220	465.2	385590	162.9
Highest	Nassau County, New York	6.7	High	98312	514.3	1361350	145.8

Extreme Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Somerset County, New Jersey	5	Very High	100194	471	333654	153.3
Highest	Falls Church City, Virginia	3.2	Very High	125635	447.7	13892	137.6

Incidence Rates with Focus on Median Income:

Similarly, we also identified the highest and lowest Incidence Rate and their associated Median Income in each category

Low Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Presidio County, Texas	21.8	Low	34258	211.1	6876	66.3
Highest	Union County, Florida	24.3	Low	40207	1206.9	15234	362.8

Medium Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Aleutians West Census Area, Alaska	9.9	Medium	68387	201.3	5702	203.3
Highest	Sherman County, Oregon	14.2	Medium	53277	587.4	1680	180.4

High Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	McKenzie County, North Dakota	8.6	High	81209	254.7	12826	124.6
Highest	Nantucket County, Massachusetts	7.1	High	82596	572.8	10925	146.9

Extreme Category

	Location	Poverty Percentage	Median Income Interval	Median Income	Incidence Rate	Population	Death Rate
Lowest	Loudoun County, Virginia	3.9	Very High	122641	364.9	375629	136.5
Highest	Hunterdon County, New Jersey	4.7	Very High	103876	496.3	125488	145.1

Correlation with predictor variables:

Correlation analysis measures how two variables are related, The correlation coefficient (r) is a statistic that tells you the strength and direction of that relationship, it will be expressed as a positive or negative number between -1 and one. The value of the number indicates the strength of the relationship:

- r = 0 means there is no correlation
- r = 1 means there is a perfect positive correlation
- r = -1 means there is a perfect negative correlation

The sign of the correlation coefficient indicates whether the direction of the relationship is positive (direct) or negative (inverse). A direct relationship will see the variables increasing and decreasing together, whereas an inverse relationship will see one variable increasing while the other decreases.

Analysis:

After aggregating the data, we used the correlation analysis data tool to understand which variables had the strongest correlation with one another and the results were both conclusive and surprising. We found that the strongest correlation with incidence rate was death rate with an r of 0.44 which would make sense seeing that the more incidents that occur the more deaths would occur as well. A correlation that we found curious was incidence rate and population estimate, with an r of 0.023 it would suggest that there is little to no correlation at all, which wouldn’t seem to make sense as with more people you would assume there would be more incidents. Another shocking correlation was

incidence rate and poverty percent as it also had relatively no correlation with an r of 0.008, however we concluded that with regardless of your income, you can still develop cancer and income wouldn't relatively play a role in that. Two of the strongest correlations that were seen not involving incidence rate were related to median income; median income and poverty percent had a very strong negative correlation with r = -0.76, while median income and death rate also had a relatively strong negative correlation with r= -0.48.

	countyCode	povertyPercent	PovertyEst	medIncome	popEst2015	avgAnnCount	fiveYearTrend	deathRate	avgDeathsPerYear	studyCount	incidenceRate
countyCode	1										
povertyPercent	-0.146363984	1									
PovertyEst	-0.064265353	0.014599263	1								
medIncome	0.079242186	-0.792591663	0.122277917	1							
popEst2015	-0.06146342	-0.077350629	0.969062479	0.245803594	1						
avgAnnCount	-0.065752263	-0.094092769	0.938263803	0.258190251	0.980727959	1					
fiveYearTrend	-0.015415992	0.045624467	-0.027341765	-0.061242322	-0.033033195	-0.030774824	1				
deathRate	-0.05989227	0.425075273	-0.101259274	-0.438203845	-0.141823111	-0.130561067	0.072804059	1			
avgDeathsPerYear	-0.067421608	-0.080887415	0.942915055	0.233384418	0.976437706	0.997006718	-0.032876517	-0.112156	1		
studyCount	-0.027696524	-0.037334895	0.773550674	0.167442408	0.783321128	0.785423649	-0.021037242	-0.079614	0.786605807	1	
incidenceRate	-0.069392423	0.009791561	0.007187813	-0.005314196	0.017826578	0.059774897	0.182069882	0.454225	0.054757218	0.050475861	1

Figure 7 Correlation Coefficients of Dependent variable with predictor variables

Interpretation:

It can be assumed that some of the weak correlations can be attributed to the smaller sample size used to determine the incidence rate as changes between county populations will have a much smaller impact on the correlation as say the change between city population. Another assumption we made was that the weak correlation between incidence rate and median income could be attributed to the fact that those will lower income may not have the resources or capability to go to the hospital to report their illness, which would in turn lead their incident to go unreported during the clinical trials. It was surprising to see how weak certain correlations were with one another given the

Regression Analysis:

Cancer incidence is one of the most important factors that can determine how the disease will progress and evolve in the future. It is defined as a ratio of cancer incidence count to population count. It is often expressed as the number of cases per 100,000 population at risk. Cancer incidence rates are helpful in determining future cancer cases and the specific needs of detecting and preventing them needed by a given population. The timely need to evaluate trends and effects of cancer detection and prevention can go a long way in dealing with this public healthcare issue.

To predict future Cancer Incidence Rates, we created a Multi Linear Regression Model for it. Taking the predictor variables that we thought would most influence cancer incidence rate, we got the following result

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.561451213							
R Square	0.315227464							
Adjusted R Square	0.31350011							
Standard Error	45.53219004							
Observations	2783							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	7	2648364.649	378337.8	182.4915	8.02E-223			
Residual	2775	5753075.416	2073.18					
Total	2782	8401440.065						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	246.9912669	11.61573979	21.2635	4.46E-93	224.214901	269.7676327	224.214901	269.7676327
countyCode	-0.00018891	5.66077E-05	-3.33718	0.000857	-0.000299907	-7.79125E-05	-0.000299907	-7.79125E-05
povertyPercent	-1.002076112	0.229725523	-4.36206	1.34E-05	-1.452526335	-0.55162589	-1.452526335	-0.55162589
medIncome	0.000567656	0.000123655	4.590653	4.62E-06	0.000325191	0.00081012	0.000325191	0.00081012
popEst2015	-0.000125391	1.29595E-05	-9.67553	8.43E-22	-0.000150802	-9.99792E-05	-0.000150802	-9.99792E-05
avgAnnCount	0.032802135	0.003100722	10.57887	1.14E-25	0.026722179	0.038882091	0.026722179	0.038882091
fiveYearTrend	2.016653567	0.204985548	9.838028	1.79E-22	1.614713965	2.418593169	1.614713965	2.418593169
deathRate	1.105026746	0.036072523	30.63348	8.8E-178	1.03429505	1.175758443	1.03429505	1.175758443

Table 3 Regression Model Results

Interpretation:

Regression Model Equation for Incidence Rate:

$$\begin{aligned}\textbf{Incident Rate} = {} & 246.99 - 0.0001889(\textbf{countyCode}) - 1.00207(\textbf{povertyPercent}) \\ & + 0.000567(\textbf{medIncome}) - 0.000123591(\textbf{popEst2015}) \\ & + 0.0328(\textbf{AvgAnnCount}) + 2.01665(\textbf{FiveYearTrend}) + 1.10503(\textbf{deathrate})\end{aligned}$$

Explanation of the variables:

Country Code:

The slope coefficient of county code (-0.00018891) suggests that for every additional county code, Incident rate is expected to decline by 0.0189%. It would generally be expected that as county code rises, there should be an increase in incident rate since more people would hypothetically led to more/additional incident rate.

Poverty Percent:

The slope coefficient of poverty percent (-1.00207611) indicates poverty rate increases by one percent; Incident rate is predicted to decline by 1.002076%. It would generally be expected that as

poverty percent rises, there should be an increase in cancer incident rate since more people would hypothetically not be able to afford health care services.

Median Income:

Here, for every 1,000 increases in median Income, Incident rate is expected to increase by 0.05676%. It would generally be expected that as average/median income rises, there should be a decline in cancer incident rate since more people would hypothetically have the financial resource to have access to health care.

popEst2015:

The slope coefficient of population estimate (-0.0001254) suggests that for every additional/increase in population, cancer Incident rate is expected to decline by 0.01254%.

AvgAnnCount:

As average Annual count increases by 1 year, Incident rate is expected to increase by 3.2802%. Holding other variables constant.

fiveYearTrend:

As five-year trend increases by 1 year, Incident rate is expected to increase by 2.0165%. Holding County code, poverty percent, median income, population estimate, Average Annual count, five-year trend and death rate constant.

Death Rate:

The slope coefficient of 1.105027 suggest that as death rate increase by 1%, Cancer Incident rate is predicted to increase by 1.1% holding other variables constant.

Coefficient of Determination:

R- Square= 0.31523. That means that 31.52% of the variation in cancer incident rates are explained by the model.