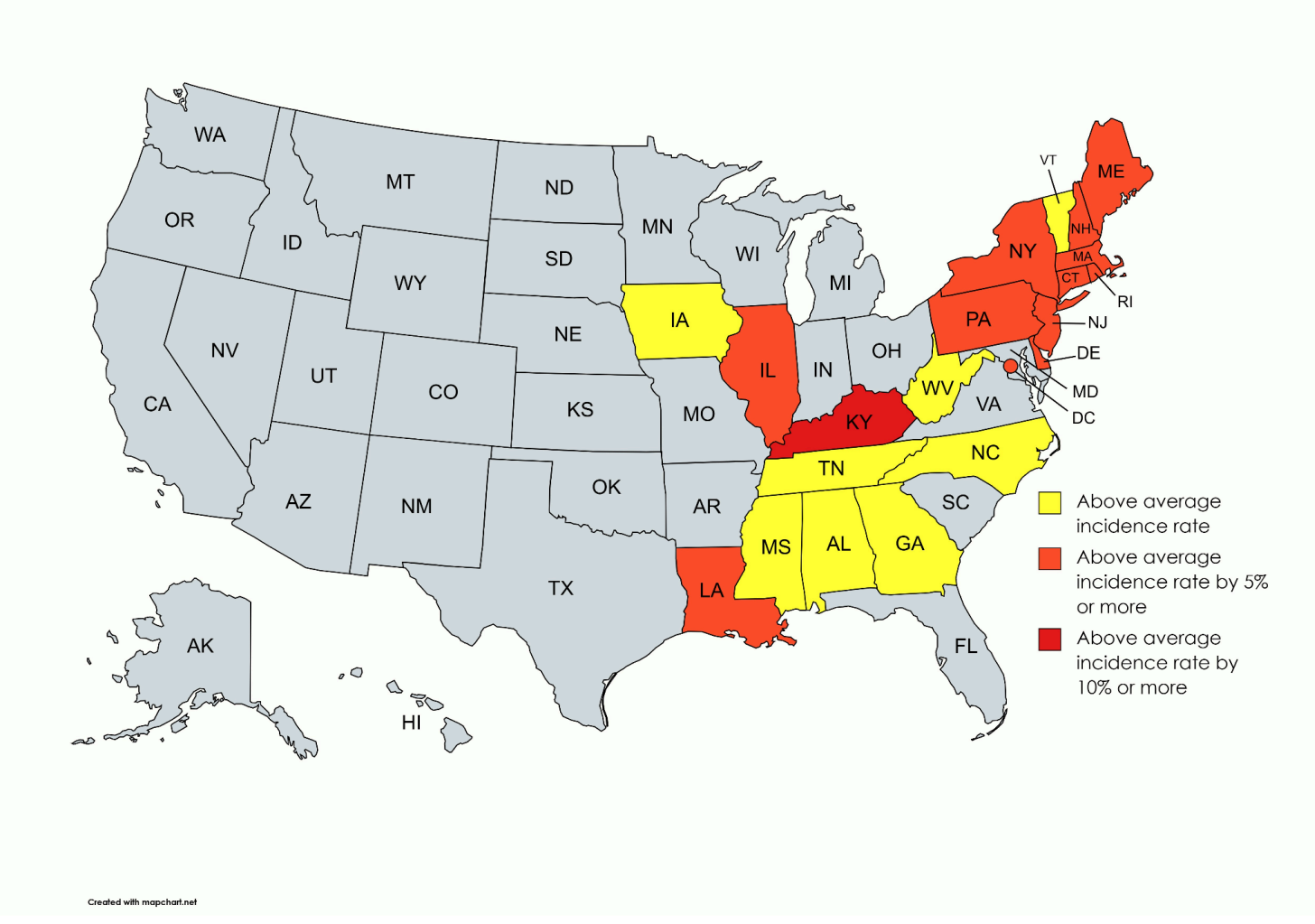
The Role of Socioeconomic Status & Region In Cancer

An analysis by Juan-Carlos Rivera, Deep Viral Mehta, Benjamin Hart, Prince Chowdhury, Dan Dewerth

Executive Summary: There are myriad factors that contribute to the incidence of cancer deaths in the United States. Much has been published on risk factors and lifestyle decisions that influence the prevalence of cancer. Preventative measures, early detection and treatment are critical towards the goal of reducing cancer deaths. In the absence of a cure, resource limitations are still a constraint on our collective response. The goal of this paper is to share insights on demographic factors that influence cancer deaths in order to inform decision makers about where these resources should be spent. Key findings include that cancer incidence can be predicted by both negative factors (things that actually cause cancer) and positive factors (things that merely detect cancer in the population).



Methodology: Available data was analyzed using statistical methods. An appendix is provided with the numerical results. In brief, we identified the States with the highest incidence of cancer and then sought to explain the higher than average rate with other dependent variables that might be driving those higher than average incidence rates.

General region and risk. Our findings point to 20 states and 1 federal district that exceed the average (mean) cancer incidence rate of 453.55 per 100,000. The north east contains 10 states that surpass the average incidence rate, 9 of those surpass the average incidence rate by **5%** or more. Kentucky surpasses the average incidence rate by **12%**.Overall, The South and North east present with a total of 18 states and 1 federal district that exceed the national average and are the regions of the country that are most cancer prone.

Study Count and Risk: Several areas are medical research hubs and have a higher incidence of clinical trials. Study count has an effect on cancer rates. Areas with high study counts (greater than 2) have a **1.05%** higher incidence rate than regions with low study counts (2 or less). Areas that conduct research at a higher rate are likely to detect and treat at a higher rate. People even migrate to these hubs for treatment and to participate in clinical trials of traditional and experimental treatment. There is, therefore, a positive bias towards some areas that are already taking action to detect and treat cancer. In an ideal world, cancer screening would be universal and access equal.

Additionally, we encourage medical networks within each state to share their data to facilitate easier research since many counties, and zip codes have had no studies performed. Furthermore, only 4 out of the 50 states (**8%**) had a mean of 5 or more studies.

Rural regions are affected the worst by lack of medical care and research but their population size is not insignificant. “More than 60 million Americans—about one-fifth of the U.S. population—live in rural areas. On average, rural residents are older and generally have worse health conditions than urban residents” *(U.S. Government Accountability Office, Why health care is harder to access in rural America 2022)*. “More than **100** (or 4% of) rural hospitals closed from 2013 through 2020. As a result, residents had to travel about **20 miles** farther for common services like inpatient care and **40 miles** farther for less common services” *(U.S. Government Accountability Office, 2022)*.

Income: Our findings show that there is a significant difference in cancer incidence rate based on median income. Concerningly, as median income increases cancer incidence rate increases to above the national average. However even at an income of $ 55,832 or higher, your cancer incidence rate is only **.575%** more than the national average of 453.55 per 100,000.

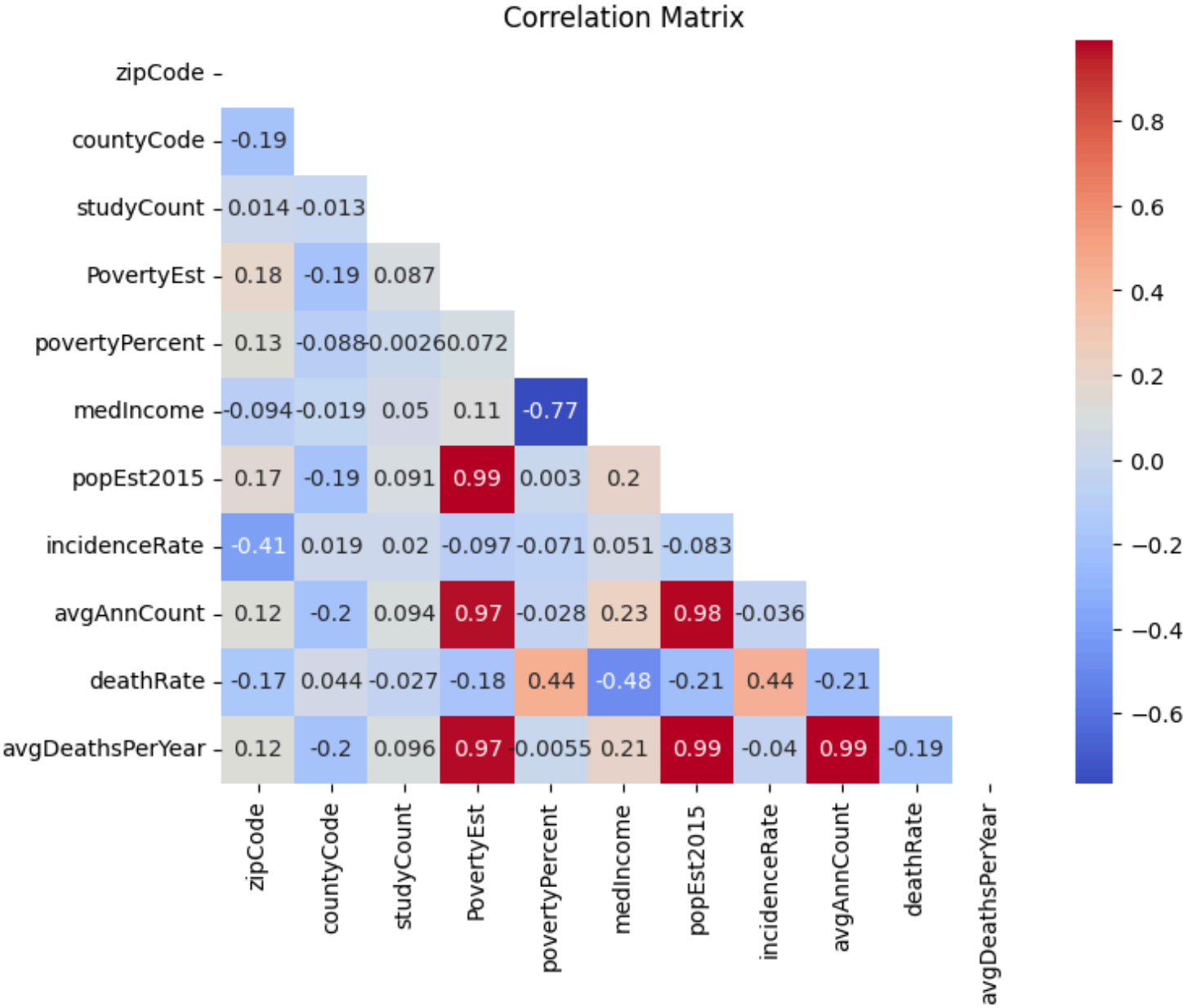
This surprising finding can be explained by several factors. Again, detection (while desirable) does drive the incidence rate up). Secondly, many people within the poor cohort are at risk of early death from all causes. “Researchers found that, compared to their wealthier counterparts, people with low socioeconomic status were **46%** more likely to die early”*(Brogan, Early death and ill health linked to low socioeconomic status: Imperial News: Imperial College London 2017).* This early death risk means that they are less likely to reach “the median age of a cancer diagnosis which is **66 years old**” *(National Cancer Institute, Risk factors: Age 2021)*.

Along this line of thinking, it is worth nothing that the dependent variable we study throughout this paper (incidence rate) has only a weak (below .45) positive correlation with death rate. This is in line with our findings and several things we know about cancer treatment

today. As cancer incidence increases some will inevitably lead to cancer deaths. However, early treatment and detection is the key to avoiding cancer incidence from converging into deaths.

Stable vs. Falling: If you are located in a region with a stable recent trend in incidence rates your risk is greater than areas where the recent trend in incidence rates is falling. However this risk is only increased by **1.52%** compared to regions that have a falling recent trend. Additionally, areas with a stable incidence rate had an **8.1%** higher mean death rate when compared with areas that had a falling incidence rate. Which can be attributed to the reality that areas with more cancer incidence inevitably will have higher rates of death from said cancer. Also, this confirms the findings of our correlation analysis concerning incidence rate and death rate.

A general predictive model using multiple regression: To preface our model it is necessary to define multicollinearity, which "is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model" *(Hayes, Multicollinearity: Meaning, examples, and faqs 2023)*. Multicollinearity can lead to misleading results when an analyst attempts to determine how well each independent variable can be used most effectively to predict the dependent variable in a statistical model *(Hayes, 2023)*. As such, we took into account the potential correlations between key independent variables we studied. The diagram below shows some of those independent variables and was instructive to the eventual model we built.


Correlation Matrix
zipCode
countyCode
studyCount
PovertyEst
povertyPercent
medIncome
popEst2015
incidenceRate
avgAnnCount
deathRate
avgDeathsPerYear
-0.19
0.014 -0.013
0.18 -0.19 0.087
0.13 -0.088 0.0026 0.072
-0.094 -0.019 0.05 0.11 -0.77
0.17 -0.19 0.091 0.99 0.003 0.2
-0.41 0.019 0.02 -0.097 -0.071 0.051 -0.083
0.12 -0.2 0.094 0.97 -0.028 0.23 0.98 -0.036
-0.17 0.044 -0.027 -0.18 0.44 -0.48 -0.21 0.44 -0.21
0.12 -0.2 0.096 0.97 -0.0055 0.21 0.99 -0.04 0.99 -0.19
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Correlations Among Key Independent Variables

Median Income vs. Poverty Percentage:

- **Strong negative correlation (above 0.75).**
- **As median income increases, the poverty percentage decreases.**

2015 Population Estimate vs. Poverty Estimation:

- **Strong positive correlation.**
- **An increase in the estimated population leads to more people below the poverty line.**

2009-2013 Average Annual Count vs. Poverty Estimation:

- **Strong positive correlation.**
- **More average annual incidents are associated with a higher number of people below the poverty line.**

Average Deaths Per Year vs. Poverty Estimation:

- **Strong positive correlation.**
- **An increase in average deaths per year corresponds to a higher number of people below the poverty line.**

Average Annual Count vs. 2015 Population Estimate:

- **Strong correlation.**
- **An increase in the average annual count is linked to a higher estimated population within a county.**

Average Deaths Per Year vs. 2015 Population Estimate:

- **Strong correlation.**
- **As average deaths per year increase, the estimated population within a county also increases.**

Average Deaths Per Year vs. Average Annual Count:

- **Strong correlation.**
- **An increase in average deaths per year is associated with an increase in average annual incidents.**

Considering the aforementioned information we tested a total of 17 different model variations to find the model that best predicted incidence rate. Our best model allowed us to

predict **53.70%** of the variation within incidence rates. Our model used median income, average annual count, death rate, five year trend, poverty estimation, and poverty percent as continuous predictors (variables that have an infinite number of values between any two values). The recent trend and state served as categorical variables (variables that contain a finite number of categories / distinct groups). Other combinations yielded multicollinearity issues and redundant results across the analysis.

To hypothesize the remaining **46.30%** we suggest looking into the lifestyle factors of and environmental factors of each region specifically. Considering lifestyle factors,"alcohol consumption accounts for about **6%** of all cancers and **4%** of all cancer deaths in the United States"*(American Cancer Society, Alcohol use an d cancer 2020)*. Smoking is another key lifestyle factor to monitor, "those who smoke are as much as **30** times more likely to get lung cancer than those who do not" *(MD Lynne Eldridge, What percentage of smokers get lung cancer? 2021)*. Diet, nutrition, activity, and body mass play vital roles as well in data that needs to be monitored. "American Cancer Society researchers estimate **18%** of cancer cases and **16%** of cancer deaths are related to a combination of eating poorly, drinking too much alcohol, not getting enough physical activity, and being overweight" *(American Cancer Society, Diet, exercise, and your cancer risk 2019)*.

In regards to additional environmental factors that we suggest be reviewed in each region air pollution is a key metric. "Different types of air pollution have been linked to a variety of cancers"*(Miller et al., Can air pollution cause cancer? what you need to know about the risks. 2021)*. Additionally water pollution is a key metric that needs to be reviewed. "Increased cancer risks were linked to **22 carcinogens** found in the drinking water. Contaminants included arsenic; radioactive materials, such as uranium and radium; and disinfectant byproducts, which are

substances produced when chlorine and other additives are used in the treatment process" *(Evans et al., Cumulative risk analysis of carcinogenic contaminants in United States drinking water 2019)*.

By combining these factors with the existing data modeled around population, poverty, and regional analysis our model can be improved to predict a greater percentage of the variability within incidence rate.

Our key takeaway is that further research is needed to fully distinguish the reasons for higher incidence that are due to negative factors in the environment (whether caused by poverty, lifestyle, or environmental risks) and positive factors that simply lead to greater detection (such as screenings). Though people with higher median income have slightly higher incidence rates they have the means to deal with cancer and reduce the chance of falling into the death rate statistic should cancer arise. Furthermore, high risk regions should be targeted for increased monitoring and study in order to understand what missing factors are leading to their incidence rate increases beyond what is stated within our model.

Bibliography

American Cancer Society. (2019, April 12). Diet, exercise, and your cancer risk. https://www.cancer.org/cancer/latest-news/diet-exercise-and-your-cancer-risk.html

American Cancer Society. (2020, June 9). Alcohol use and cancer. https://www.cancer.org/cancer/risk-prevention/diet-physical-activity/alcohol-use-and-cancer.html

Brogan, C. (2017, January 31). Early death and ill health linked to low socioeconomic status: Imperial News: Imperial College London. Imperial News. https://www.imperial.ac.uk/news/177249/early-death-health-linked-socioeconomic-status/

CITY OF HOPE. (2021, June 11). Study says carcinogens in drinking water linked to thousands of cancers. City of Hope. https://www.cancercenter.com/community/blog/2020/05/drinking-water-cancer-risk

Evans, S., Campbell, C., & Naidenko, O. V. (2019, September 18). Cumulative risk analysis of carcinogenic contaminants in ... - cell press. Cumulative risk analysis of carcinogenic contaminants in United States drinking water. https://www.cell.com/heliyon/pdf/S2405-8440(19)35974-2.pdf

Hayes, A. (2023a, February 25). Multicollinearity: Meaning, examples, and faqs. Investopedia. https://www.investopedia.com/terms/m/multicollinearity.asp

Hayes, A. (2023b, July 29). Correlation: What it means in finance and the formula for calculating it. Investopedia. https://www.investopedia.com/terms/c/correlation.asp

MD Lynne Eldridge. (2021, July 20). What percentage of smokers get lung cancer?. Verywell Health. https://www.verywellhealth.com/what-percentage-of-smokers-get-lung-cancer-2248868

Miller, M., Milevski, L., Song, L., Younes, L., Kofman, A., & Shaw, A. (2021, November 2). Can air pollution cause cancer? what you need to know about the risks. ProPublica. https://www.propublica.org/article/can-air-pollution-cause-cancer-risks

National Cancer Institute. (2021, March 5). Risk factors: Age. https://www.cancer.gov/about-cancer/causes-prevention/risk/age

Park, A. (2016, July 18). Race and poverty: How they contribute to early death. Time. https://time.com/4410610/race-poverty-health-death/#tbl-em-lnqo1awmp0rjviquhz

U.S. Government Accountability Office. (2022, August 9). Why health care is harder to access in rural America. U.S. GAO. https://www.gao.gov/blog/why-health-care-harder-access-rural-america

Appendix

Statistics

Variable	State	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
incidenceRate	AK	178	421.25	69.96	201.30	373.50	441.40	475.10	501.10
	AL	639	459.33	26.32	391.80	438.80	460.70	475.90	555.60
	AR	588	435.60	37.01	364.10	401.90	433.20	462.30	532.10
	AZ	402	369.25	43.11	269.90	339.20	379.10	401.40	444.90
	CA	1750	427.89	25.98	312.80	412.60	420.30	447.00	490.80
	CO	512	399.14	43.66	234.00	381.70	408.00	434.30	454.70
	CT	279	482.50	18.35	431.40	480.90	488.80	490.40	500.10
	DC	31	483.70	0.000000	483.70	483.70	483.70	483.70	483.70
	DE	67	498.19	16.68	485.90	485.90	493.30	493.30	527.20
	FL	980	436.95	58.80	259.60	409.20	430.80	470.70	1206.90
	GA	723	456.93	38.20	295.70	439.60	463.00	480.20	550.70
	HI	92	411.56	12.91	392.20	402.70	418.30	425.20	425.20
	IA	931	468.79	28.19	390.30	446.80	470.80	487.20	548.40
	ID	273	431.54	47.66	345.30	383.50	435.50	469.00	538.60
	IL	1380	482.91	26.70	392.30	468.60	480.10	502.60	551.50
	IN	770	453.15	26.00	396.80	435.70	451.10	476.00	535.70
	KS	695	453.55	0.000000	453.55	453.55	453.55	453.55	453.55
	KY	760	517.33	34.56	343.20	496.00	518.80	536.20	639.70
	LA	512	486.81	28.27	379.00	470.80	487.10	499.60	586.20
	MA	534	477.32	29.32	413.90	470.30	473.10	497.30	572.80
	MD	466	452.39	38.47	392.90	420.50	457.00	481.60	532.10
	ME	432	480.25	19.89	435.20	472.80	479.60	497.70	507.30
	MI	974	448.98	56.25	310.10	406.20	463.90	495.70	541.30
	MN	880	453.55	0.000000	453.55	453.55	453.55	453.55	453.55
	MO	1015	442.57	45.75	287.40	421.80	451.60	475.20	529.50
	MS	417	471.22	42.50	339.00	443.10	468.10	501.00	561.90
	MT	348	450.37	45.21	348.10	420.20	446.65	481.05	587.00
	NC	803	459.46	31.28	370.90	442.60	457.60	482.30	524.40
	ND	376	440.19	57.63	254.70	403.70	456.90	475.30	579.70
	NE	559	434.05	42.81	221.50	409.00	431.10	463.50	529.00
	NH	248	485.15	26.10	435.50	475.70	482.20	507.60	518.80
	NJ	594	494.01	34.34	399.50	468.10	493.10	529.60	559.10
	NM	364	375.17	38.12	303.00	355.60	366.50	393.00	558.50
	NV	174	453.55	0.000000	453.55	453.55	453.55	453.55	453.55
	NY	1767	497.52	28.14	435.20	477.40	500.80	514.30	577.40
	OH	1191	452.34	31.27	316.50	439.30	458.70	473.10	514.70
	OK	644	442.32	35.99	354.70	418.20	449.40	465.50	516.30
	OR	415	442.77	32.61	358.50	419.40	441.70	459.50	587.40
	PA	1787	481.62	25.41	393.50	470.70	484.00	496.10	560.40
	RI	77	479.71	9.56	469.00	473.30	473.30	492.30	492.30
	SC	420	451.81	26.36	383.00	430.90	459.00	470.40	497.90
	SD	343	436.01	55.65	284.80	403.80	426.40	468.10	577.00
	TN	617	469.90	27.25	407.40	453.30	470.60	484.40	538.70
	TX	1902	411.91	41.15	211.10	393.10	424.25	434.90	537.20
	UT	278	398.96	36.50	261.10	377.60	404.10	417.90	489.90
	VA	871	427.37	58.98	265.70	381.80	421.60	463.30	1014.20
	VT	254	467.04	19.13	436.90	461.60	463.70	476.20	548.50
	WA	592	452.26	44.79	293.60	423.80	461.80	490.90	518.40
	WI	768	447.84	46.73	278.50	428.82	455.00	477.00	572.80
	WV	703	469.87	46.22	234.00	447.80	469.60	506.60	591.00
	WY	176	407.56	39.31	280.90	397.10	411.30	436.73	484.20

Correlations

Means

medIncome	N	Mean	StDev	95% CI
very low	7859	450.800	56.825	(449.752, 451.848)
low	7609	454.605	44.192	(453.540, 455.670)
high	7278	454.160	42.301	(453.071, 455.249)
very high	7595	456.175	44.297	(455.109, 457.241)

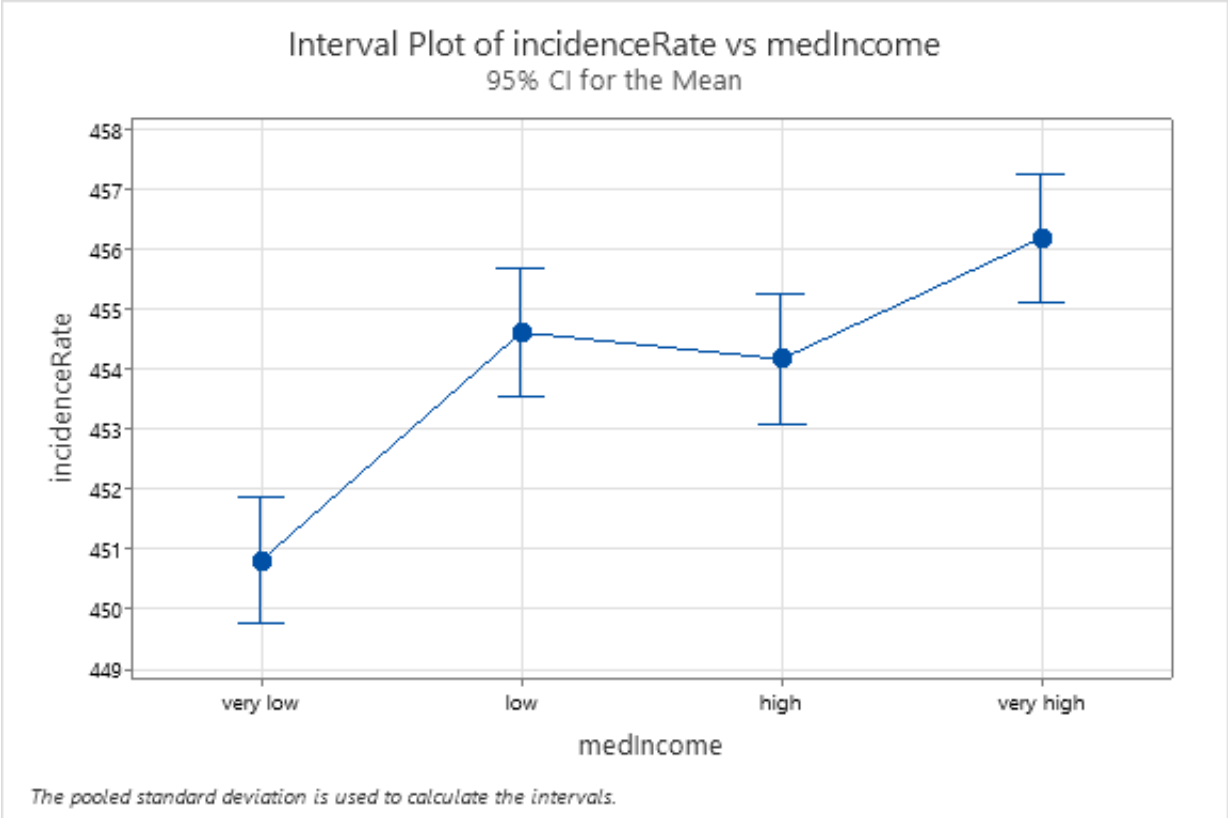
Pooled StDev = 47.3982

	int	PovertyEst	povertyPercent	medIncome
	87			
	03	0.072		
	50	0.111	-0.766	
	91	0.989	0.003	0.204
	20	-0.097	-0.071	0.051
	94	0.966	-0.028	0.232
fiveYearTrend	0.005	-0.019	-0.007	-0.058	0.043	-0.064
deathRate	-0.166	0.044	-0.027	-0.178	0.440	-0.483
avgDeathsPerYear	0.122	-0.195	0.096	0.973	-0.006	0.208

	popEst2015	incidenceRate	avgAnnCount	fiveYearTrend	deathRate
countyCode					
studyCount					
PovertyEst					
povertyPercent					
medIncome					
popEst2015					
incidenceRate	-0.083				
avgAnnCount	0.983	-0.036			
fiveYearTrend	-0.062	0.176	-0.058		
deathRate	-0.213	0.436	-0.213	0.094	
avgDeathsPerYear	0.988	-0.040	0.993	-0.058	-0.193

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
medIncome	3	119104	39701	17.67	0.000
Error	30337	68154720	2247		
Total	30340	68273824			



The pooled standard deviation is used to calculate the intervals.

Regression Equation

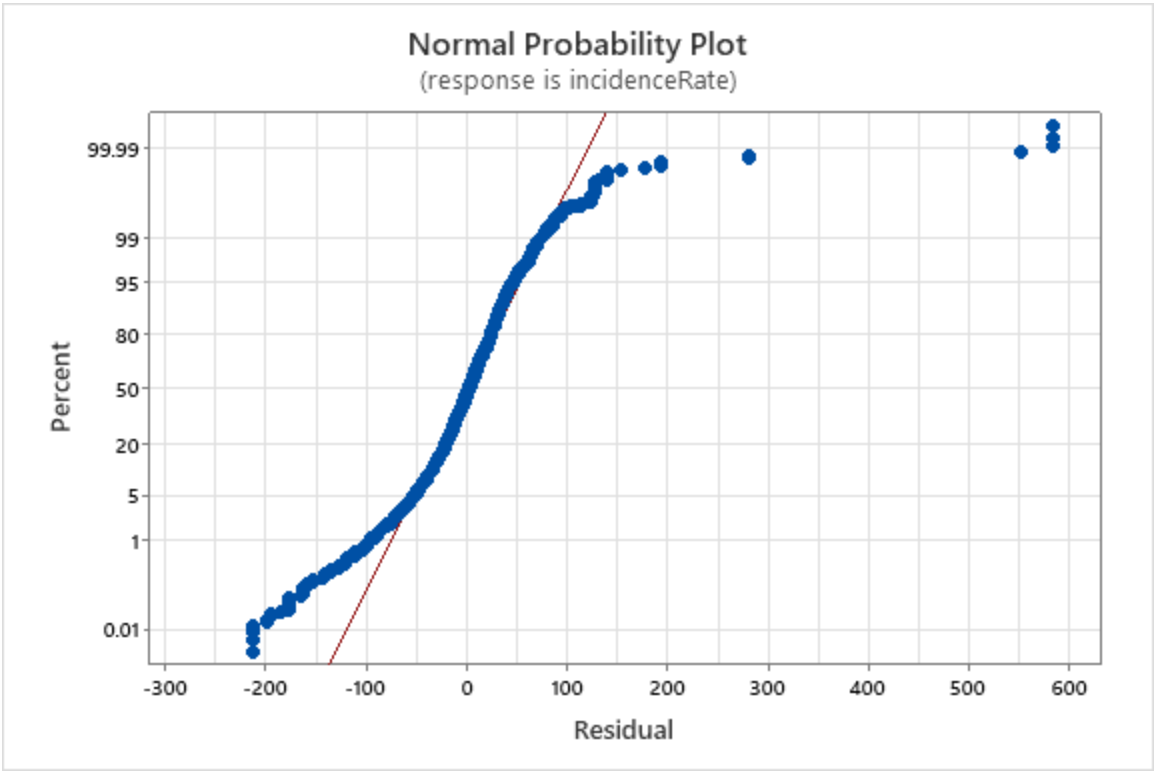
incidenceRate = 227.60 + 0.000305 medIncome + 1.0048 deathRate - 0.000090 PovertyEst
- 1.0659 povertyPercent + 0.004587 avgAnnCount + 1.2196 fiveYearTrend
+ 0.0 recentTrend_falling + 15.24 recentTrend_rising
+ 0.559 recentTrend_stable + 0.0 State_AK + 49.76 State_AL + 20.33 State_AR
- 2.08 State_AZ + 40.69 State_CA + 28.25 State_CO + 78.45 State_CT
+ 67.08 State_DC + 85.04 State_DE + 42.60 State_FL + 59.33 State_GA
+ 32.85 State_HI + 68.41 State_IA + 49.70 State_ID + 68.03 State_IL
+ 38.73 State_IN + 81.36 State_KY + 69.45 State_LA + 68.05 State_MA
+ 39.87 State_MD + 71.05 State_ME + 45.43 State_MI + 32.10 State_MO
+ 55.67 State_MS + 69.84 State_MT + 62.33 State_NC + 51.01 State_ND
+ 41.68 State_NE + 79.25 State_NH + 79.95 State_NJ + 7.97 State_NM
+ 92.90 State_NY + 40.76 State_OH + 27.30 State_OK + 49.31 State_OR
+ 72.53 State_PA + 77.42 State_RI + 45.97 State_SC + 48.77 State_SD
+ 50.78 State_TN + 22.72 State_TX + 33.07 State_UT + 23.51 State_VA
+ 64.14 State_VT + 52.81 State_WA + 46.58 State_WI + 47.04 State_WV
+ 18.71 State_WY

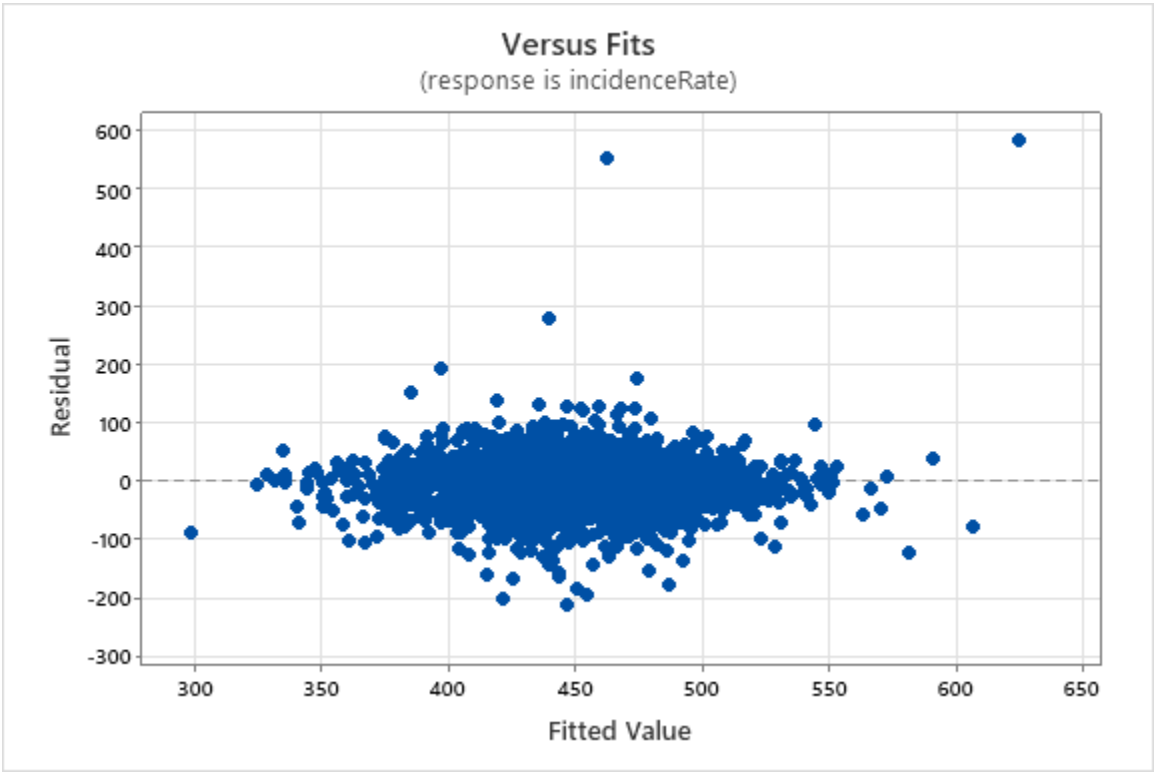
Analysis of Variance

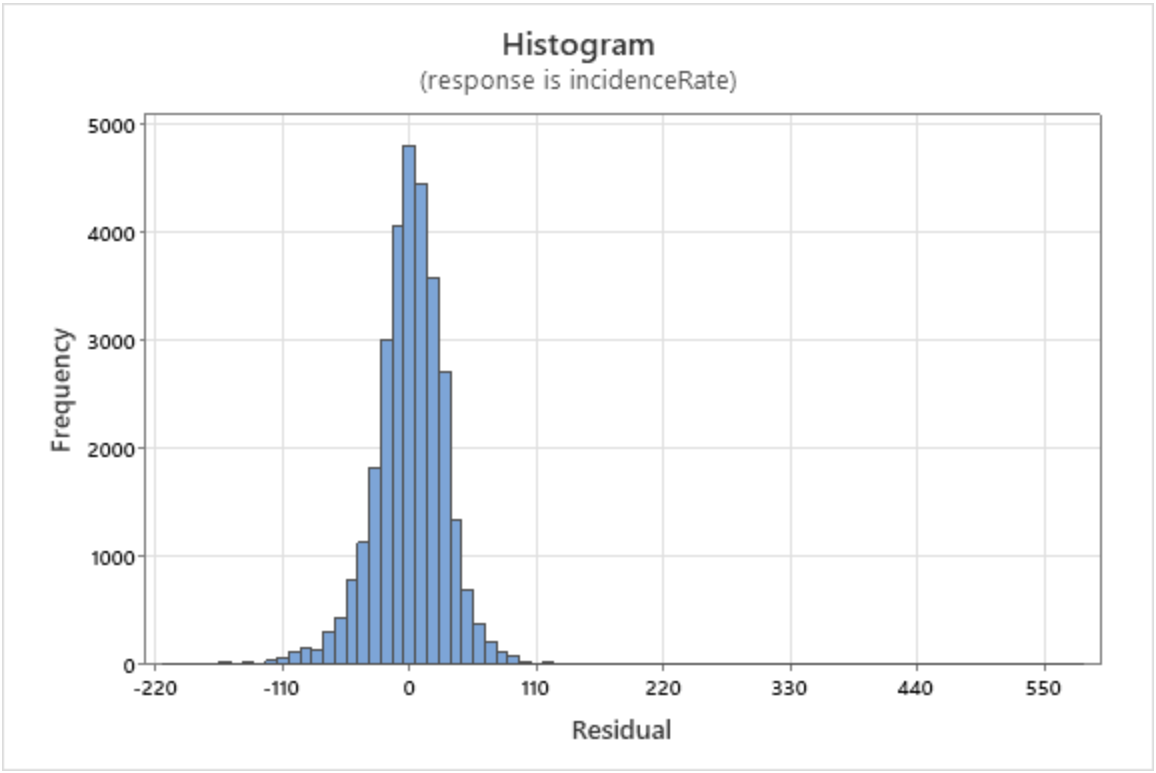
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	55	37153479	675518	645.86	0.000
medIncome	1	114833	114833	109.79	0.000
deathRate	1	8605109	8605109	8227.30	0.000
PovertyEst	1	396414	396414	379.01	0.000
povertyPercent	1	317033	317033	303.11	0.000
avgAnnCount	1	480258	480258	459.17	0.000
fiveYearTrend	1	379053	379053	362.41	0.000
recentTrend	2	49269	24635	23.55	0.000
State	47	12530416	266605	254.90	0.000
Error	30524	31925696	1046		
Lack-of-Fit	2740	31925696	11652	*	*
Pure Error	27784	0	0		
Total	30579	69079175			

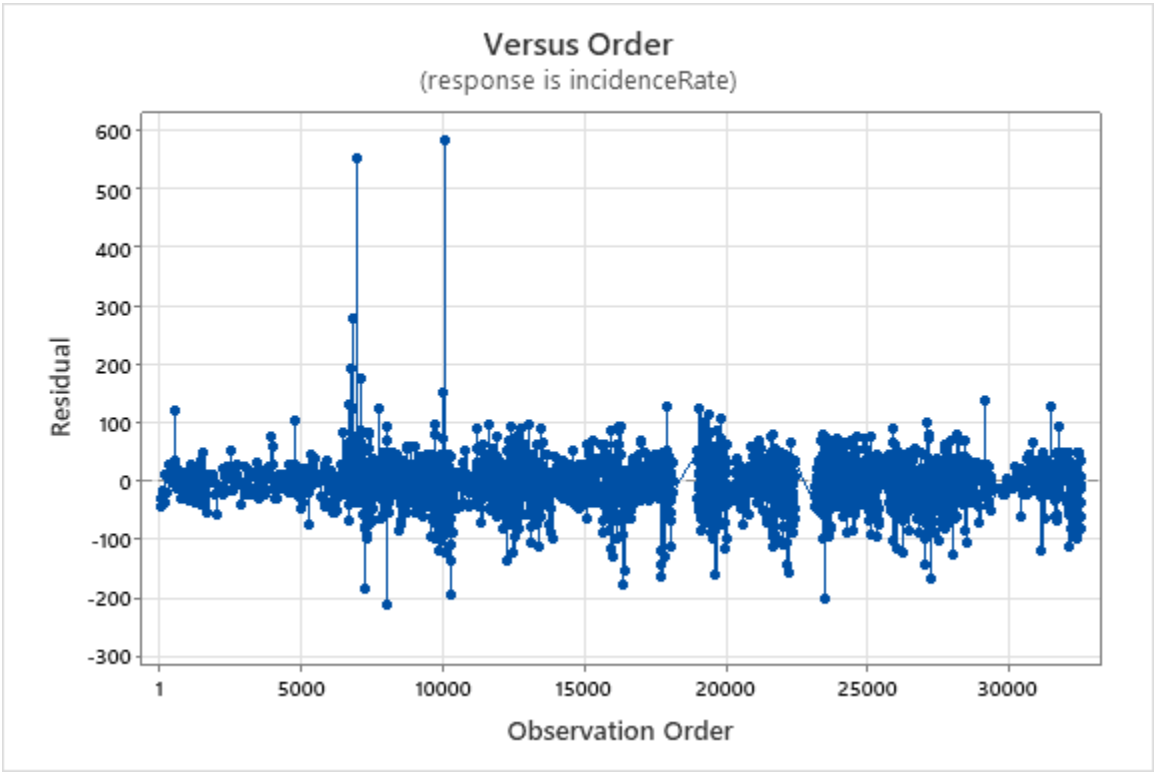
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.3407	53.78%	53.70%	53.58%


Normal Probability Plot
(response is incidenceRate)
Percent
99.99
99
95
80
50
20
5
1
0.01
-300
-200
-100
0
100
200
300
400
500
600
Residual


Versus Fits
(response is incidenceRate)
Residual
600
500
400
300
200
100
0
-100
-200
-300
300
350
400
450
500
550
600
650
Fitted Value


Histogram
(response is incidenceRate)
Frequency
5000
4000
3000
2000
1000
0
-220
-110
0
110
220
330
440
550
Residual


Versus Order
(response is incidenceRate)
Residual
600
500
400
300
200
100
0
-100
-200
-300
1
5000
10000
15000
20000
25000
30000
Observation Order

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
High study count	2339	458.0	39.6	0.82
Low study count	30212	453.2	47.4	0.27

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
5.59	2881	0.0000000121163756

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
incidenceRate_stable	23385	455.5	48.8	0.32
incidenceRate_falling	6956	448.6	42.1	0.51

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	0.000000000000000