Average Per Capita Cancer Mortalities Analysis using Multi-Linear Regression

David A. Iniguez

March 11, 2024

1 Abstract

In this paper we discuss and analyze a cancer data set (https://data.world/nrippner/ols-regression-challenge) using Ordinary Least Squares Multi-Linear Regression. The goal is to create an interpretable model to to predict mean per-capita (100,000) cancer mortalities with 32 possible predictive variables and gleam some insight from the factors that influence cancer mortailites among a population.

2 Data Cleaning

We begin by looking at the features given in the data set.

1. deathRate: Mean per capita (100,000) cancer mortalities
2. avgAnnCount: Mean number of reported cases of cancer diagnosed annually
3. avgDeathsPerYear: Mean number of reported mortalities due to cancer
4. incidenceRate: Mean per capita (100,000) cancer diagoses
5. medianIncome: Median income per county
6. popEst2015: Population of county
7. povertyPercent: Percent of populace in poverty
8. studyPerCap: Per capita number of cancer-related clinical trials per county
9. binnedInc: Median income per capita binned by decile
10. MedianAge: Median age of county residents
11. MedianAgeMale: Median age of male county residents

12. MedianAgeFemale: Median age of female county residents
13. Geography: County name
14. AvgHouseholdSize: Mean household size of county
15. PercentMarried: Percent of county residents who are married
16. PctNoHS18-24: Percent of county residents ages 18-24 highest education attained: less than high school
17. PctHS18-24: Percent of county residents ages 18-24 highest education attained high school diploma
18. PctSomeCol18-24: Percent of county residents ages 18-24 highest education attained: some college
19. PctBachDeg18-24: Percent of county residents ages 18-24 highest education attained: bachelor's degree
20. PctHS25-Over: Percent of county residents ages 25 and over highest education attained: high school diploma
21. PctBachDeg25-Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree
22. PctEmployed16-Over: Percent of county residents ages 16 and over employed
23. PctUnemployed16-Over: Percent of county residents ages 16 and over unemployed
24. PctPrivateCoverage: Percent of county residents with private health coverage
25. PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance)
26. PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage
27. PctPublicCoverage: Percent of county residents with government-provided health coverage
28. PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone
29. PctWhite: Percent of county residents who identify as White
30. PctBlack: Percent of county residents who identify as Black
31. PctAsian: Percent of county residents who identify as Asian

32. PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian

33. PctMarriedHouseholds: Percent of married households

34. BirthRate: Number of live births relative to number of women in county

My first step in the data cleaning process was to split the data into testing and training set before any kind of inspection to prevent data-leakage. With the testing set being 30% of the data and the training set being the other 70%. To ensure reproducibility I set the random seed to 42.

Now I begin inspecting the training set which has 2,132 rows and 34 columns one of which is the target variable deathRate. Upon initial inspection I notice that the MedianAge column has 23 values in the range between 400 and 700 which does not seem like a reasonable median age for a given population of humans. So I removed those values from the data.

Next I begin to look for missing data and find 3 columns that contain missing data: PctSomeCol18-4, PctEmployed16-Over, PctPrivateCoverageAlone. The percent of missing data in each column is: PctSomeCol18-4 with 75% of its' data missing, PctPrivateCoverageAlone with 20% of its' data missing, and PctEmployed16-Over with 4% of its' data missing. Here I opted to remove the PctSomeCol18-4 and PctPrivateCoverageAlone columns from the data set entirely in the interest of performing Complete-Case Analysis. As for the missing data in the PctEmployed16-Over I simply removed the rows with missing data from the dataset, but did not remove the column entirely.

I also removed the Geography column from the data since I conclude it to contain no predictive information. Now the data set has 30 features (plus 1 target variable column) and 2,015 rows.

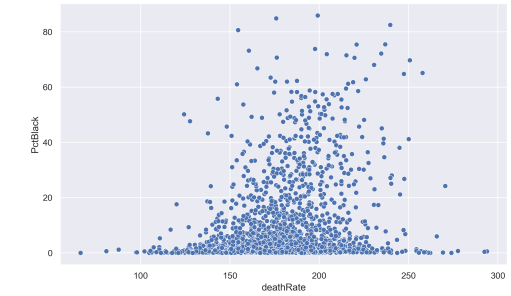
3 Exploratory Data Analysis (EDA)

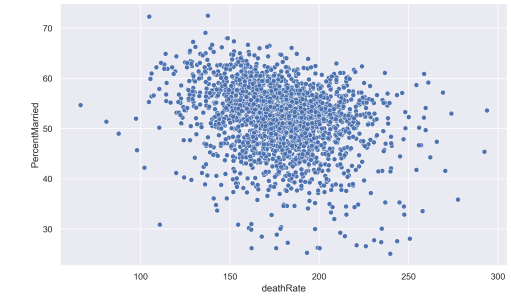
Since I have 30 features available there are a total of 2^{30} possible models to choose from which is a number that may grow exponentially if any new features are created utilizing transformations of other features. Since it is computationally impossible for me to explore all possible models, I limit my final model to only include 10 features at most to predict the deathRate target variable.

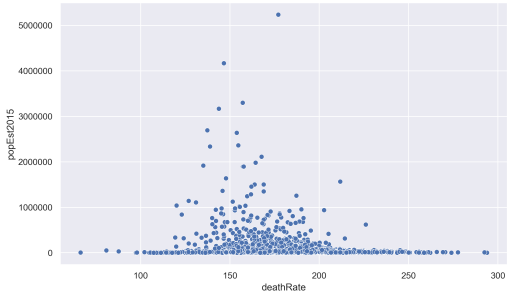
I began my investigation by looking for features with large pearson correlation values and I choose the the 5 most positively and 5 most negatively correlated values to be the initial features to use as predictive variables. These 10 features that are most correlated with the deathRate varibles are: 'PctPublicCoverageAlone', 'incidenceRate', 'PctHS25-Over','povertyPercent','PctPublicCoverage','PctBachDeg25-Over' ,'medIncome', 'PctEmployed16-Over', 'PctPrivateCoverage', and 'Pct-

MarriedHouseholds.'

The next stage of my EDA was to look for non-linear relationship with the target variable and the features. I do this by creating 30 scatter plots where each plot has a feature on the y-axis and the deathRate variable on the x-axis. A small sample of these plots look like:







Most plots had similar trends where at most we see weak linear trends, simialr to the pearson correlation coefficients. As for non-linear trends I could not find any obvious polynomial relationship in the data.

My next course of thought was to apply a log transformation to the data. This again yielded no new obvious relationships either in correlation coefficients or scatter plots with resepct to the deathRate target variable. SO, at the end of the EDA I concluded that the features with the most predicive power are the high correlation features. Looking at the descriptive statstics we see nothing out of the ordinary or erroneous based on the context of the features:

	count	mean	std	min	25%	50%	75%	max
PctPublicCoverageAlone	2015.0	19.228784	5.983793	2.600000	15.00000	18.800000	23.000000	46.600000
incidenceRate	2015.0	447.898225	52.884153	201.300000	420.90000	453.549422	480.450000	1014.200000
PctHS25_Over	2015.0	34.968586	6.968704	8.300000	30.60000	35.400000	39.700000	54.800000
povertyPercent	2015.0	16.862333	6.302165	3.900000	12.20000	15.900000	20.400000	47.000000
PctPublicCoverage	2015.0	36.293846	7.685226	11.200000	30.85000	36.500000	41.700000	65.100000
PctBachDeg25_Over	2015.0	13.177072	5.190441	2.700000	9.40000	12.300000	15.900000	39.700000
medIncome	2015.0	47005.762779	11768.135985	22640.000000	39023.50000	45159.000000	52480.500000	122641.000000
PctEmployed16_Over	2015.0	54.234342	8.231423	17.600000	48.80000	54.400000	60.400000	80.100000
PctPrivateCoverage	2015.0	64.357221	10.489683	27.200000	57.40000	64.900000	72.050000	88.800000
PctMarriedHouseholds	2015.0	51.298355	6.449612	23.885628	47.84093	51.692308	55.362114	71.703057

4 Model Selection

For the model selection, I opted to explore 3 different models. The model of large correlations from the EDA stage, forward selection, and backward selection as my 3 models each containing at most 10 features. While also using 10 fold cross-validation with the Root Mean Squared Error as my standard of measure while comparing the models.

For this stage the forward selection model achieved an average RMSE of 19.19, the backward selection model achieved an average RMSE of 19.22, and the model that used the high correlation features achieved and average RMSE of 19.62.

Next I inspected the summary outputs for each model and analyzed the R^2 values and the t-test results of each feature. After looking at the summary results taking into account the RMSE, R^2, and significance I councluded the forward model to be the best model. It summary output is as follows:

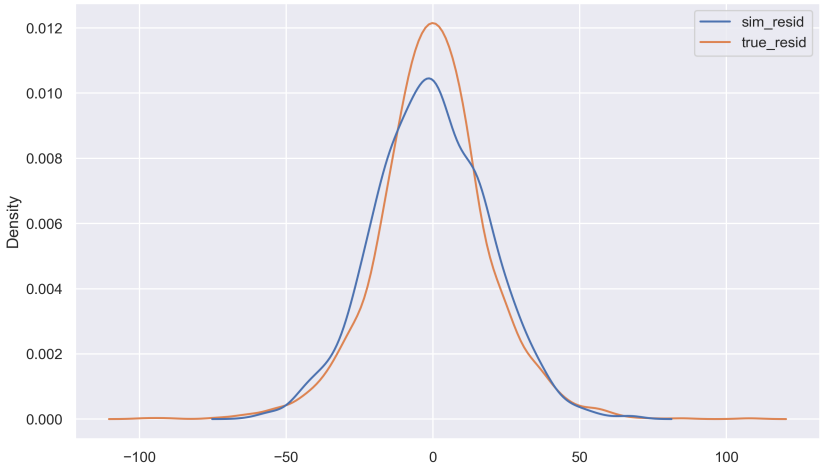
Dep. Variable:	deathRate	**R-squared:**	0.504
Model:	OLS	**Adj. R-squared:**	0.501
Method:	Least Squares	**F-statistic:**	203.6
Date:	Sat, 17 Dec 2022	**Prob (F-statistic):**	2.33e-296
Time:	06:52:42	**Log-Likelihood:**	-8793.2
No. Observations:	2015	**AIC:**	1.761e+04
Df Residuals:	2004	**BIC:**	1.767e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	160.9685	7.138	22.551	0.000	146.970	174.967
avgAnnCount	-0.0009	0.000	-2.540	0.011	-0.002	-0.000
incidenceRate	0.1943	0.009	22.223	0.000	0.177	0.211
PctHS18_24	0.2963	0.054	5.443	0.000	0.190	0.403
PctHS25_Over	0.3112	0.108	2.874	0.004	0.099	0.524
PctBachDeg25_Over	-1.3378	0.168	-7.947	0.000	-1.668	-1.008
PctPrivateCoverage	-0.8676	0.089	-9.795	0.000	-1.041	-0.694
PctEmpPrivCoverage	0.5385	0.084	6.411	0.000	0.374	0.703
PctOtherRace	-0.7792	0.132	-5.892	0.000	-1.039	-0.520
PctMarriedHouseholds	-0.6366	0.081	-7.891	0.000	-0.795	-0.478
BirthRate	-0.8331	0.217	-3.838	0.000	-1.259	-0.407

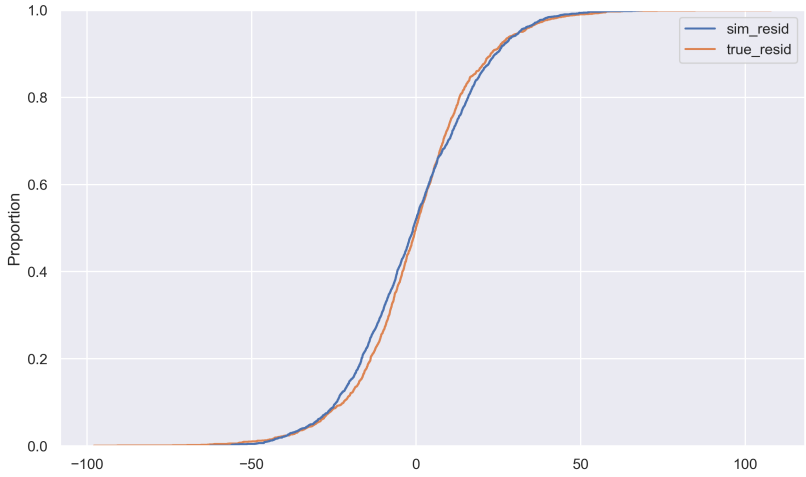
We notice that about 50% of the variance seen in the deathRate variable can be accounted for by our model and our average error was about 20 based on the RMSE with all features being statistically significant.

Looking at model diagnostics we see that the data is distributed normally. In the KDE and ECDF plots below we see that the residuals are distributed fairly normally. Ther are discrpencies around the mean, but thanks to the central limit theorem we don't have to worry about this. The most importnat part here is that we see the residuals taper off to zero at the same rate as the simulated residual data which it in fact does as evident by the plots below. Indicating that are residuals are normally distributed.

KDE Plot

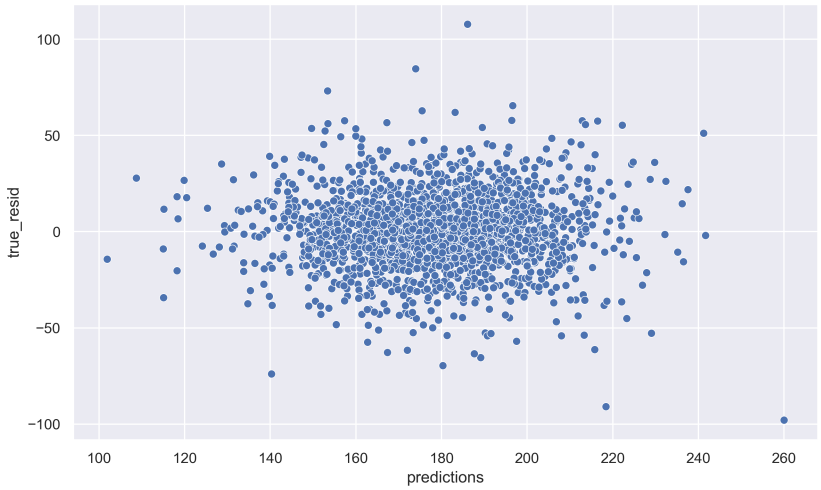


ECDF Plot



Lookng at plot below we can see no conic shape indicating there is constant varaince. The outlying points.

prediction vs. true residual plot



The model also seems to retain independence in the residuals when analyzing residuals plots. Based on these results above the model seems like a good candidate to gleam insight into the mean per-capita cancer mortalites of a population.

5 Interpretation

Looking at the summary table above we can see that the largest coefficient is the PctBachDeg25-Over feature. From it we can gleam that, with all other factors constant we can say with 95% confidence that with every 5% increase in the amount of individuals that are 25 or older and have obtained a bachlor's degree, on average we will see between a 5 and 8 decrease in average per capita cancer mortailites. Which may be because these individuals may have jobs that offer health insurance allowing individuals an avenue to stay healthy.

From the PctPrivateCoverage coefficient we can say that, on average for every 10% increase in individuals with private health insurance, we will see a decrease of 10 in average per capita cancer mortalities assuming all other factors constant. Which matches up with the logic that individuals with health insurance are more likely to stay healthy since they can see a doctor and get proper medical care either increasing their chances of avoiding a cancer diagnosis all together or getting proper medical care in the case of a diagnosis.

However, we see something contrary when looking at the PctEmpPrivCoverage coefficient. According to that coefficient, on average for ever 10 percent increase of county residents with employee-provided private health coverage, we expect to see an increase of 5 per capita average cancer moralities assuming all other factors constant. This warrents further investigation of the data to discover why these contrary conclusions are being reached.

As for the intercept I do not believe it has an interpretation that makes sense since there can be no non-zero average per-capita mortalities for populations of zero.

I hope you enjoyed the analysis.

Thank You For Reading.