Cancer Mortality Rates for US Counties (2024)

OLS Regression Challenge - MLR

Contributors

Quinn Campfield
Vani Singh
Hadley Dixon
Raghava Srinivasan

October, 2024

Project Overview

This project builds a multivariate ordinary least squares regression model to predict cancer mortality rates for US counties.

Source

Data is sourced from the website data.world as a component of the OLS Regression Challenge. Read in-depth about this OLS Regression Challenge here. Data for this competition was originally aggregated from several sources, including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. Pre-preparation of this data by the original uploader can be found here.

Variable Descriptions

The dataset used in this project, 'cancer_reg.csv', has 3,047 entries (rows), each representing a different county in the United States. The dataset includes 34 variables (columns) that provide demographic, economic and health-related information in that county. The dependent variable is **TARGET_deathRate**, measured as the mean cancer mortality rate per capita between 2010 and 2016.

There are two types of variables being tracked, including both predictor variables and the response variable.

(a) 2010-2016 County Averages
(b) 2013 County Census Estimates

See the Appendix for a complete data dictionary.

Evaluation Metrics

The evaluation metric used is adjusted R^2 (R_a^2) and Mean Squared Prediction Error (MSPE).

The final product is a model equation that predicts cancer mortality rates, as well as the accompanying code file.

Context and Motivation

The main motivation for selecting this dataset was to prepare for our practicums. Specifically, one member is beginning a research project with UCSF, as a member of their Radonc BioNx team. The UCSF PhosphoAtlas+ project aims to investigate phosphorylation networks, which play a crucial role in understanding cellular signaling and identifying potential therapeutic targets, particularly in the context of cancer treatment resistance. UCSF's work is at the forefront of cutting-edge computational techniques in the field of cancer research.

This project will equip our members with technical skills in data handling and analysis, feature selection and interpretability, modeling and prediction, as well as qualitative skills such as effective communication and collaboration efforts in a team setting, comprehensive summary deliverables, and real-world contextualization of cancer-related data.

Procedure

Research Question

What are the key demographic, geographic, socioeconomic, and healthcare factors that most significantly predict cancer mortality rates across US counties?

Summary of Methods

Exploratory Data Analysis (EDA)

EDA is the process of visually and statistically analyzing the data to summarize its main characteristics. In MLR, EDA helps you understand the distribution of the data, detect patterns, spot anomalies (such as outliers and leverage points), and examine relationships between predictor variables and the dependent variable. This allows assessment of the assumptions of linear regression: linearity between the predictors and the response variable, fixed nature of the dependent variables (independence), and multivariate normal distribution of the error terms.

Outliers and Leverage

Outliers are data points that deviate significantly from the rest. Leverage points are observations with extreme predictor values that can disproportionately influence the model. Identifying and addressing these points is essential to prevent them from skewing the regression results, ensuring a more robust and generalizable model.

Box-Cox Transformations

The Box-Cox transformation is a statistical technique that transforms non-normal dependent variables into a normal shape. In MLR, applying this transformation helps meet normality assumption, stabilizing variance and making relationships more linear. This will improve the model fit and predictive power.

ANOVA

An Analysis of Variance (ANOVA) is a statistical method used to test whether changes in one or more independent variables significantly affects a dependent variable. It partitions total

variability in the dependent variable into two components: explained variance and unexplained variance.

Model Selection (AIC, BIC, C_p, PRESS, Adjusted R^2)

Akaike Information Criterion (AIC) is a criterion used to compare models based on the trade-off between goodness-of-fit and model complexity, penalizing models for more parameters by a factor of 2. A lower AIC suggests a better-fitting model with fewer parameters, helping in selecting an optimal model that balances complexity and fit.

Bayesian Information Criterion (BIC) is a criterion used to compare models based on the trade-off between goodness-of-fit and model complexity, penalizing models for more parameters by a factor of log(n). Similar to AIC, a lower BIC suggests a better-sitting model, but enforces a stricter penalty on the number of parameters, favoring simpler models. BIC is useful for selecting models that avoid overfitting, making it ideal when there are many predictors.

Adjusted R^2 (R_a^2) is a modification of the R^2 criterion that adjusts for the number of predictors, preventing overfitting. Adjusted R^2 provides a more reliable measure of the model's explanatory power, accounting for unnecessary variables. A high R_a^2 indicates better predictive performance of the model.

Validation & Testing (R_a^2, MSPE)

Model validation involves splitting the data into training and validation sets, assessing how well the model generalizes to new, unseen data. This helps ensure that the chosen model performs well on out-of-sample data, preventing overfitting and improving reliability in real-world applications.

Adjusted R^2 (R_a^2) See above.

Mean Squared Prediction Error (MSPE) measures the average squares difference between predicted and actual values on unseen data. A low MSPE indicates that the model is making accurate predictions on out-of-sample data, highlighting its practical utility.

Multiple Regression Analysis

Preliminary EDA

To begin EDA, we looked at frequency plots for every potential predictor to get an idea of their distributions and look for any places where we could implement feature engineering. The frequency plots for most variables were relatively normal, occasionally with some skew or heavy skew. View the Appendix for further insight.

We also looked at scatter plots for every predictor versus our dependent variable, cancer death rate, to check for linearity. From here we can see some extreme values in **`medianage`** (row 2, column 5) and **`avganncount`** (row 1, column 1) that we will look to address. Both **`geography`** (row 3, column 3) and **`binnedInc`** (row 2, column 4) have a strange axis, treating values as unique categories. Both **`geography`** and **`binnedInc`** will be manipulated in a way we find meaningful. Quite a few plots show relationships between the predictor and dependent variable, but there is a potential for influential points in some cases.

Feature Engineering

We needed to do some feature engineering to make some variables usable, remove some null values found in columns, and remove any values of a variable that don't make sense or are impossible in the context of the problem, i.e. median age of a counties population being over 150.

Three variables had null values randomly in occasional entries. To handle this, we decided to impute the missing values using different strategies based on the distribution and nature of each variable. For **pctSomeCol18_24**, since this represents a demographic feature and is likely to follow a certain distribution across counties, we replaced the null values with the **mean** of the respective column, as this was justifiable given its normal or near-normal distribution. However, for **pctEmployed16_Over**, employment rates can vary significantly by region, so we opted for **median** imputation instead of the mean, as this approach is more robust against potential outliers. Finally, for **pctPrivateCoverageAlone**, since health coverage percentages also tend to vary significantly by region, we used **median** imputation to account for any regional outliers while still providing a reasonable estimate for missing values.

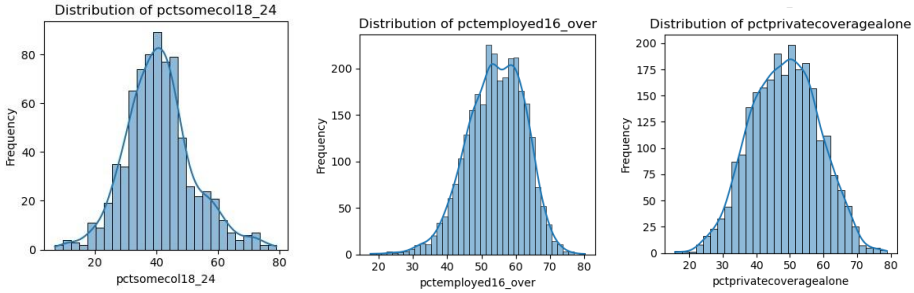


Figure 1: Distribution of pctsomecol_24, pctemployeed16_over and pctprivatecoveragealone

The variable **`medianAge`** had entries that did not make sense, such as having a median age for a county of over 200. To handle this we decided to replace any value over 150 with the mean of the variable. This again was justified because the values below 150 followed a normal distribution and there are few outliers that we don't expect an impact on the distribution of the dataset after this adjustment.

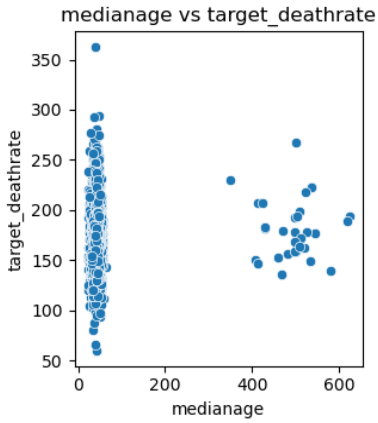


Figure 2: Scatterplot of medianage vs target_deathrate

The variable **`avgAnnCount`** which represents the mean number of reported cases of cancer diagnosed annually had some extreme outliers that were unnecessarily large so we decided to crop the variable at 15,000. We could have cropped further to get closer to a reasonable percentile but we felt that this value would eliminate the most extreme values while preserving some of the extremes in the column.

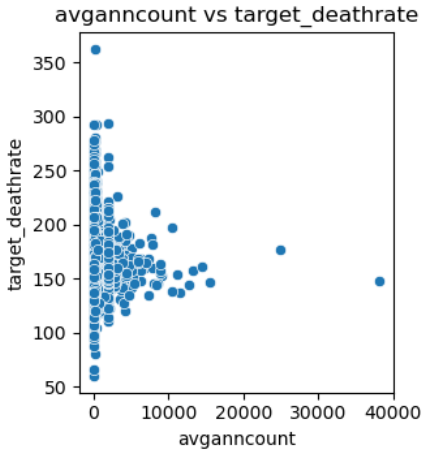


Figure 3: Scatterplot of avganncount vs. target_deathrate

The **`geography`** variable in the original dataset represented a county and state in the form ‘County, State’ because every row had a unique value in this column there was no use in using it. To make it usable we used the US geographical census regions of Northeast, Midwest, South, and West to add a new value for each entry of a geographical region by mapping the state of the county to the region it belongs to. This allowed us to create a categorical variable from this new column that correlated with our dependent variable.

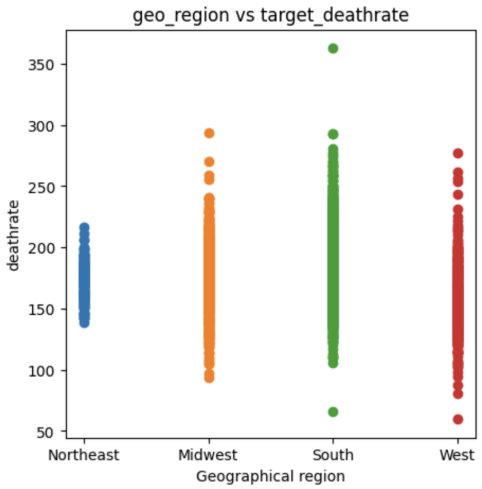


Figure 4: Scatterplot of region vs target_deathrate

The **`binnedInc`** variable represented the median income per capita binned by decile from the 2013 census of the county. This gave us a variable with a lower and upper bound value in each entry, this was simple enough to split into two individual variables for the binnedInc lower bound and the binnerInc upper bound values.

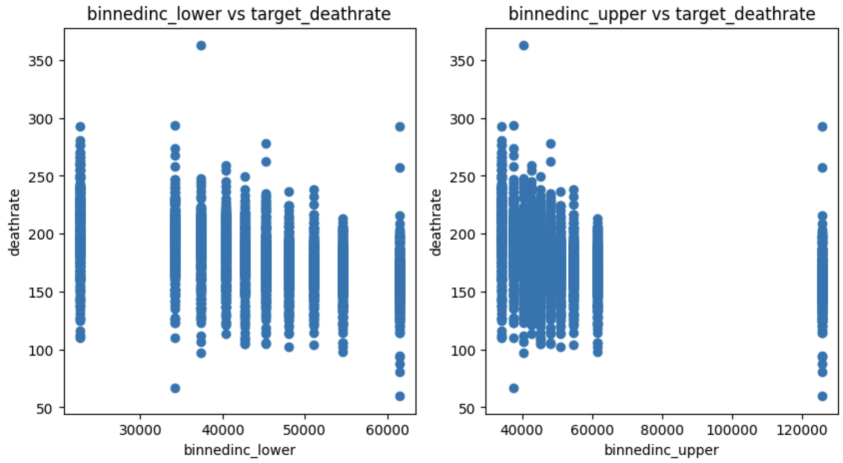


Figure 5: Scatterplots of binedInc_upper and binnedInc_lower vs. target_deathrate

We decided that categorizing the **`medincome`**, median income of the county, variable would also be viable because then it would be easier to interpret the meaning of the column. It's difficult to interpret individual values on their own and we thought there would be a more clear relationship between death rate and low, medium, high, and very high buckets for the median income of a county. We used the 30, 70, and 95th percentiles as the divisions between the groupings.

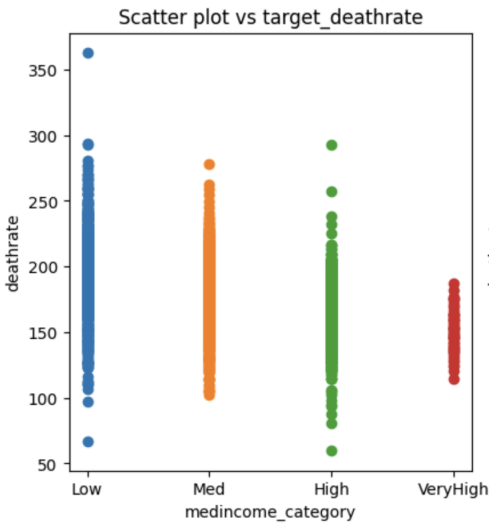


Figure 6: Scatterplot of Medianincome_category vs target_deathrate

Multicollinearity

To assess every predictor further in relation to the target dependent variable we calculated the correlation values for all of them. Traditionally above 0.7 correlation is considered highly correlated, between 0.3 and 0.7 is considered to have medium correlation and below 0.3 is exhibiting low correlation. The lower bound of 0.3 felt a bit restrictive on our dataset so we adjusted the cut-off from 0.3 to 0.25. This will allow more variables to fit into the group of having medium correlation, adjusting this value changes the number of variables with medium or high correlation from 13 to 21. Because the majority of our predictors did not show a high correlation with our response variable, it is reasonable to include predictors that show moderate linearity in their scatter plots, to allow for a more thorough model selection process. See the Appendix for a full correlation table.

Variance Inflation Factor (VIF)

Multiple variables in this dataset have some relationship with the other so we wanted to look at specific cases of that and check their VIF values to see if they exhibit multicollinearity. The groupings of closely related variables are the percentage of race in the population, median age between genders, percentages of education levels between ages, percentage of employment and unemployment above 16 years old in the population, and percentage of different insurance coverage types in the population.

Racial Percentage of the Population

Taking a look at the variance inflation factor between the percentage of the population that is white, black, asian and other we find that there are very low VIF scores between them suggesting they aren't multicollinear.

VIF	Feature
1.455521	pctwhite
1.195237	pctblack
1.252887	pctasian
1.328156	pctotherrace

Table 1: VIF amongst race variables

Median Age by Genders

Taking a look at the VIF values between **median age**, **median age male**, and **median age female** we find there are extremely high VIF values and we should look into removing one or multiple of these values. We decided that removing both median age male and median age female was the best option because we didn't feel those values added anything significant that median age wasn't already supplying.

VIF	Feature
3100.500922	medianage
1335.697619	medianagemale
1129.906684	medianagefemale

Table 2: VIF amongst age variables

Education Level Percentages by Age

Looking at the VIF values between different levels of degree at different age ranges we find decently high VIF values suggesting some multicollinearity and prompting us to remove some values. We removed the percentage highschool 18-24 (**pcths18_24**), percentage of some college 18-24 (**pctsomecol18_24**), percentage highschool 25 and over (**pcths25_over**), percentage bachelor's degree 25 and over (**pctbachdeg25_over**) the reason for removing those ones in particular was they had the highest VIF values in this test.

VIF	Feature
6.338839	pctnohs18_24
20.004559	pcths18_24
5.013162	pctsomecol18_24

42.697223	pcths25_over
15.445700	pctbachdeg25_over

Table 3: VIF amongst educations variables

Employment and Unemployment Percentage for Population Over 16

The VIF values for percent **`pctemployed16_over`** and **`pctunemployed16_over`** did not show large signs of multicollinearity when considering the VIF values obtained. In general, we are always hoping to stay below that level 5 VIF score and although these were close we don't think it makes sense to remove them at this point.

VIF	**Feature**
4.04912	pctemployeed16_over
4.04912	pctunemployed16_over

Table 4: VIF amongst employment variables

Outliers and Leverage Points

We removed outliers in two variables during our EDA process. The two variables that we removed are median age (**medianage**) and average annual count of cases (**avgAnnCount**). The outliers in median age were incorrect where multiple counties had median ages of the population over 150 years old, suggesting a data entry error. The extremely high values in average annual count were cropped down to a lower value of 15,000 which was still above the 99 quantile of ~6,000. The values cropped remain significantly large but they were providing more skew than we wanted in our dataset at values well above 15,000.

We then went into looking at influential points. To do this we fit the full model, including every predictor we had at this point into a model. Once we had this model we looked at the influential points from our model and got the Cook's distance for each point. To determine if the point was influential we compared Cook's distance of the point against the Cook's Distance threshold value of $\frac{4}{n}$ where n is the number of entries in our dataset. If the absolute value of a point Cook's distance is larger than the threshold value we would consider it an influential point and remove the point from our dataset. We recognize that to fully justify removing these points we should investigate each point individually. This is acknowledged in our potential problems section.

Cook's D Statistic	**Value**
Threshold : $\frac{4}{n}$	1.31e-03

Mean	4.88e-04
Min	9.77e-11
25% quantile	1.43e-05
50% quantile	7.91e-05
75% quantile	2.97e-04
Max	1.16e-01

Table 5: Quantile Values for Cook's D Statistic

Heteroskedasticity

We assessed the presence of heteroskedasticity by analyzing residuals against various numeric features in the dataset. By first extracting the residuals and fitted values from our initial model, we then fit a locally weighted regression line to help visualize the pattern in residuals. We looked for any patterns in the residuals, which would indicate the presence of heteroskedasticity, such as a fan shape or increasing spread. See the Appendix for full variable assessment.

Box-Cox Transformation

To address areas of heteroskedasticity we observed, we applied a Box-Cox transformation to our data, as a way to stabilize variance and make the data more normally distributed. Initially, we performed this across all predictors, simultaneously removing outliers and leverage points. See the Appendix for a full variable assessment.

However quickly realized that this transformation was only necessary on a certain subset of the data.

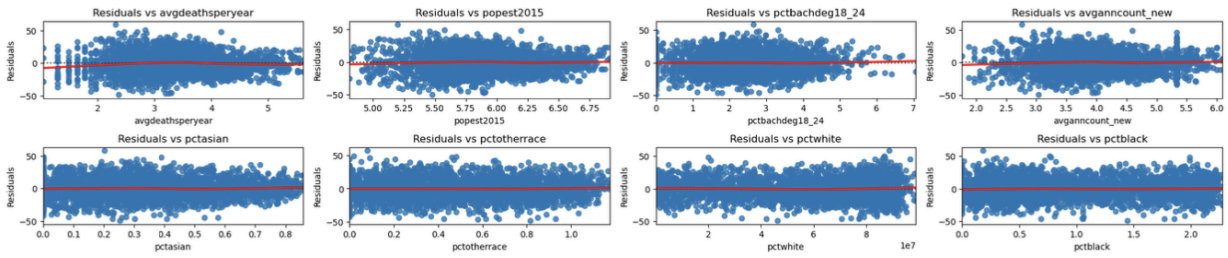


Figure 7: Box-Cox Transformations on Subset of Predictors

After selectively applying the Box-Cox transformation, due to the persistent heteroskedasticity of **studypercap**, we opted to drop this predictor.

Overview

Our dataset was updated accordingly, to account for the removal of outlier observations and transformed features. One final linearity and normality check was conducted using various diagnostic plots, before beginning to build our final model. See Appendix for full diagnostics.

Model Selection

Data Splitting

With the cleaned and preprocessed data, we now have a usable dataset to build our MLR model.

To begin, the dataset was split into training, validation, and test sets using a 70-10-20 split. 70% of the data (training) was used in model training and selection, to narrow down first order models into the top 2 best performing models, according to our criterion. 10% of the data (validation) was used for model selection to determine the overall best performing model. The remaining 20% of the data (test) was used for the final model evaluation.

Stepwise Regression

Two stepwise regression functions were defined to perform variable selection based on AIC and BIC. These processes were performed iteratively until no further improvement was observed.

Forward Selection
Starting with no variables, predictors are added one by one based on which addition most improves the model according to AIC/BIC.

Backwards Elimination
Starting with all candidate variables, predictors are removed one by one based on which removal most improves the model according to AIC/B

AIC Model

```
target_deathrate ~ binnedinc_lower + incidencerate +
geo_region_South + geo_region_West + geo_region_Northeast +
pctunemployed16_over + pctempprivcoverage + pctprivatecoverage +
pctemployed16_over + avghouseholdsize_category_Med +
pctpubliccoverage + pctwhite + popest2015 +
medincome_category_VeryHigh + pctnohs18_24 + percentmarried +
pctmarriedhouseholds + medianage + avghouseholdsize_category_High
+ povertypercent
```

OLS Regression Results

Dep. Variable:	target_deathrate	R-squared:	0.596
Model:	OLS	Adj. R-squared:	0.592
Method:	Least Squares	F-statistic:	168.9
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.00
Time:	21:17:38	Log-Likelihood:	-7774.0
No. Observations:	1852	AIC:	1.558e+04
Df Residuals:	1835	BIC:	1.568e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	150.9920	11.671	12.937	0.000	128.102	173.882
binnedinc_lower	-0.0004	7.58e-05	-5.001	0.000	-0.001	-0.000
incidencerate	0.2037	0.009	23.922	0.000	0.187	0.220
geo_region_South	2.8429	1.105	2.573	0.010	0.676	5.010
geo_region_West	-14.7845	1.368	-10.809	0.000	-17.467	-12.102
pctprivatecoverage	-0.7805	0.101	-7.694	0.000	-0.979	-0.582
pctempprivcoverage	0.8004	0.096	8.359	0.000	0.613	0.988
pctemployed16_over	-0.7051	0.108	-6.527	0.000	-0.917	-0.493
geo_region_Northeast	-9.8296	1.658	-5.930	0.000	-13.080	-6.579
pctpubliccoverage	0.3052	0.130	2.345	0.019	0.050	0.560
pctunemployed16_over	0.4657	0.174	2.676	0.008	0.124	0.807
avghouseholdsize_category_Med	3.4864	1.066	3.269	0.001	1.395	5.578
medincome_category_VeryHigh	-5.0320	2.838	-1.773	0.076	-10.597	0.533
percentmarried	1.0122	0.190	5.335	0.000	0.640	1.384
pctmarriedhouseholds	-0.8837	0.182	-4.843	0.000	-1.242	-0.526
medianage	-0.3433	0.130	-2.650	0.008	-0.597	-0.089
avghouseholdsize_category_High	2.6935	1.399	1.925	0.054	-0.051	5.438

Omnibus:	5.768	Durbin-Watson:	2.025
Prob(Omnibus):	0.056	Jarque-Bera (JB):	5.679
Skew:	0.130	Prob(JB):	0.0585
Kurtosis:	3.077	Cond. No.	1.40e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.4e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 9: OLS Summary Output for AIC Model

Note that, in the AIC model (refer to Figure 9), the p-values for medianincome_category_VeryHigh and avghouseholdsize_category_High exceed 0.05. This suggests that there may not be a significant relationship between high median income and high average household size with the target death rate. However, we chose to retain these variables in the model, as their p-values were only slightly above 0.05.

BIC Model

```
target_deathrate ~ binnedinc_lower + incidencerate +
geo_region_West + geo_region_Northeast + pctunemployed16_over +
pctempprivcoverage + pctprivatecoverage + pctemployed16_over +
avghouseholdsize_category_Med
```

OLS Regression Results

Dep. Variable:	target_deathrate	R-squared:	0.586
Model:	OLS	Adj. R-squared:	0.584
Method:	Least Squares	F-statistic:	289.2
Date:	Sat, 12 Oct 2024	Prob (F-statistic):	0.00
Time:	21:27:40	Log-Likelihood:	-7796.4
No. Observations:	1852	AIC:	1.561e+04
Df Residuals:	1842	BIC:	1.567e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	160.1410	5.969	26.827	0.000	148.433	171.848
binnedinc_lower	-0.0005	6.78e-05	-6.701	0.000	-0.001	-0.000
incidencerate	0.2164	0.008	26.281	0.000	0.200	0.233
geo_region_West	-16.3348	1.176	-13.892	0.000	-18.641	-14.029
pctprivatecoverage	-0.8690	0.082	-10.626	0.000	-1.029	-0.709
pctempprivcoverage	0.5927	0.082	7.209	0.000	0.431	0.754
pctemployed16_over	-0.5269	0.082	-6.440	0.000	-0.687	-0.366
geo_region_Northeast	-12.2687	1.495	-8.209	0.000	-15.200	-9.337
pctunemployed16_over	0.4847	0.163	2.981	0.003	0.166	0.803
avghouseholdsize_category_Med	2.3477	0.817	2.874	0.004	0.745	3.950

Omnibus:	5.382	Durbin-Watson:	2.026
Prob(Omnibus):	0.068	Jarque-Bera (JB):	5.295
Skew:	0.127	Prob(JB):	0.0708
Kurtosis:	3.067	Cond. No.	7.08e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.08e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 10: OLS Summary Output for BIC Model

ANOVA Tests

An ANOVA type-2 test was conducted on the full model to identify variables that were not statistically significant, i.e. p-value > 0.05).

Insignificant Predictors

```
avgdeathsperyear, popest2015, povertypercent, pctnohs18_24,
pctbachdeg18_24, pctprivatecoveragealone,
pctpubliccoveragealone, pctwhite, pctblack, pctasian,
pctotherrace, birthrate, avganncount_new, binnedinc_upper,
medincome_category_Med, medincome_category_High
```

These variables were removed from the set of predictors for subsequent model selection.

Refined Stepwise Regression

With the reduced set of predictors, the stepwise selection process was repeated. From this, we determined the best performing models according to AIC and BIC.

AIC Model

```
target_deathrate ~ binnedinc_lower + incidencerate +
geo_region_South +
geo_region_West + pctprivatecoverage + pctempprivcoverage +
pctemployed16_over + geo_region_Northeast + pctpubliccoverage +
pctunemployed16_over + avghouseholdsize_category_Med +
medincome_category_VeryHigh + percentmarried +
pctmarriedhouseholds + medianage + avghouseholdsize_category_High
```

BIC Model

```
target_deathrate ~ binnedinc_lower + incidencerate +
geo_region_West +
pctprivatecoverage + pctempprivcoverage + pctemployed16_over +
geo_region_Northeast + pctunemployed16_over +
avghouseholdsize_category_Med
```

Validation

After defining our best performing models, we conducted validation set evaluation using MSPE and Adjusted R^2.

AIC Model	**BIC Model**
MSPE: 243.34143969959075 R_a^2: 0.6141829399196173	*MSPE:* 249.0320624465023 R_a^2: 0.6109761701099814

The best performing model according to both criteria is the one determined by AIC. This will be our final model.

Testing

Using our final model, we conducted a test set evaluation using MSPE and Adjusted R^2.

Final Model
MSPE: 277.7587508650746 R_a^2: 0.5736617562231975

Discussion

Summary of Findings

Our analysis found that the strongest factors in predicting cancer mortality rates were the minimum income of people in the particular decile, the average cancer diagnoses per capita, the states from southern United States, the states from western United States, the percent of county residents with private health coverage, the percent of county residents with employee-provided private health coverage, the percent of county residents ages 16 and over-employed, the states from northeastern United States, the percent of county residents with government-provided health coverage, the percent of county residents ages 16 and over unemployed, the average household size from 33.33rd percentile till 66.67th percentile, the median income category that falls between the 98th percentile and the maximum median income, the percent of county residents who are married, the percent of married households, the median age of county residents, and average household size from 66.67th percentile till maximum average household size.

Our evaluation metrics allow us to assess the performance of our model. The MSPE value indicates that the model's predictions are reasonably accurate, though some deviations from the actual values still exist. The adjusted R^2 value indicates that approximately 57.37% of the variance in cancer mortality rates across counties can be explained by the model's predictors. While this is a moderate level of explanatory power, there is still room for improvement.

A scatterplot comparing actual versus predicted values further explains the model's predictive performance. The ideal fit line, where predicted equals actual, is plotted in red. The model's predictions follow this line, though with some spread. This plot underscores that while the model captures the general trend in cancer mortality rates, there are deviations in individual predictions, suggesting potential sources of error that the model does not account for.

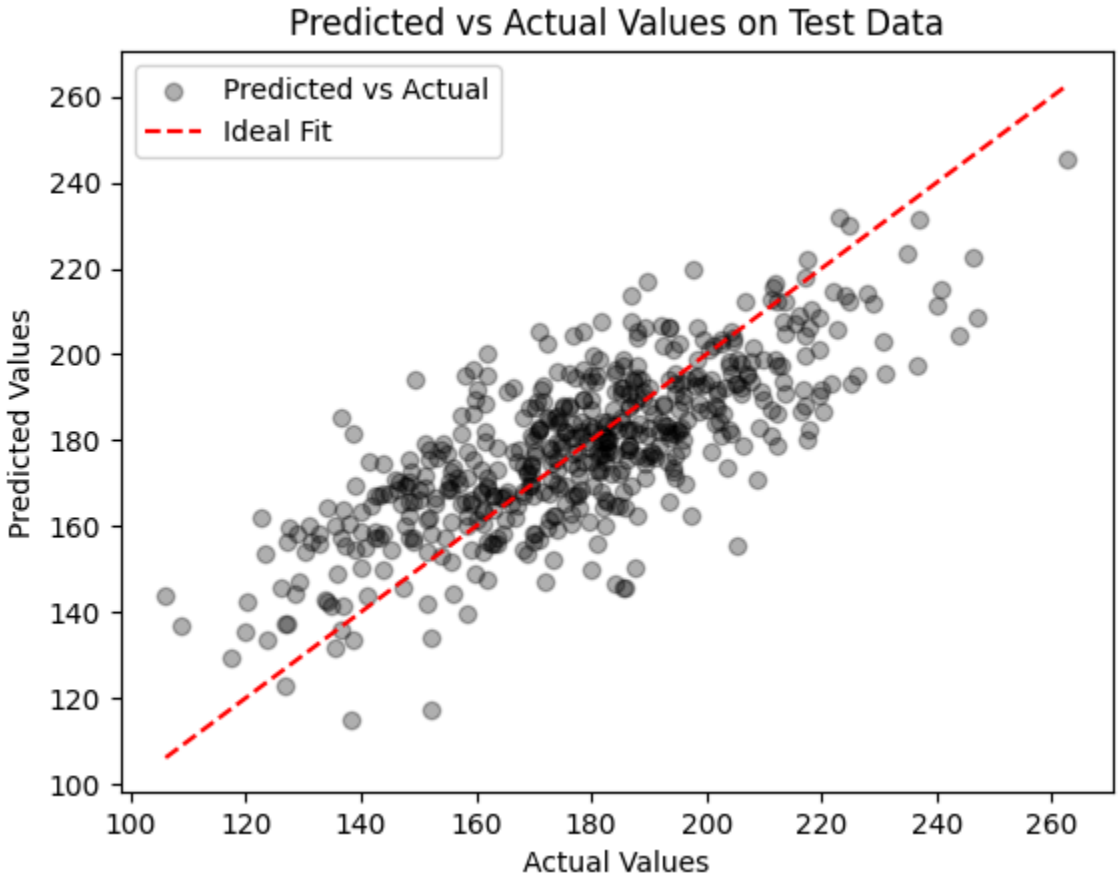


Figure 11: Predicted vs. Actual Values on Test Data with Model Fit Line

The scatterplot below of the residuals versus fitted values shows a generally well-fitting regression model. The residuals appear to be centered around zero, indicating that the model's assumptions of linearity are likely met. There is no clear pattern or trend, which suggests that the model is not overfitting or underfitting the data. Additionally, the spread of residuals seems relatively constant across the range of fitted values, supporting the assumption of homoscedasticity. There are no significant outliers or clusters that could point to problematic

predictions. However, there is a slight variation in the residual spread at the ends. Overall, this diagnostic plot suggests that the model is performing well and provides a good fit to the data.

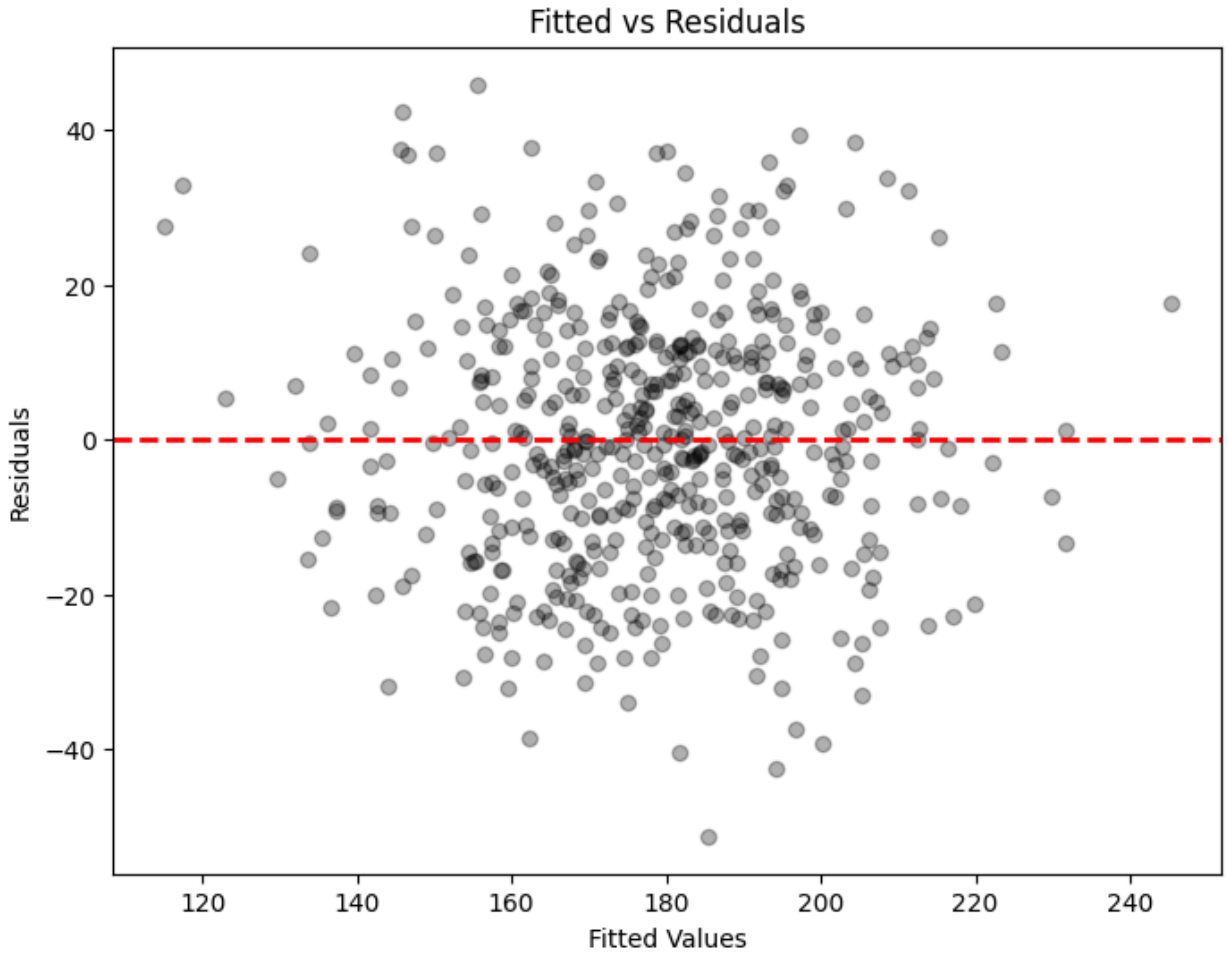


Figure 12: Fitted Values vs. Residuals on Test Data

Potential Problems

Three key issues emerged that raise concerns about the predictive power of our model.

First, during the initial EDA process, we observed minimal correlation between individual predictors and the response variable, which is problematic because it indicates that the chosen predictors may not have strong linear relationships with the outcome. This lack of correlation can lead to a weak explanatory model, where the predictors collectively do not account for a significant proportion of the variance in the response variable. Even if the final model fits the training data well, it may fail to capture the underlying patterns that genuinely influence the response variable, leading to poor generalizability.

Second, while our decision to remove influential points entirely was partially motivated by clear input errors, such as an average country age of 400, it was improper practice to remove all influential points without individually investigating each one.

Lastly, when evaluating the model on the test data, we noticed a decrease in the Adjusted R^2 value compared to the training and validation sets. This suggests potential overfitting, where the model has become overly complex by capturing noise in the training data rather than meaningful trends. Overfitting reduces the model's ability to generalize to unseen data, as the model performs well on the training set but struggles to make accurate predictions on new data, undermining its predictive reliability.

Real-Life Context

Cancer is the leading cause of death in the United States, and mortality rates vary significantly across different countries. This project has the potential to provide insights into several areas.

(1) *Healthcare Accessibility*: Cancer mortality rates differ across regions due in part to varying access to healthcare, demographic differences, socioeconomic factors, and lifestyle habits. This project could help identify particularly vulnerable counties.
(2) *Policy Making*: Government bodies and organizations could use data like this to craft policies aimed at reducing cancer mortality. Predictive models help in forecasting future cancer trends, which informs long-term planning for healthcare services, research funding, and preventative measures.
(3) *Public Health Interventions:* By understanding which factors contribute most to high cancer mortality, public health officials can allocate resources more effectively. This project could support targeted interventions, screening programs, and education efforts in high-risk areas.

Next Steps

There are a variety of strong follow-ups to this project that could further explore cancer-related data in a predictive model. This would involve delving deeper into the variables that influence not just mortality rates, but the effectiveness of different cancer treatments across various demographics and regions. Specific models include:

(1) *Machine Learning Models*, such as random forests, gradient boosting machines, or neural networks, which can be used to model more complex non-linear interactions between variables that our current model might miss
(2) Survival Analysis models, which can be used to predict the time until an event (death, remission, etc.)

Appendices

A1: Data Dictionary

(a): 2010-2016 County Averages

(b): 2013 County Census Estimates

TARGET_deathRate: Dependent variable. Mean *per capita* (100,000) cancer mortalities (*a*)

avgAnnCount: Mean number of reported cases of cancer diagnosed annually (*a*)

avgDeathsPerYear: Mean number of reported mortalities due to cancer (*a*)

incidenceRate: Mean *per capita* (100,000) cancer diagnoses (*a*)

medianIncome: Median income per county (*b*)

popEst2015: Population of county (*b*)

povertyPercent: Percent of populace in poverty (*b*)

studyPerCap: *Per capita* number of cancer-related clinical trials per county (*a*)

binnedInc: Median income per capita binned by decile (*b*)

MedianAge: Median age of county residents (*b*)

MedianAgeMale: Median age of male county residents (*b*)

MedianAgeFemale: Median age of female county residents (*b*)

Geography: County name (*b*)

AvgHouseholdSize: Mean household size of county (*b*)

PercentMarried: Percent of county residents who are married (*b*)

PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school (*b*)

PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma (*b*)

PctSomeCol18_24: Percent of county residents ages 18-24 highest education attained: some college (*b*)

PctBachDeg18_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree (*b*)

PctHS25_Over: Percent of county residents ages 25 and over highest education attained: high school diploma (*b*)

PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree (*b*)

PctEmployed16_Over: Percent of county residents ages 16 and over employed (*b*)

PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed (*b*)

PctPrivateCoverage: Percent of county residents with private health coverage (*b*)

PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance) (*b*)

PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage (*b*)

PctPublicCoverage: Percent of county residents with government-provided health coverage (*b*)

PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone (*b*)

PctWhite: Percent of county residents who identify as White (*b*)

PctBlack: Percent of county residents who identify as Black (*b*)

PctAsian: Percent of county residents who identify as Asian (*b*)

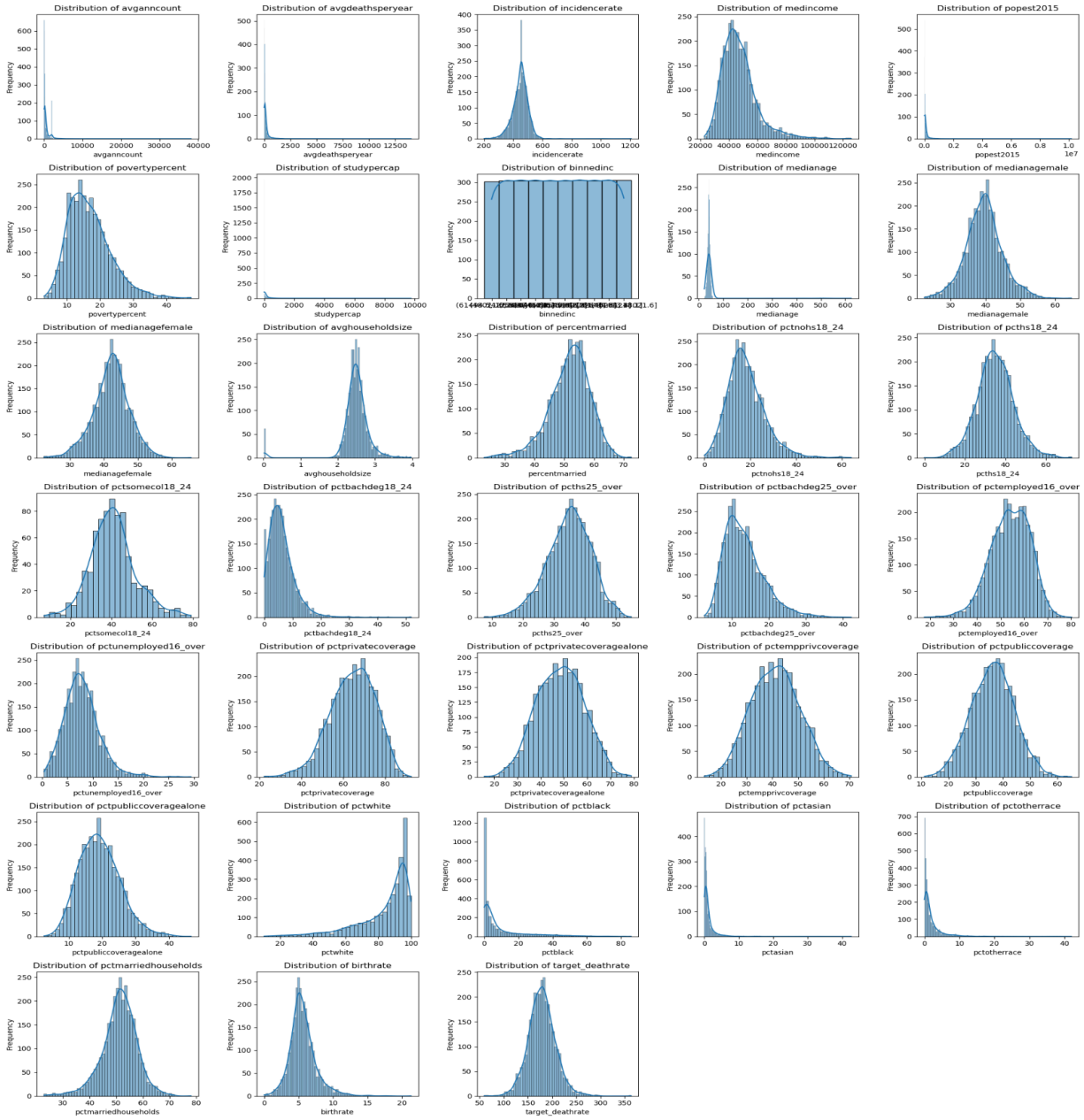
PctOtherRace: Percent of county residents who identify in a category that is not White, Black, or Asian (*bg*)

PctMarriedHouseholds: Percent of married households (*b*)

BirthRate: Number of live births relative to the number of women in the county (*b*)

A2: EDA Supplementals

A2 I: Preliminary Distributions



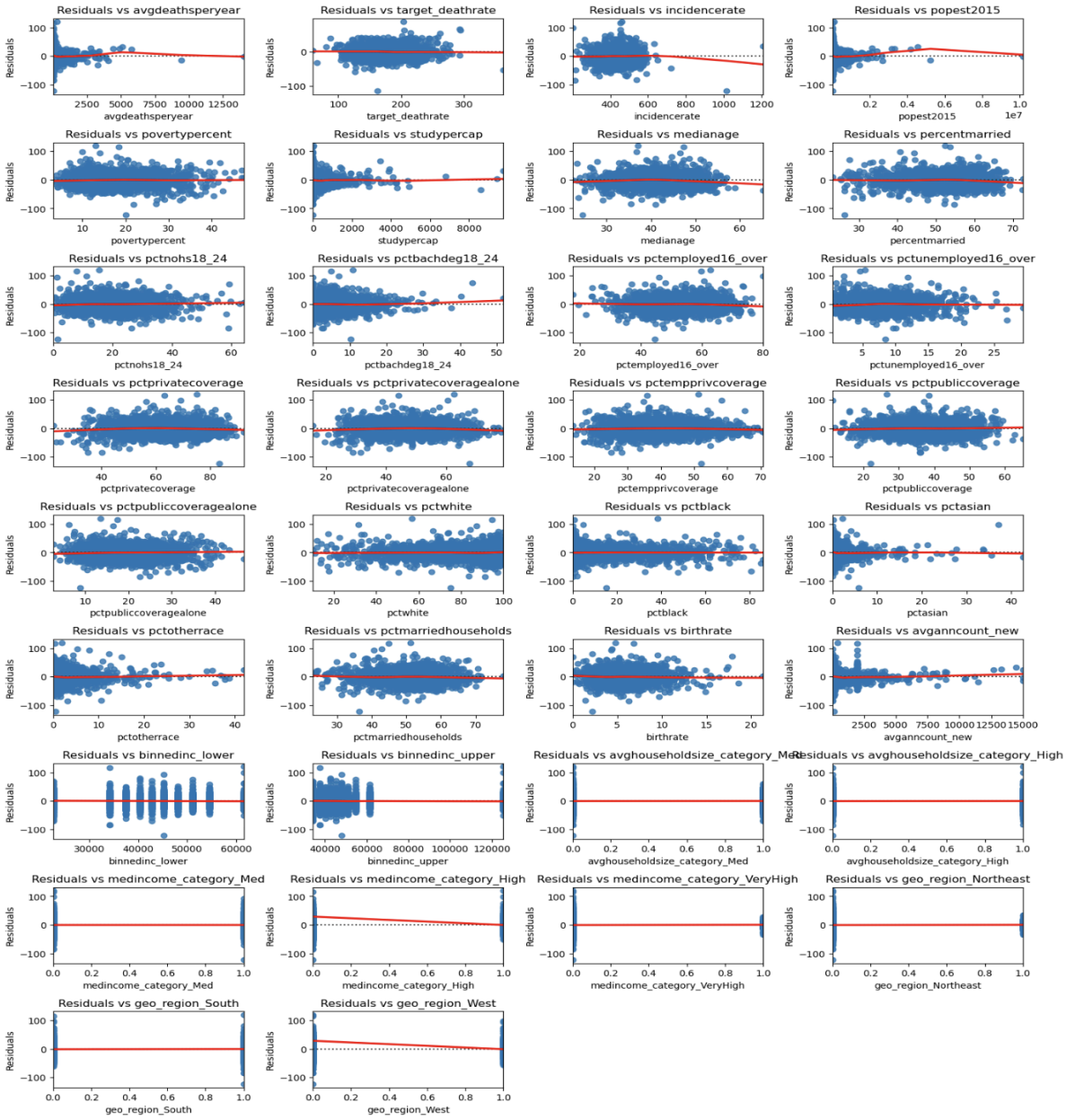
A2 II: Correlation Table

Variable Name	Correlation	Strength
target_deathrate	1	strong
incidencerate	0.4494316976	medium

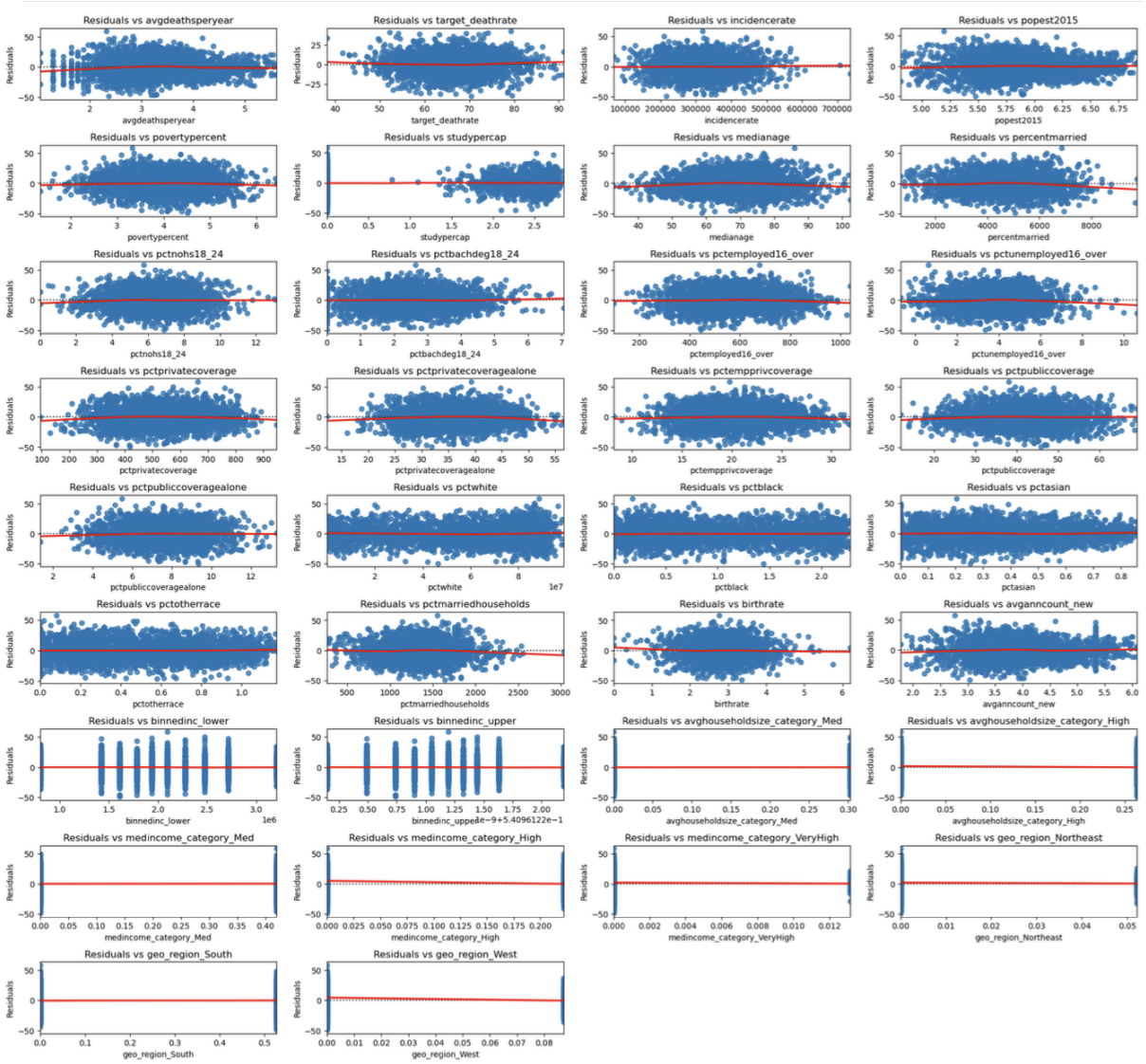
pctpubliccoveragealone	0.4493575758	medium
povertypercent	0.4293889803	medium
pcths25_over	0.4045890758	medium
pctpubliccoverage	0.4045716563	medium
pctunemployed16_over	0.3784124421	medium
geo_region_South	0.338125641	medium
pcths18_24	0.2619759403	medium
pctblack	0.2570235605	medium
avghouseholdsize_category_Med	0.1680614862	weak
pctnohs18_24	0.08846261004	weak
medianagefemale	0.01204838597	weak
medianage	-0.003763881028	weak
medianagemale	-0.02192942908	weak
studypercap	-0.02228501077	weak
medincome_category_Med	-0.03178675318	weak
geo_region_Northeast	-0.06124929722	weak
avghouseholdsize_category_High	-0.06524371737	weak
birthrate	-0.08740696984	weak
avgdeathsperyear	-0.09071515999	weak
pctsomecol18_24	-0.09490143372	weak
popest2015	-0.1200730957	weak
medincome_category_VeryHigh	-0.1454236912	weak
avganncount_new	-0.1599885547	weak
pctwhite	-0.1773999803	weak
pctasian	-0.1863311045	weak
pctotherrace	-0.1898935711	weak
percentmarried	-0.2668204636	medium
pctempprivcoverage	-0.2673994282	medium
pctbachdeg18_24	-0.2878174102	medium
medincome_category_High	-0.2892676824	medium
pctmarriedhouseholds	-0.2933253405	medium
geo_region_West	-0.2962657209	medium
pctprivatecoveragealone	-0.326067207	medium
binnedinc_upper	-0.3371848431	medium
pctprivatecoverage	-0.3860655068	medium

pctemployed16_over	-0.3974325408	medium
binnedinc_lower	-0.4480561831	medium
pctbachdeg25_over	-0.4854773181	medium

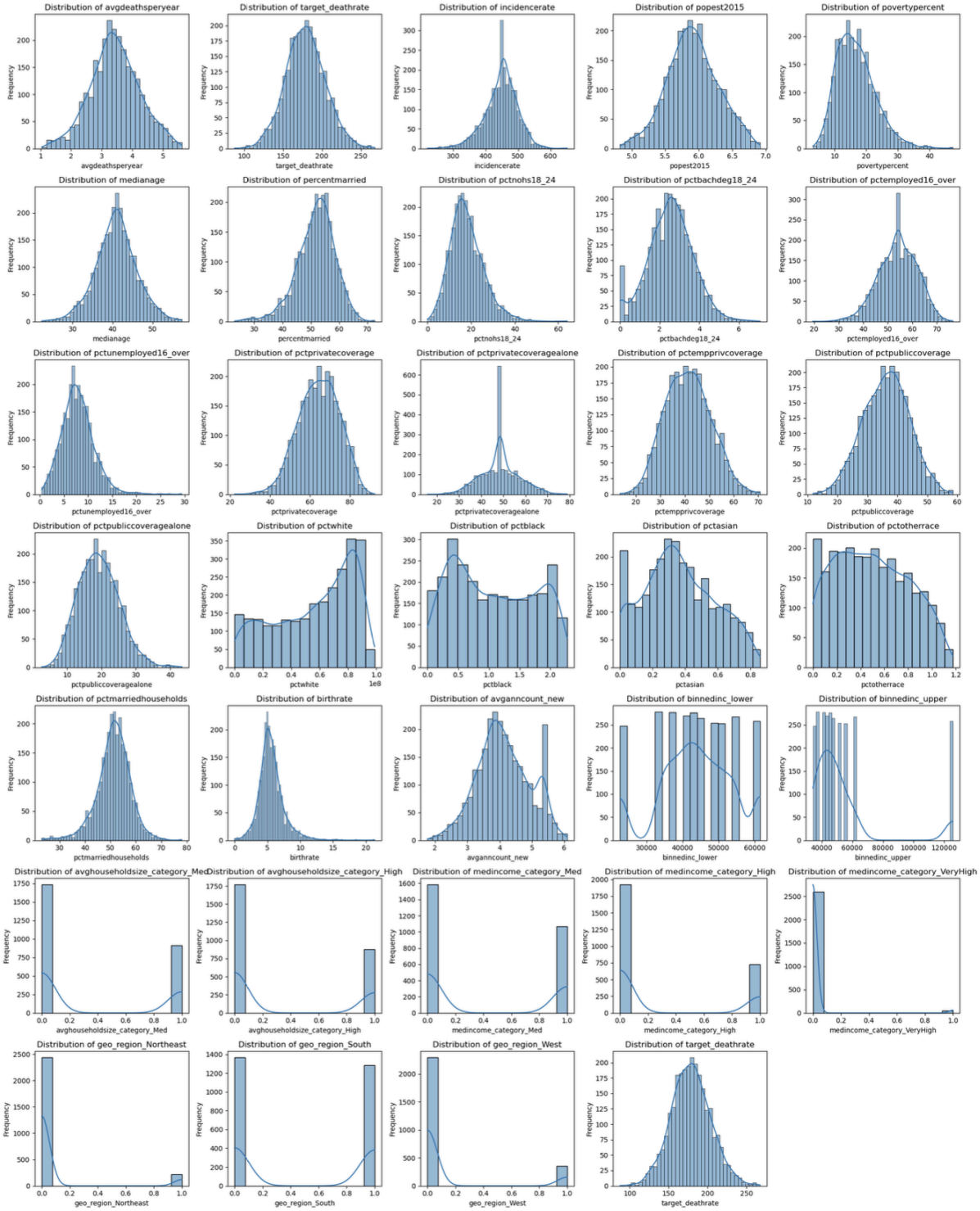
A2 III: Heteroskedasticity Plots

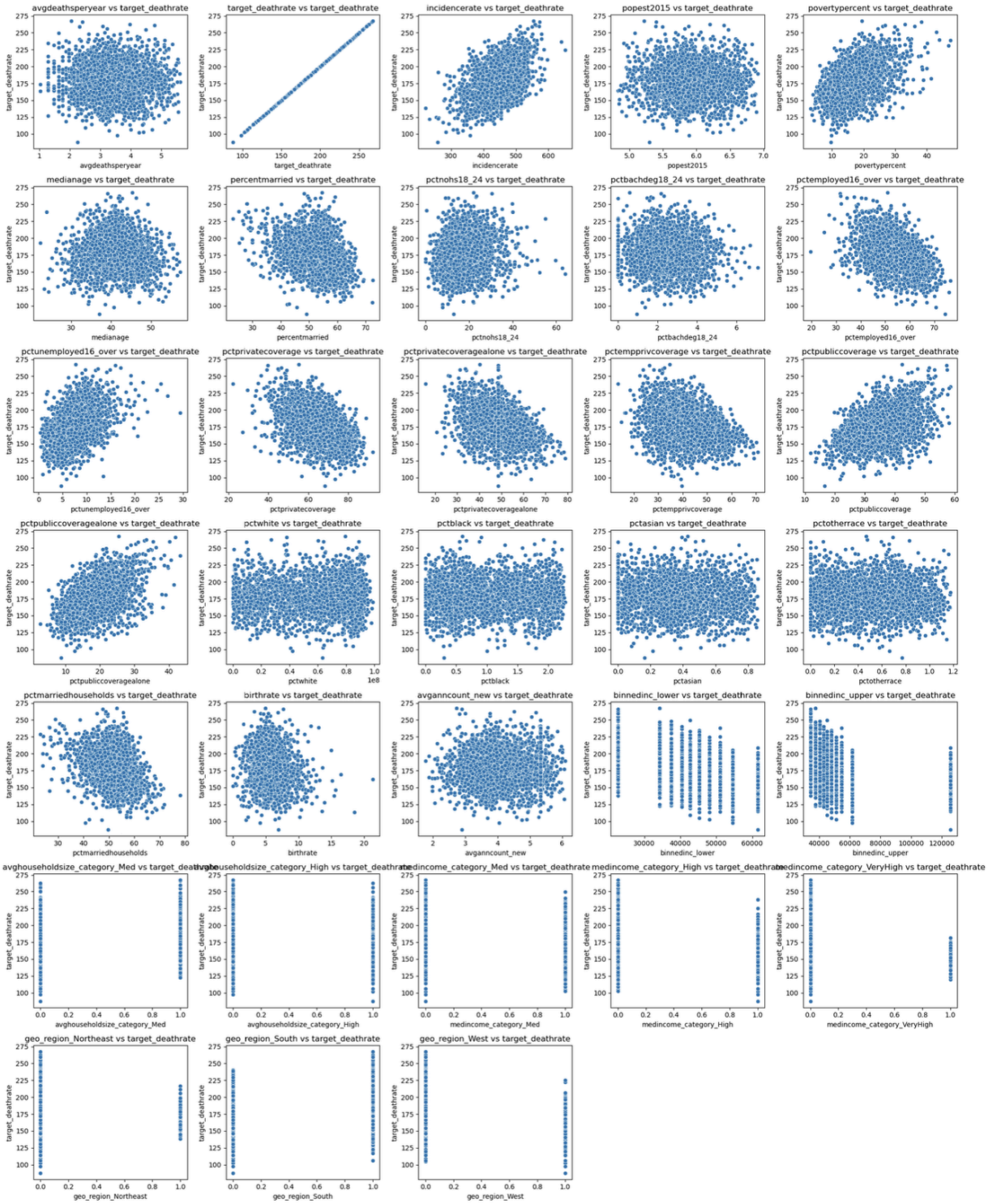


A2 IV: Box-Cox Plots



A2 V: Final Diagnostics




avgdeathsperyear vs target_deathrate
target_deathrate vs target_deathrate
incidencerate vs target_deathrate
popest2015 vs target_deathrate
povertypercent vs target_deathrate
medianage vs target_deathrate
percentmarried vs target_deathrate
pctnohs18_24 vs target_deathrate
pctbachdeg18_24 vs target_deathrate
pctemployed16_over vs target_deathrate
pctunemployed16_over vs target_deathrate
pctprivatecoverage vs target_deathrate
pctprivatecoveragealone vs target_deathrate
pctempprivcoverage vs target_deathrate
pctpubliccoverage vs target_deathrate
pctpubliccoveragealone vs target_deathrate
pctwhite vs target_deathrate
pctblack vs target_deathrate
pctasian vs target_deathrate
pctotherrace vs target_deathrate
pctmarriedhouseholds vs target_deathrate
birthrate vs target_deathrate
avganncount_new vs target_deathrate
binnedinc_lower vs target_deathrate
binnedinc_upper vs target_deathrate
avghouseholdsize_category_Med vs target_deathrate
avghouseholdsize_category_High vs target_deathrate
medincome_category_Med vs target_deathrate
medincome_category_High vs target_deathrate
medincome_category_VeryHigh vs target_deathrate
geo_region_Northeast vs target_deathrate
geo_region_South vs target_deathrate
geo_region_West vs target_deathrate
target_deathrate

Breakdown of Work

Raghava Srinivasan

- EDA
- Data cleaning & engineering

Vani Singh

- Model Selection
- Model Building

Quinn Campfield

- Model selection
- Final project writeup (EDA Section)

Hadley Dixon

- MLR model starter code
- Final Project Writeup (All sections)