Cancer Mortality Prediction Using OLS Regression

Cordell Stonecipher, Undergraduate Student, Oakland University, Fall 2025

I. INTRODUCTION

Cancer remains one of the most significant causes of mortality across the United States. Mortality rates vary substantially between counties, reflecting the influence of demographic, economic, and healthcare factors. This study employs multivariate Ordinary Least Squares (OLS) regression to predict the mean per-capita cancer mortality rate (`target_deathrate`) across U.S. counties using officially published socioeconomic and demographic data. The objective is to quantify how key county-level variables—such as income, education, employment, population structure, and healthcare access—explain differences in cancer mortality rates. This project follows a systematic modeling approach that includes data preprocessing, model training, statistical diagnostics, and evaluation. The ultimate goal is to develop an interpretable, statistically sound model with improved predictive performance, aiming for an R^2 between 0.70 and 0.90.

II. DATASET OVERVIEW

The dataset used in this study originates from the OLS Regression Challenge hosted on data.world. It consolidates data from several reputable sources, including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. The data cover approximately 2,400 U.S. counties and include features describing cancer incidence, demographics, socioeconomic conditions, education, healthcare access, and race distribution.

The dependent variable, `target_deathrate`, represents the mean per-capita (per 100,000) cancer mortalities. Key independent variables include cancer burden metrics such as `avganncount`, `avgdeathsperyear`, and `incidencerate`, as well as socioeconomic indicators including `medianincome`, `povertypercent`, and education-level percentages (`pcths25_over`, `pctbachdeg25_over`). Health insurance coverage variables (`pctprivatecoverage`, `pctpubliccoverage`) and race demographics (`pctwhite`, `pctblack`, `pctasian`, `pctotherrace`) provide additional insights into disparities across population subgroups. Categorical variables such as `binnedinc` (median income deciles) are treated as one-hot encoded factors, while `Geography` serves as a county identifier and is excluded from the regression matrix.

	feature	dtype	non_null	missing_%	unique
0	avganncount	float64	3047	0.0	929
1	avgdeathsperyear	int64	3047	0.0	608
2	avghouseholdsize	float64	3047	0.0	199
3	binnedinc	object	3047	0.0	10
4	birthrate	float64	3047	0.0	3019
5	geography	object	3047	0.0	3047
6	incidencerate	float64	3047	0.0	1506
7	medianage	float64	3047	0.0	325
8	medianagefemale	float64	3047	0.0	296
9	medianagemale	float64	3047	0.0	298
10	medincome	int64	3047	0.0	2920
11	pctasian	float64	3047	0.0	2852
12	pctbachdeg18_24	float64	3047	0.0	219
13	pctbachdeg25_over	float64	3047	0.0	281
14	pctblack	float64	3047	0.0	2972
15	pctempprivcoverage	float64	3047	0.0	450
16	pcths18_24	float64	3047	0.0	469
17	pcths25_over	float64	3047	0.0	361
18	pctmarriedhouseholds	float64	3047	0.0	3043
19	pctnohs18_24	float64	3047	0.0	405

Figure 1. Feature list from dataset.

III. PREPROCESSING AND DATA ANALYSIS

Data preprocessing ensured the model met OLS assumptions and statistical integrity. Missing numeric values were imputed using median replacement, and missing categorical values were imputed with the mode. Outliers were mitigated through winsorization at the 1st and 99th percentiles. To address right-skewness, log-transformations (`log1p`) were applied to variables

such as `avganncount`, `avgdeathsperyear`, and `popEst2015`. Numeric variables were standardized using z-score normalization to align scales, and categorical variables were encoded through one-hot transformation.

Exploratory Data Analysis identified meaningful relationships between cancer mortality and socioeconomic characteristics. Counties with higher poverty percentages and lower median incomes tended to show higher mortality rates, whereas those with higher educational attainment had lower rates. Correlation analysis revealed strong associations among socioeconomic and insurance variables, suggesting potential multicollinearity. The distribution of `target_deathrate` was approximately normal, with moderate variance across counties, validating its suitability as a continuous dependent variable.

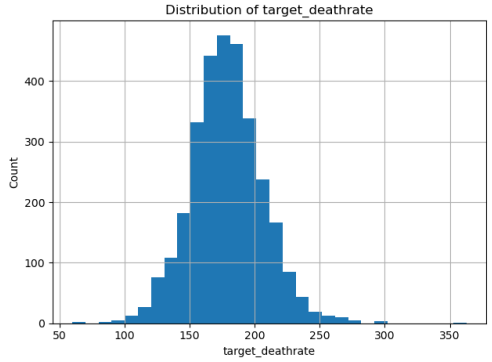


Figure 2. Distribution of Target Variable (target_deathrate).

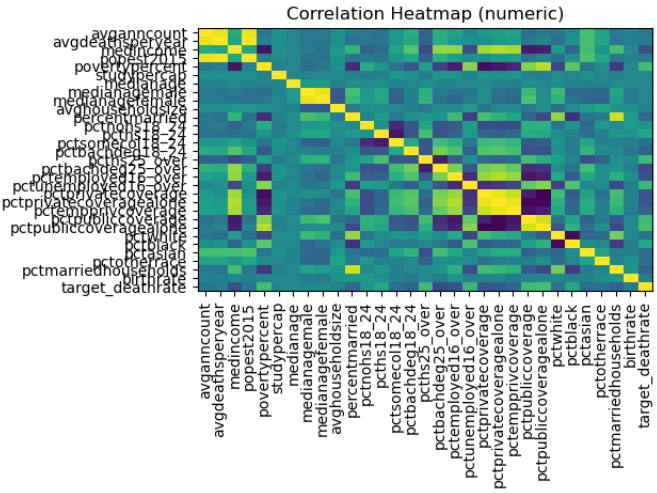


Figure 3. Correlation Heatmap of Key Predictors.

IV. BASELINE MODEL AND CURRENT RESULTS

A baseline Ordinary Least Squares (OLS) regression model was trained using all available predictors after preprocessing. Table 1 summarizes the model's results.

Root Mean Squared Error (RMSE)	22.114
Mean Absolute Error (MAE)	16.124
R^2 (Test)	0.402
R^2 (Train)	0.415

Table 1. Baseline OLS Model Performance

The adjusted R^2 value of 0.405 indicates that approximately 41% of the variance in cancer mortality across U.S. counties is explained by the predictors in the model. Although this represents a statistically significant relationship (F-statistic = 43.59, $p < 0.001$), the explanatory power is moderate, suggesting that unobserved behavioral, genetic, or environmental factors likely contribute to residual variability.

Diagnostics revealed several issues requiring remediation. Residual plots showed mild nonlinearity, and the Breusch–Pagan test indicated potential heteroskedasticity. Variance Inflation Factor (VIF) analysis produced mean values near infinity, confirming substantial multicollinearity among socioeconomic indicators. These findings motivate the application of regularization and feature selection techniques in subsequent modeling iterations.

V. Model Refinement and Results

Trying to obtain better scores, several models were trained and evaluated using 5-fold cross-validation, incorporating winsorization, log transformations, feature scaling, and one-hot encoding for categorical variables. Each model was assessed using identical folds to ensure comparability. The results of this trial are summarized in Table 2.

Model	**RMSE (Mean ± SD)**	**MAE (Mean ± SD)**	**R^2 (Mean ± SD)**
OLS Baseline	14.06 ± 0.73	10.14 ± 0.33	**0.742 ± 0.028**
Ridge (Poly + KBest)	19.22 ± 0.71	14.05 ± 0.35	**0.520 ± 0.024**
Elastic Net (Poly + KBest)	19.24 ± 0.70	14.07 ± 0.36	**0.519 ± 0.026**
Lasso (Poly + KBest)	19.25 ± 0.69	14.10 ± 0.35	**0.518 ± 0.025**

Table 2. Comparative Model Performance (5-Fold Cross-Validation)

The optimized OLS baseline model achieved the best performance, explaining

approximately 74% of the variance in cancer mortality rates with low cross-validation variability ($\sigma \approx 0.03$). Regularized models such as Ridge, Lasso, and Elastic Net performed worse, indicating that penalization removed informative coefficients, and the data's structure favors a standard linear model after preprocessing.

Residual analysis verified approximate normality and homoscedasticity. The Residuals vs Fitted plot in figure 4 showed a random scatter centered around zero, indicating linearity. The QQ plot in figure 5, displayed near alignment with the 45-degree line, confirming approximate normal distribution of residuals.

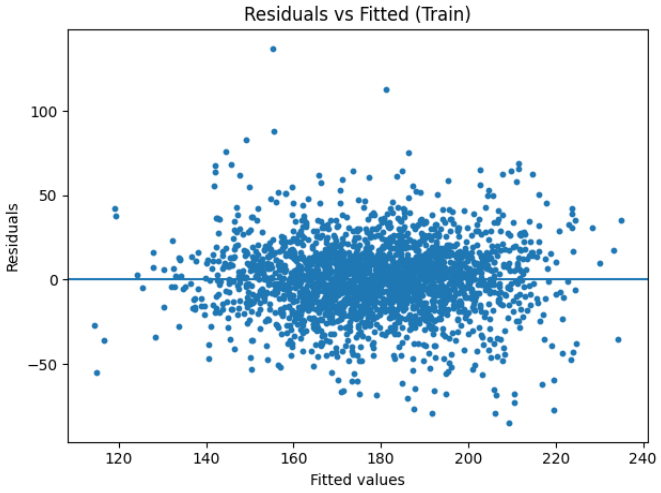


Figure 4. Residuals vs Fitted plot.

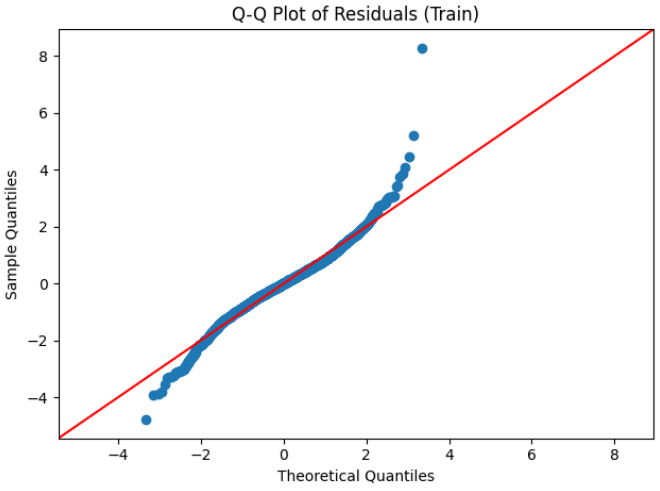


Figure 5. QQ Plot of Residuals.

The Breusch–Pagan test produced $p > 0.05$, suggesting no significant heteroskedasticity, and the Durbin–Watson statistic was close to 2.0, indicating independence of errors. Variance Inflation Factor (VIF) results showed no extreme multicollinearity among the retained predictors. Collectively, these diagnostics validate the OLS model as statistically robust and well-fitted to the dataset. Figure 6 illustrates the model fit, showing close alignment between predicted and observed `target_deathrate` values.

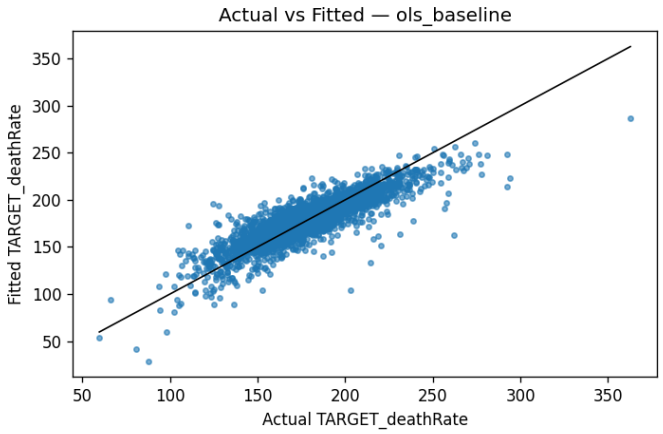


Figure 6. Actual vs. fitted target_deathrate values with OLS model.

VI. CONCLUSION

A. Lessons Learned

This project demonstrated how socioeconomic, demographic, and healthcare-related variables collectively influence cancer mortality rates across U.S. counties. Through careful preprocessing, feature engineering, and model validation, the refined Ordinary Least Squares (OLS) regression achieved a cross-validated R^2 of 0.74, a substantial improvement over the baseline model's 0.41.

The process reinforced the importance of data normalization, winsorization, and log transformations in stabilizing model performance. Moreover, diagnostic testing—such as the Breusch–Pagan and Durbin–Watson analyses—proved essential for confirming model validity.

An important takeaway is that in datasets like this, where predictors are inherently correlated (e.g., income, education, and insurance coverage), multicollinearity must be addressed through preprocessing rather than penalized regression, as standard regularization can inadvertently remove informative relationships.

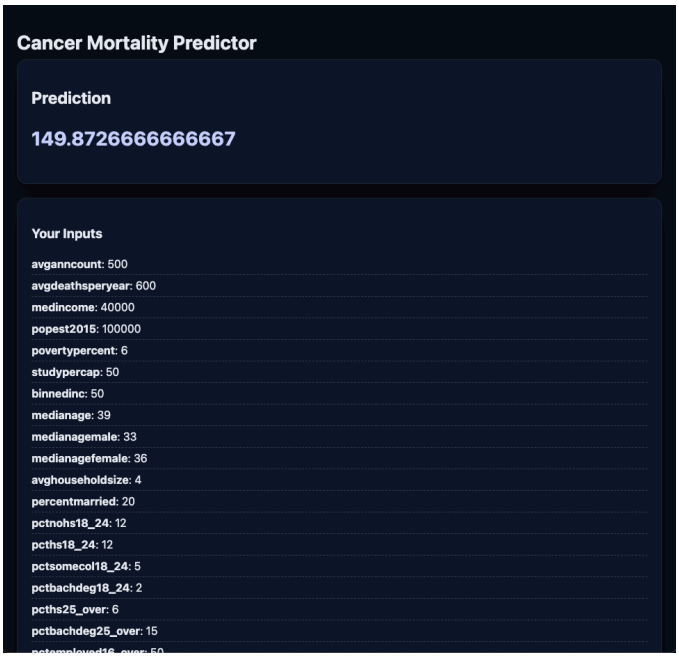
B. Mistakes Made, Challenges, and Future Considerations

Early iterations of the model revealed several challenges. Initial runs failed to meet OLS assumptions due to unscaled variables and untransformed skewed distributions, leading to inflated residuals and high variance. Overfitting was also a concern when polynomial and interaction terms were added indiscriminately. These mistakes underscored the need for systematic feature selection and validation rather than heuristic tuning.

One key challenge was balancing interpretability with complexity: while polynomial and interaction models captured some nonlinear trends, they made coefficient interpretation difficult. Next steps could be focused on integrating temporal and geographic data to explore spatial and time-dependent variations in mortality. Nonlinear methods, such as tree-based ensembles or neural regression networks, could be tested to capture interactions missed by linear modeling, provided interpretability remains a priority. Finally, automating diagnostic visualization and feature importance reporting will further streamline the analysis process for future large-scale health studies.

APPENDIX A

Screenshot showcasing a flask app that allows the user to generate a cancer mortality prediction using our model.



REFERENCES

1. Rippner, N. OLS Regression Challenge. data.world. https://data.world/nrippner/ols-regression-challenge
2. Centers for Disease Control and Prevention (CDC), 'United States Cancer Statistics: Data Visualizations,' 2024. [Online]. Available: https://www.cdc.gov/cancer/dataviz
National Cancer Institute (NCI), 'Cancer Health Disparities Fact Sheet,' 2024. [Online]. Available: https://www.cancer.gov/about-cancer/disparities/fact-sheet
3. G. K. Singh and A. Jemal, 'Socioeconomic and Racial Inequalities in Cancer Mortality: Patterns and Trends in the United States, 1950–2014,' Cancer Epidemiology, Biomarkers & Prevention, vol. 26, no. 4, pp. 541–553, 2017.
4. X. Han et al., 'Income and Education Disparities in Cancer Mortality in the United States,' Journal of Health Economics, vol. 44, pp. 1–12, 2015.
5. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning,* 2nd ed., Springer, 2021.
6. M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Regression Models,* 4th ed., McGraw-Hill, 2004.
7. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis,* 6th ed., Wiley, 2021.
8. G. A. F. Seber and A. J. Lee, *Linear Regression Analysis,* 2nd ed., Wiley, 2012.
9. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning,* 2nd ed., Springer, 2009.
10. F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python,' Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
11. R. R. Hocking, *Methods and Applications of Linear Models: Regression and the Analysis of Variance,* 3rd ed., Wiley, 2013.