A Multiple Linear Regression Approach to Predicting Synthetic Opioid Death Rates

Alan Baxley
The Pennsylvania State University
ajb8481@psu.edu

Ryan Brennan
The Pennsylvania State University
rmb6448@psu.edu

***Abstract*—Synthetic opioid overdoses, particularly those involving fentanyl, are a critical public health crisis. This study uses multivariate linear regression to model synthetic opioid deaths as a function of deaths associated with heroin, cocaine, psycho-stimulants, and natural/semi-synthetic opioids. Using CDC data, the model accounts for 59.27% of the variance in synthetic opioid deaths ($R^2 = 0.5927$), with all predictors contributing significantly. While residual analysis revealed heteroscedasticity (White's Test $p = 0.000$), suggesting non-linearity, the model provides valuable insights into synthetic opioid fatalities and offers a surrogate mechanism for estimating deaths in under-reported regions. Future work should explore non-linear models, such as XGBoost, and incorporate additional predictors to improve accuracy and inform targeted public health interventions.**

I. Introduction

Drug overdose deaths are an ever-growing public health crisis, with synthetic opioids emerging as one of the most dominant and deadly contributors. According to Brent et al., the opioid epidemic has claimed more than 800,000 lives in the US, more than the combined total of all deaths in all US wars and armed conflicts and, since approximately 2013, the death rate has increased more than 10-fold with the introduction of the inexpensive and highly potent synthetic opioid fentanyl [1]. Such a crisis requires analysis of relevant data so that trends can be identified and policy makers can act in the best interest of those being affected.

This study examines the connection between drug-related fatalities from specific substances and deaths caused by synthetic opioids, excluding methadone. The dataset used, VSRR Provisional Drug Overdose Death Counts [2], was obtained from the Centers for Disease Control and Prevention an includes data on overdose deaths across counties from various U.S. states. Our dataset has entries starting from April 12, 2015 to May 12, 2024. However, not every county is able to report synthetic opioid deaths reliably. Consequently, the goal of this analysis is twofold. First, to determine how deaths associated with heroin, cocaine, psycho-stimulants, and natural/semi-synthetic opioids serve as predictors for synthetic opioid fatalities and second, to provide a model to produce a surrogate aggregate of synthetic opioid deaths for the defined time period for counties that were unable to report reliably.

II. Methodology

A multiple linear regression model was implemented to examine the connection between drug-related fatalities from specific substances and deaths caused by synthetic opioids, excluding methadone. Multiple regression was chosen to account for the combined influence of multiple explanatory variables, providing a more nuanced understanding of the factors contributing to synthetic opioid deaths. Coefficients were obtained from the closed-form solution for ordinary least squares (OLS). The regression model is represented as follows:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \tag{1}$$

- $\hat{y}$: Predicted number of synthetic opioid deaths
- β_0: Intercept term
- X_1: Heroin-related deaths
- X_2: Cocaine-related deaths
- X_3: Psycho-stimulant-related deaths
- X_4: Natural opioid-related deaths

The dataset was filtered down to 2162 observations to focus on these specific substances because they are all tangentially related to synthetic opioids, such as fentanyl, in the sense that they are all pharmacologically involved in the manipulation of dopamine pathways. The metrics related to heroin, cocaine, psycho-stimulants, and natural/semi-synthetic opioids were included as explanatory variables. Methadone was excluded from the response variable to focus on synthetic opioids more commonly linked to recent overdose trends, particularly fentanyl and its derivatives. This decision also avoids confusing synthetic opioid death trends with methadone's role in opioid treatment.

The completeness of the dataset and its reliability were ensured, with the removal of missing values. In our analysis, we assumed that there is a linear relationship between predictors and response, residuals are normally distributed, and there is no perfect multicollinearity among the explanatory variables. A number of metrics and heuristics were applied to confirm these assumptions and to evaluate how well the model performed. We utilized F-statistics, T-statistics, standard errors, and the corresponding P-values to perform hypothesis testing for global usefulness, coefficient of determination (R^2), and the regression parameters.

Residual analysis was performed to check the model's accuracy and ensure the assumptions held true. Multiple linear regression models rely on a number of assumptions being satisfied in order to provide a reliable approximation to the true association between a response and explanatory variables and these assumptions describe the probability distributions

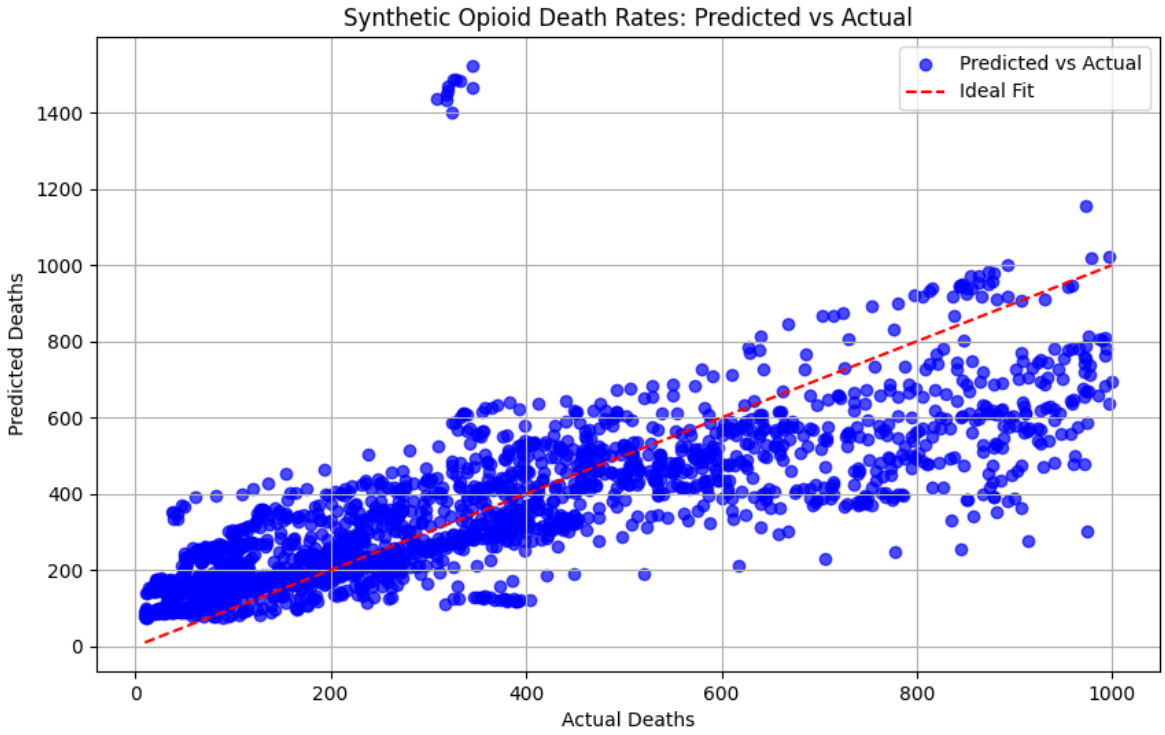

Fig. 1. Synthetic Opioid Death Rates

of the random errors in the model [3]. The random error is defined as $\hat{e} = y - \hat{y}$. The following four assumptions regarding $\mathbf{e}$ [3] must hold:

1) The probability distribution of $\mathbf{e}$ at each set of values $(X_1, X_2, ...X_k)$ has a mean of zero.
2) The probability distribution of $\mathbf{e}$ at each set of values $(X_1, X_2, ...X_k)$ has constant variance or homoscedasticity.
3) The probability distribution of $\mathbf{e}$ at each set of values $(X_1, X_2, ...X_k)$ is normal.
4) The value of $\mathbf{e}$ for one observation is independent of the value of $\mathbf{e}$ for any other observation.

The condition number was also calculated to detect multicollinearity, as higher values could indicate instability in the model caused by correlations among predictors. We implemented a statistical test, White's test, to check for heteroscedasticity in the probability distribution $\mathbf{e}$ to cross reference visual analysis of the residual scatter plot. If the p-value from White's test is significant, it suggests that the homoscedasticity assumption has been violated. Two visualizations were created to support these diagnostics: a Predicted vs. Actual plot, which shows how well the model's predictions matched the observed deaths, and a Residuals vs. Predicted plot, which provides the means for visual analysis of the four random error assumptions.

III. Results

The results from for our multivariate model using synthetic opioid deaths as the response variable and heroin-related deaths, cocaine-related deaths, psycho-stimulant-related deaths, and natural opioid-related deaths as the explanatory variables are presented in this section. The R^2 value of 0.5927 indicates that approximately 59.27% of vari-

TABLE I
Regression Results Summary

Statistic	**Value**
R^2	0.5927
Adjusted R^2	0.5920
F-statistic	784.8
Condition Number	5.933
Coefficients (β)	**Value (SE, t-statistic, p-value)**
Intercept (β_0)	53.65 (7.35, 7.30, 4.16×10^{-13})
Heroin (β_1)	-0.198 (0.036, -5.52, 3.84×10^{-8})
Cocaine (β_2)	1.585 (0.041, 38.78, 0.0)
Psycho-stimulants (β_3)	0.521 (0.019, 27.08, 0.0)
Natural Opioids (β_4)	-0.139 (0.034, -4.12, 3.94×10^{-5})
Residual Analysis	
Mean Residual	-6.02×10^{-12}
Residual Standard Deviation	168.36
White's Test Statistic	1662
White's Test p-Value	0.000

ation in synthetic opioid deaths can be accounted for by a linear association between synthetic opioid deaths and the explanatory variables heroin-related deaths, cocaine-related deaths, psycho-stimulant-related deaths, and natural opioid-related deaths. The condition number, 5.933, is low indicating little to no correlation among the explanatory variables. Figure 1 displays the predicted deaths vs the actual deaths with the x-axis being the $\mathbf{y}$-values from the dataset and the y-axis being $\hat{y}$ from equation (1).

IV. Model Evaluation

To verify our initial assumption of linearity between the predictors and response variables and demonstrate that our model does have explanatory power over simply using the mean of synthetic opioid deaths, we performed several hypothesis tests as well as a visual analysis of the residual distribution.

A. Hypothesis Tests

1) Global Usefulness Test: To determine that at least one of the explanatory variables has a significant relationship with the response variable, we performed a global usefulness test as follows:

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
- $H_1 : \exists i$ such that $\beta_i \neq 0$

The F-distribution was used with $\alpha = 0.05$ (95% confidence) and degrees of freedom ($\mathbf{v}$) were calculated as

$$v = k/(n - k - 1) \tag{2}$$

where $\mathbf{k}$ is the number of regression parameters and $\mathbf{n}$ is the number of observations. Hence, $v = 4/2157$ and our global F-statistic from Table 1 is 784.8. We use the fact that $F_{0.05,(4,2157)} < F_{0.05,(4,120)} = 2.447$. Thus, $F = 784.8 > F_{0.05,(4,120)} > F_{0.05,(4,2157)}$ and we can reject the null hypothesis with at least 95% confidence.

2) Coefficient of Determination: The coefficient of determination, R^2, can be mathematically defined as:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3}$$

where:

- RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, the Residual Sum of Squares, represents the sum of the squared differences between observed values y_i and their corresponding predicted values $\hat{y}_i$.
- TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$, the Total Sum of Squares, represents the total variance in the observed values, where $\bar{y}$ is the mean of the observed y values.

[3]
A hypothesis test would be set up as follows:

- $H_0 : R^2 = 0$
- $H_1 : R^2 > 0$

The coefficient of determination, R^2, quantifies the proportion of variance in the response variable y that is explained by the predictors in the regression model. The hypothesis test for R^2 being significantly greater than zero ($H_0 : R^2 = 0$) is inherently equivalent to the Global Usefulness Test, as $R^2 = 0$ would imply that none of the predictors contribute to explaining the response variable. Since the F-statistic (784.8) in the Global Usefulness Test significantly exceeds the critical value ($F_{0.05,(4,2157)} < F_{0.05,(4,120)} = 2.447$), we conclude with at least 95% confidence that $R^2 > 0$. This result confirms that the model is globally significant and that the predictors collectively explain a meaningful proportion of the variance in synthetic opioid deaths.

3) Regression Parameters: We performed hypothesis tests on the individual coefficients, β_i, to determine their significance as follows:

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$

We performed two tailed tests with $\alpha = 0.05$ and 2157 degrees of freedom. We use the fact that $t_{0.025,2157} < t_{0.025,120} = 1.98$ and, consequently, define our critical regions as $t > 1.98$ or $t < -1.98$. Table 1 reports our t-values for each parameter as:

- β_1 -5.52
- β_2 38.78
- β_3 27.08
- β_4 -4.12

Thus, $\forall i\ t_i > 1.98$ or $t_i < -1.98$ and we can reject $H_0\ \forall i\ \beta_i$ with at least 95% confidence.

B. Residual Analysis

Figure 2 shows the residual distribution relative to the predicted $\hat{y}$ values of our model. Upon performing a visual analysis, we ascertain that the residuals appear to be evenly distributed around the hyperplane. This is verified by mean residual value of $-6.02 * 10^{-12}$ in Table 1, which is very close to 0. However, the variance of the residual distribution does not appear constant as the variance amplitude from the hyperplane increases significantly from predicted values 200 to 400. This is verified by the White's Test P-value result of 0.000 in Table 1. This P-value result means we can reject the null hypothesis that the residual distribution is homoscedastic at any reasonable level of significance (even $\alpha = 0.01$). The clustering of the residuals do appear to be relatively normal and there are discernible patterns to the residual distribution. Thus, our model satisfies three out of four of the assumptions needed to provide a reliable linear association between the response and predictors.

V. Discussion

The results of our analysis demonstrate that the multivariate regression model provides more explanatory power than simply using the mean ($\bar{y}$) of synthetic opioid deaths. Specifically, the Global Usefulness Test and R^2 Test confirm that the predictors collectively account for approximately 59.27% of the variance in synthetic opioid deaths ($R^2 = 0.5927$), providing strong evidence of a meaningful linear association between the response variable and the explanatory variables (heroin-related deaths, cocaine-related deaths, psycho-stimulant-related deaths, and natural opioid-related deaths).

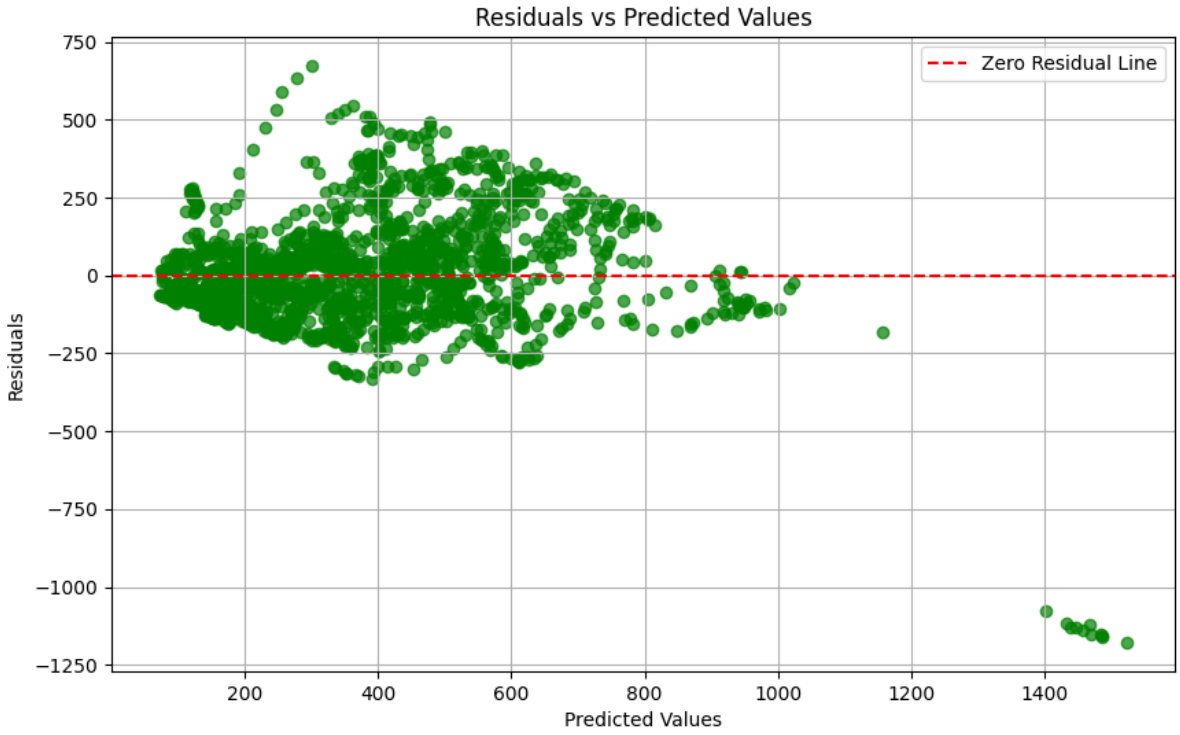

Fig. 2. Synthetic Opioid Death Rates Residual Plot

Furthermore, the hypothesis tests on individual regression coefficients (β_i) indicate that each predictor contributes significantly to the explanatory power of the model. At least 95% confidence can be assigned to the conclusion that none of the coefficients (β_i) are zero, further supporting the notion that the predictors have individual explanatory significance.

Taken together, these results suggest that the sample data is unlikely to come from a population where $R^2 = 0$ or all $\beta_i = 0$. Thus, we conclude that there is strong evidence of a degree of linear association between synthetic opioid deaths and the explanatory variables. This supports the use of the model as a tool for understanding and predicting trends in synthetic opioid deaths, albeit with the limitation of heteroscedasticity, as evidenced by White's Test. Corrective measures, such as robust standard errors or non-linear transformations, may enhance the reliability of statistical inferences.

VI. Conclusion

In this analysis, we explored the linear relationship between the number of deaths from synthetic opioids and heroin, cocaine, psycho-stimulants, and natural/semi-synthetic opioids using a multivariate linear regression model. The analysis demonstrated that the model provides significant explanatory power, accounting for 59.27% of the variation in synthetic opioid deaths ($R^2 = 0.5927$). This highlights a meaningful linear association between synthetic opioid deaths and the predictors, validated through hypothesis testing for both global usefulness and individual regression parameters.

While the predictors collectively and individually contribute significantly to the model, the residual analysis revealed evidence of heteroscedasticity, as indicated by White's Test ($p = 0.000$). This limitation suggests that the relationship between synthetic opioids deaths and the predictors is not strictly linear. Future analyses could explore non-linear models such as XGBoost, which excel in handling complex interactions and non-linear relationships between predictors and response variables.

Despite the lack of homoscedastic residual distribution, our model provides a valuable framework for understanding the factors contributing to synthetic opioid deaths. Policymakers and public health officials can utilize this semi-linear relationship and identify regions at risk and develop targeted interventions. Furthermore, the model offers a surrogate mechanism to estimate synthetic opioid deaths in counties where reliable data is unavailable, thereby enhancing the comprehensiveness of overdose surveillance.

Future research could expand this work by incorporating additional explanatory variables, such as socio-economic factors, to further improve the model's explanatory power and predictive accuracy. By acknowledging the identified limitations and exploring new areas of analysis, this work can continue to support efforts to mitigate the opioid crisis

and save lives.

Supplementary Material and Code: The supplementary material and code are available at: https://github.com/ZeroTolerance225/Multiple_Regression_Drug_Overdose

REFERENCES

[1] J. Brent and S. T. Weiss, "The opioid crisis—not just opioids anymore," *JAMA Network Open*, vol. 5, no. 6, pp. e2 215 432–e2 215 432, 06 2022. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2022.15432

[2] C. for Disease Control and Prevention, "Vsrr provisional drug overdose death counts," 2024. [Online]. Available: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm

[3] I. Pardoe, *Applied Regression Modeling*, 3rd ed. Hoboken, NJ, USA: Wiley, 2021.