A Statistical Exploration of Factors Driving Mortality Rates in Small American Cities

Jason Noble

School of Engineering & Technology, National University

ANA605, Analytic Models and Data Systems

Matthew Vanderbilt

August 18, 2024

Abstract

This study explores factors influencing mortality rates in small American cities, and focused on the role of income and healthcare availability. The analysis investigated how income impacts the availability of doctors and hospitals, and how these factors, in turn, affect mortality rates. The primary variable of interest was the mortality rate, with income hypothesized to influence it indirectly through healthcare availability. Using multiple linear regression models, we first analyzed the direct effect of income on healthcare availability. We then examined the impact of these healthcare variables on mortality rates and assessed a combined model to evaluate both direct and indirect effects of income on mortality. This approach allowed us to explore the complex interactions between socioeconomic factors and health outcomes in small American cities.

Method

The dataset includes 5 variables derived from sources documenting socioeconomic and healthcare factors across 53 small American cities from: Life in America's Small Cities, by G.S. Thomas. **X1: Death Rate** –represents the mortality rate, measured as the number of deaths per 1,000 residents in each city. It serves as the primary variable of interest in the analysis, reflecting overall public health outcomes in these areas. **X2: Doctor Availability** – measures the availability of doctors, expressed as the number of doctors per 100,000 residents. It is used to assess the level of medical care accessible to the population, which is hypothesized to influence mortality rates. **X3: Hospital Availability** – measures the availability of hospital facilities, expressed as the number of hospitals per 100,000 residents. Like doctor availability, this variable reflects the healthcare infrastructure in each city. **X4: Annual Per Capita Income** – represents the average income per person, measured in thousands of dollars earned annually. A key socioeconomic indicator, hypothesized to influence healthcare availability and thus mortality rates. **X5: Population Density** – measures the number of people per square mile in each city. Population density was hypothesized to be inversely related to healthcare accessibility and public health outcomes: As density increased (inner city), infrastructure would stay constant, but would translate to less healthcare options available to the people. These variables were collected to explore the interplay between socioeconomic status, healthcare resources, and mortality rates in small American cities, providing an initial look at factors that may contribute to variability in health outcomes.

Exploratory Data Analysis

I looked at outliers with boxplot() to see if there was any data that needed to be cleaned, but did not see any that should be excluded.

Processed the descriptive statistics for the variables, Table 1. Here's a summary of the findings as visualized in Figure 1. **X1**: Has a normal distribution skewed to the left. **X2**: Shows a bimodal distribution heavily skewed to the right. **X3**: Heavily grouped around 500 per 100,000 and long right tail indicates a high disparity in hospital availability. **X4**: Normal distribution, indicating economic diversity, with a slight skew to the right. **X5**: Considerable population variability with another right skew to the data.

The correlation matrix, Figure 2, reveals that there is some correlation between X2 (doctor availability) and X3 (hospital availability), that will need to be explored during collinearity analysis.

Figure 1

Histograms of Variables, showing the density of each variable

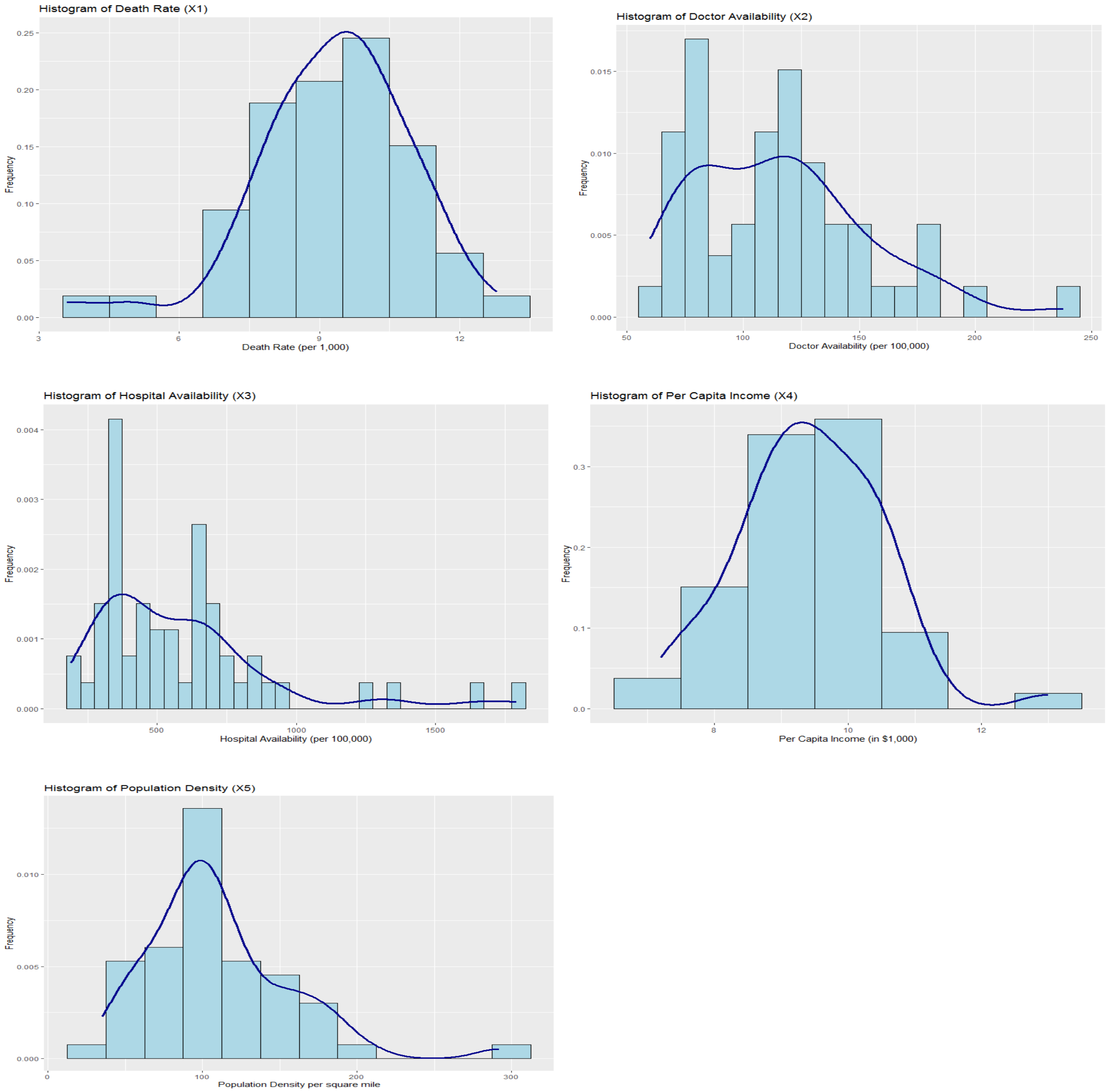


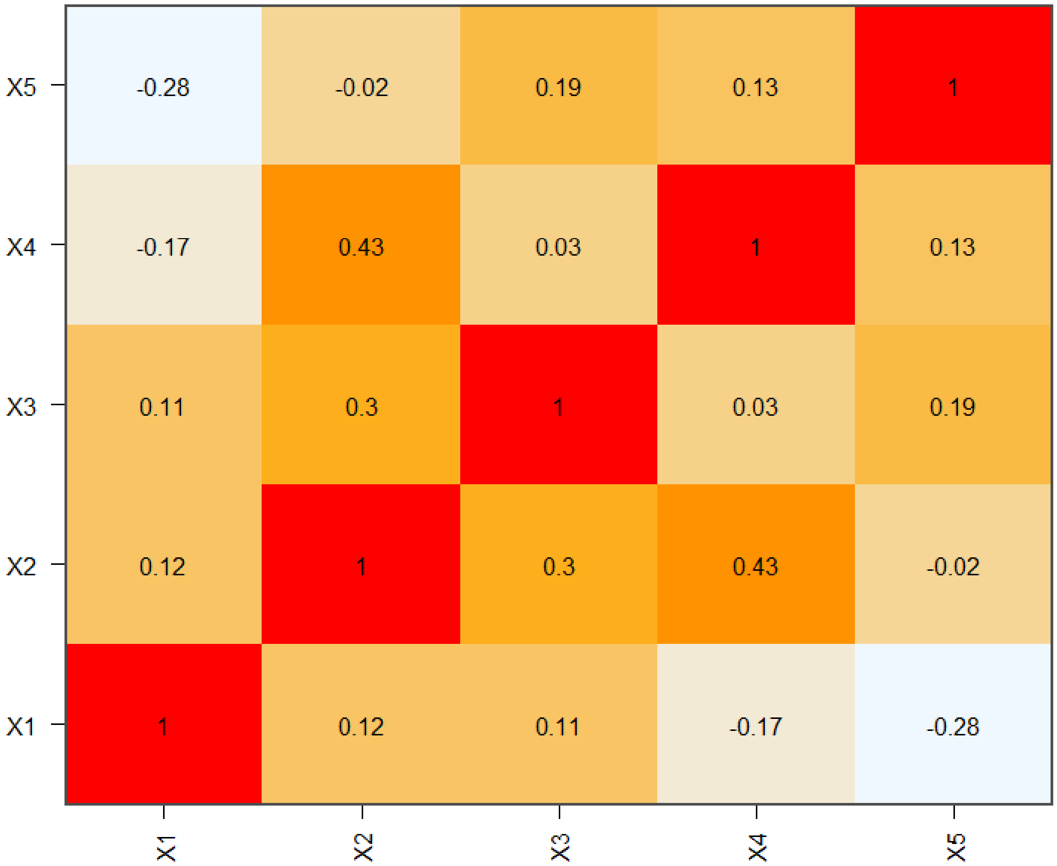
Table 1

Descriptive Statistics for data Variables

Variable	X1	X2	X3	X4	X5
Min	3.6	60	190	7.2	35
Q1	8.3	82	353	8.8	80
Median	9.4	114	525	9.5	103
Q3	10.3	134	686	10.3	129
Max	12.8	238	1792	13	292
Mean	9.31	116.09	589.79	9.44	110.64
sd	1.66	37.89	332.62	1.08	47.18
n	53	53	53	53	53

Figure 2

Correlation Matrix Heatmap



Linear Regression Analysis

Performed a multiple linear regression with **X1** (death rate) as the dependent variable and X2, X3, X4, X5 as independent variables. Table 2 shows the results: With an intercept estimate of 12.27, this represents the predicted value of X1 when other variables are zero. A very significant p-value indicates a strong level for the death rate independent of the other factors.

X2: This suggests a slight increase in death rate with higher doctor availability, but it is not statistically significant. **X3: T**his small positive coefficient indicates a minimal and non-significant effect of hospital availability on the death rate. **X4:** This negative coefficient suggests that higher income is associated with lower death rates, but this relationship is not statistically significant. **X5:** The negative coefficient indicates that higher population density might be associated with lower death rates, and is borderline significant (closest to 0.05 threshold). The Multiple **R^2** means that the model explains about 14.37% of the variance in death rates, which is relatively low. The **Adjusted R^2** adjusts for the number of predictors, and means that the model explains about 7.24% of the variance. The low **F-statistic** and **p-value** mean the overall model is not statistically significant, indicating that the combination of these predictors does not significantly explain the variation in death rates.

Table 2

Analysis of Variance Table (Type III SS)

Variable	Error	SS	df	MS	F	PRE	p
Model	(e reduced)	20.65	4	5.16	2.01	.144	.108
X2		2.91	1	2.91	1.14	.02	.29
X3		1.68	1	1.68	.65	.01	.42
X4		5.08	1	5.08	1.98	.04	.17
X5		9.61	1	9.61	3.75	.07	.06
Error	(from model)	123.07	48	2.56	N/A	N/A	N/A
Total	(empty model)	143.728	52	2.76	N/A	N/A	N/A

Model: X1 ~ X2 + X3 + X4 + X5

Collinearity Investigation

Income and Doctor Availability: Income significantly increases doctor availability. This is a strong and statistically significant positive effect on doctor availability, with a coefficient of 15.57503 and a p-value of 0.000675. This suggests that higher income levels are associated with increased doctor availability.

Residual Analysis

All **VIF** values are around 1.20, below 5, indicating that multicollinearity is not a significant issue in this model. Therefore, no variables need to be removed based on multicollinearity. Figure 3 shows a normally distributed, but with the consistent skew to the right, which could indicate the normality assumption is slightly off. The **Q-Q plot** suggests that the residuals deviate from the line of normality, particularly at the tails, which further confirms the skewness observed in the histogram.

Figure 3

Histogram of Residuals

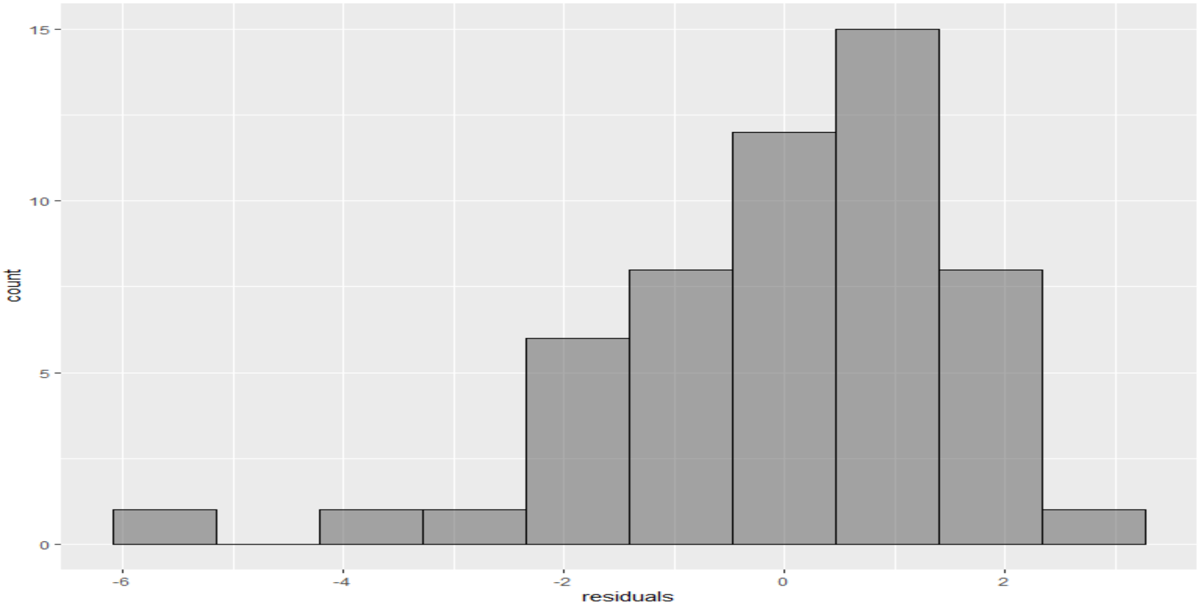
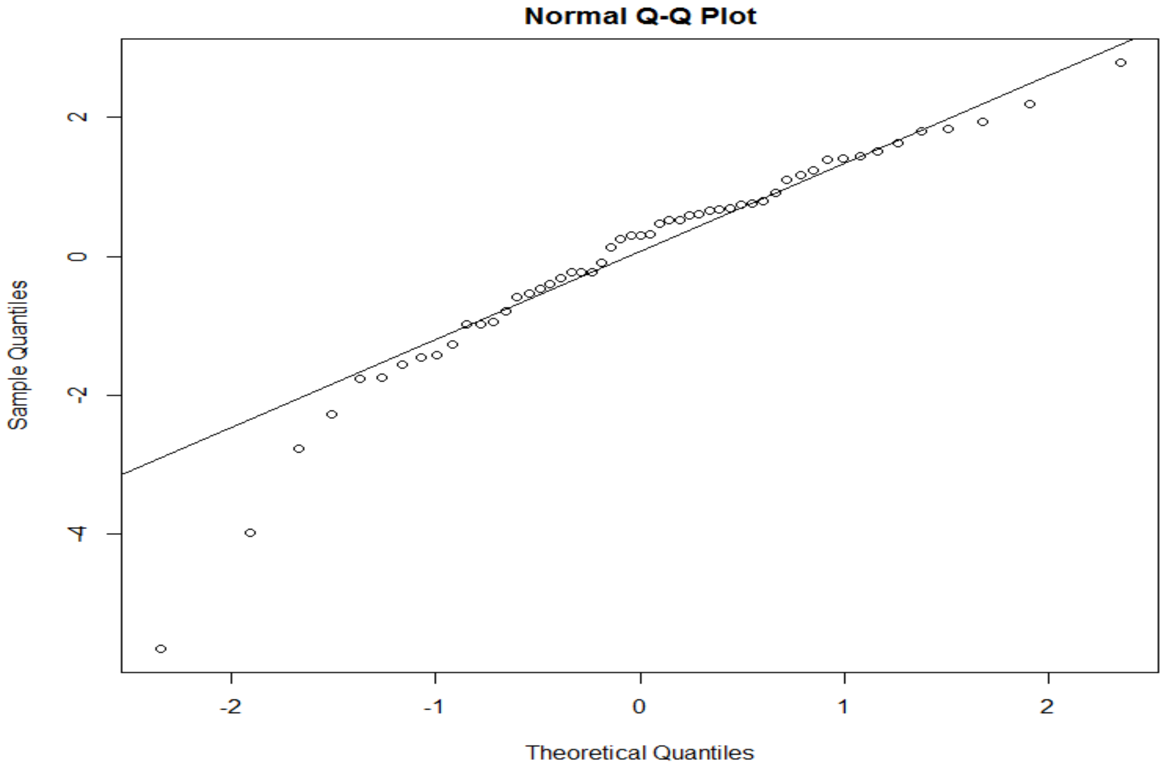


Figure 4

QQ Plot



Summary

The multiple linear regression analysis reveals that the model explains only a small portion of the variance in the death rate (14.4%), and none of the independent variables are statistically significant. The VIF analysis shows no serious multicollinearity issues. However, the residual analysis suggests some deviations from the assumptions of normality, which may indicate that the model could be improved or that the relationship between the variables is not entirely linear. The results suggest that other variables might better explain the variability in the death rate, or that the relationship between these predictors and the death rate is more complex than what a simple linear model can capture.