Statistical Analysis of Serious Crimes: A Comparison of Frequentist and Bayesian Approaches

> ***Abstract:*** *This report presents a statistical analysis of factors influencing serious crime rates across different counties. Through both frequentist and Bayesian modeling approaches, we identify the number of hospital beds as the most significant predictor of serious crime rates among the variables examined. The analysis includes exploratory data analysis, model comparison, residual diagnostics, and discussion of limitations and potential improvements. Our findings suggest that hospital beds likely serve as a proxy for urbanization, with more urbanized areas experiencing higher crime rates. Model diagnostics reveal several outliers that impact model performance, emphasizing the need for more complex modeling approaches in future work.*

1. Executive Summary

This report presents an analysis of factors influencing serious crime rates across different counties. Using demographic data, we built a statistical model to predict the number of serious crimes. Our analysis involved both traditional regression and Bayesian statistical methods, with a focus on finding the most informative single predictor. We identified that the number of hospital beds is the most significant predictor of serious crime rates. The report compares these modeling approaches, discusses limitations including outliers, and suggests potential improvements.

2. Introduction

Crime prediction models are valuable tools for policymakers and law enforcement agencies. In this analysis, we examined a dataset containing demographic information from various counties to determine which factors might be associated with serious crime rates. The primary objectives were to:

1. Identify the most significant predictor of serious crime rates
2. Compare traditional regression and Bayesian statistical approaches
3. Evaluate model diagnostics and limitations

3. Data and Methodology

3.1 Dataset

The analysis used a demographic dataset containing the following variables for each county:

- County and State identifiers
- Land area
- Population statistics (total, age groups)
- Healthcare resources (physicians, hospital beds)
- Education levels (high school and bachelor's degree graduation rates)
- Economic indicators (poverty rates, unemployment, income measures)
- Geographic region

3.2 Data Preparation

The response variable (serious crimes, Y_i) was log-transformed to address skewness and improve model fit. This transformation is common when modeling count data such as crime statistics.

```
Demographic$Log_Y_i <- log(Demographic$Y_i)
```

The log transformation serves several important purposes in our analysis. First, crime count data often follows a right-skewed distribution, with many counties having moderate crime rates and fewer counties having very high rates. The log transformation helps normalize this distribution, making it more suitable for linear regression techniques that assume normality. Second, it allows us to interpret the model coefficients in terms of percentage changes rather than absolute changes, which is often more intuitive for policy discussions. Finally, it helps stabilize the variance across the range of predictors, addressing potential heteroscedasticity issues.

4. Exploratory Data Analysis

4.1 Correlation Analysis

We began by examining correlations between the log-transformed serious crime rates and potential predictor variables. This helped identify which variables might have strong relationships with crime rates.

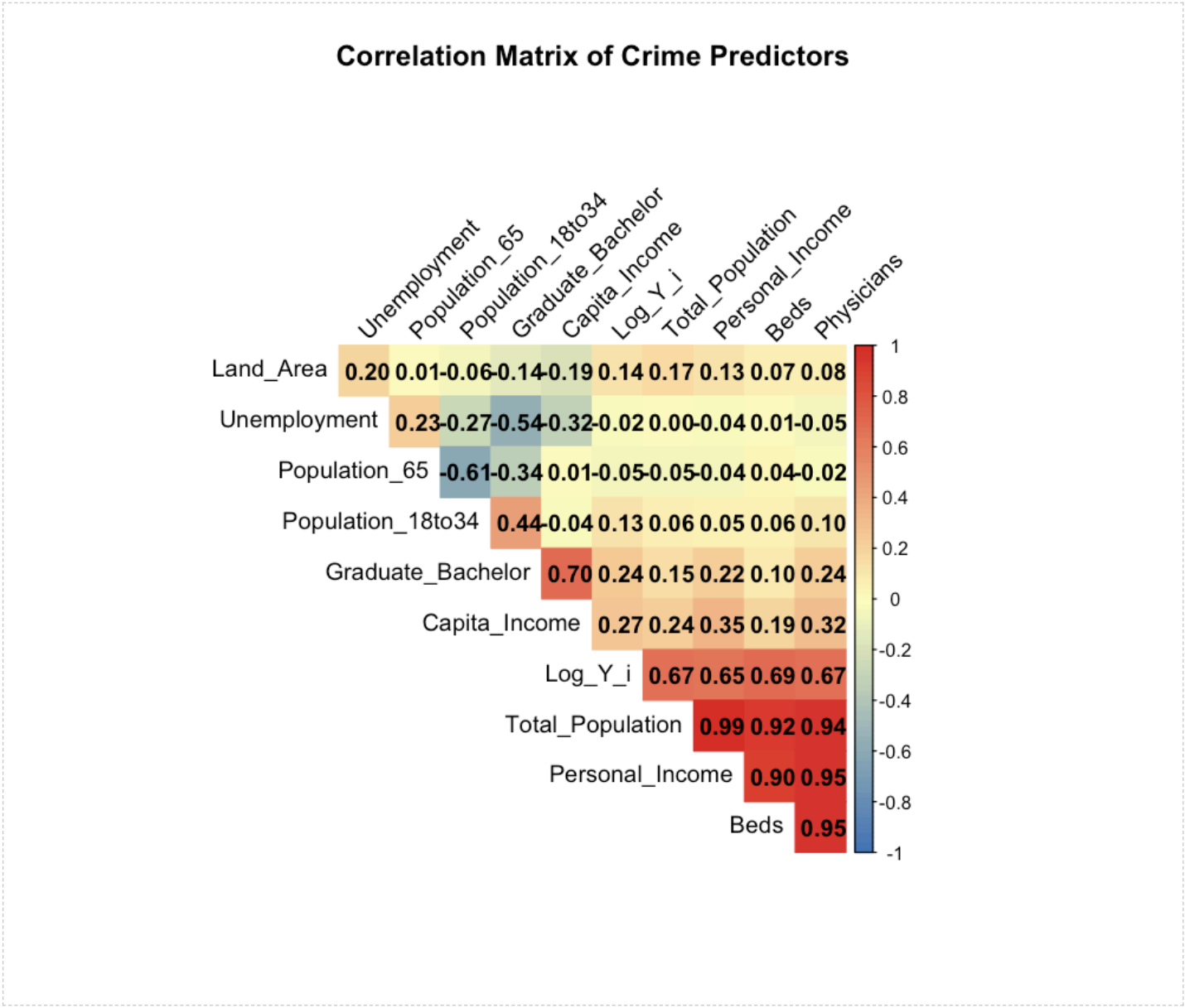

Figure 1: Correlation matrix heatmap showing relationships between log(serious crimes) and key predictors.

The correlation matrix reveals several important patterns in our data. Hospital Beds demonstrates the strongest correlation with log(serious crimes) at approximately 0.70, suggesting it may be our most informative predictor. Total Population also shows a strong positive correlation (around 0.67), which aligns with criminological theory that more populous areas tend to experience higher crime rates. Physicians shows a similar strong correlation (0.65), while Personal Income exhibits a moderate positive correlation (0.52).

Interestingly, some variables we might theoretically expect to be strongly correlated with crime rates show weaker relationships. For example, unemployment and poverty rates exhibit relatively modest correlations, which challenges some conventional assumptions about socioeconomic factors and crime. The education variables (Graduate_Highschool and Graduate_Bachelor) also show weaker correlations than expected.

These correlation patterns suggest that urbanization factors (represented by beds, physicians, and population) may be more strongly associated with crime rates in our dataset than socioeconomic factors. The multicollinearity among predictors (e.g., Beds, Physicians, and Total Population are all strongly intercorrelated) further supports this interpretation and influenced our decision to focus on a single-predictor model.

Key correlations with Log_Y_i included:

- Total Population: Strong positive correlation, suggesting more populated areas have higher crime rates
- Beds: Strong positive correlation, indicating areas with more hospital beds tend to have higher crime rates
- Personal Income: Moderate positive correlation
- Land Area: Weak positive correlation

4.2 Scatter Plot Analysis

Scatter plots were used to visualize relationships between log(serious crimes) and various predictors, allowing us to identify potential patterns and nonlinearities.

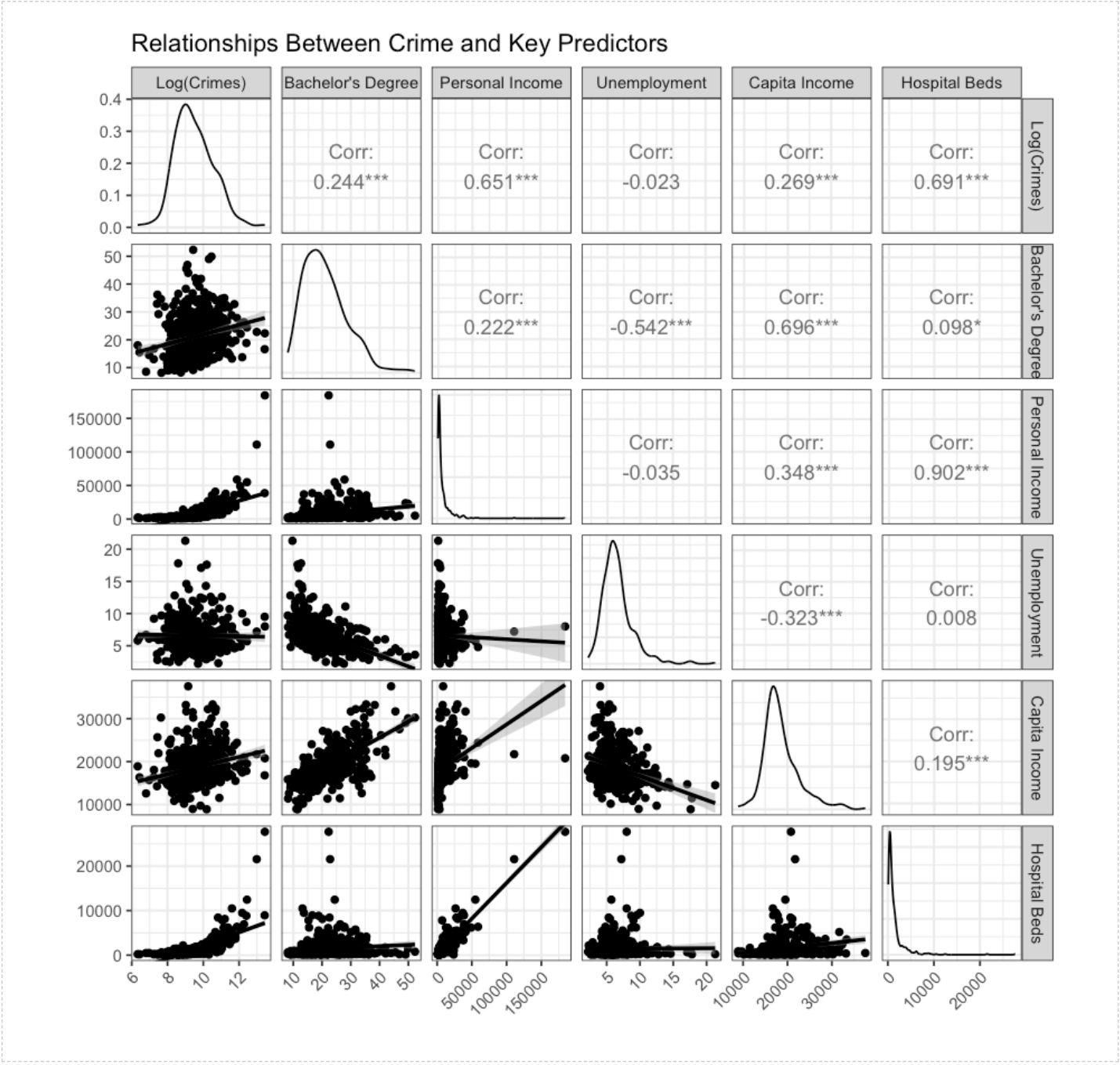

Figure 2: Scatterplot matrix showing relationships between log(serious crimes) and major predictors.

The scatterplot matrix provides visual confirmation of the correlations identified earlier while revealing additional nuances in these relationships. The scatter plots involving Log_Y_i (log of serious crimes) show varying patterns across different predictors:

The relationship between crime rates and Beds appears relatively linear but with increasing variability at higher values, suggesting heteroscedasticity. This pattern is also evident in the relationship with Personal Income and Capita Income. The relationship with Graduate_Bachelor shows a possible non-linear pattern, with diminishing effects at higher education levels.

The diagonal plots showing the distribution of each variable reveal important characteristics of our data. Most notably, several predictors (particularly Beds, Personal Income, and Graduate_Bachelor) show right-skewed distributions with extreme values that could potentially influence our model. The distribution of Log_Y_i appears more symmetric, confirming that our log transformation has helped normalize the crime data.

We also observe clustering patterns in several scatter plots, particularly for Beds and Personal Income, suggesting that counties may group into distinct categories based on their urbanization levels. These visual patterns reinforced our decision to explore Beds as our primary predictor while remaining aware of potential outliers and non-linearities.

We also examined the relationship between Personal Income and log(serious crimes) more closely:

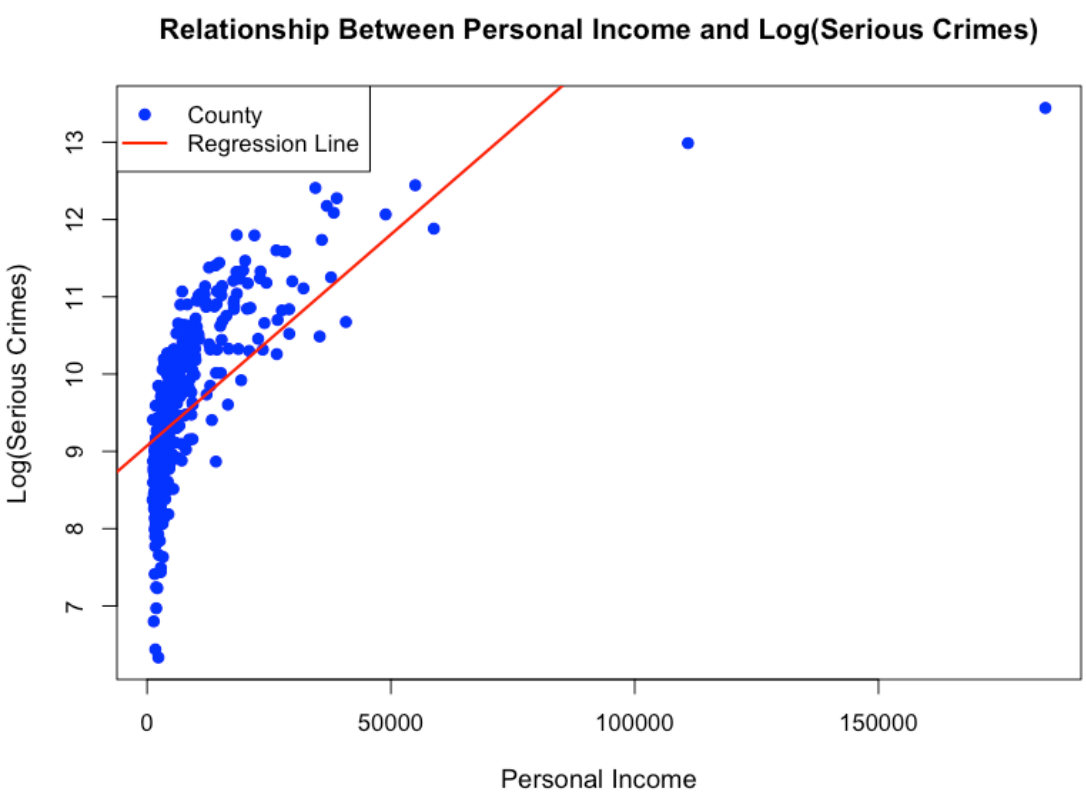

Figure 3: Relationship between Personal Income and log(serious crimes) with regression line.

This focused examination of Personal Income versus log(serious crimes) reveals a clear positive relationship that might seem counterintuitive at first glance. Conventional wisdom might suggest that higher income areas would experience lower crime rates, but our data shows the opposite trend. The red regression line indicates that as Personal Income increases, log(serious crimes) tends to increase as well.

Several features of this relationship deserve attention. First, there is substantial variability around the regression line, suggesting that Personal Income alone explains only a portion of the variation in crime rates. Second, we can observe a cluster of counties with lower Personal Income and lower crime rates, likely representing rural areas. Third, several notable outliers appear in both the upper and lower regions of the plot.

This relationship likely reflects the confounding effect of urbanization – counties with higher Personal Income tend to be more urbanized, and urbanization is associated with higher crime rates. When interpreting this relationship, we must be careful not to infer causality; Personal Income is likely serving as a proxy for urbanization rather than directly influencing crime rates.

4.3 Variable Distribution Analysis

Understanding the distribution of key variables helps identify potential modeling challenges:

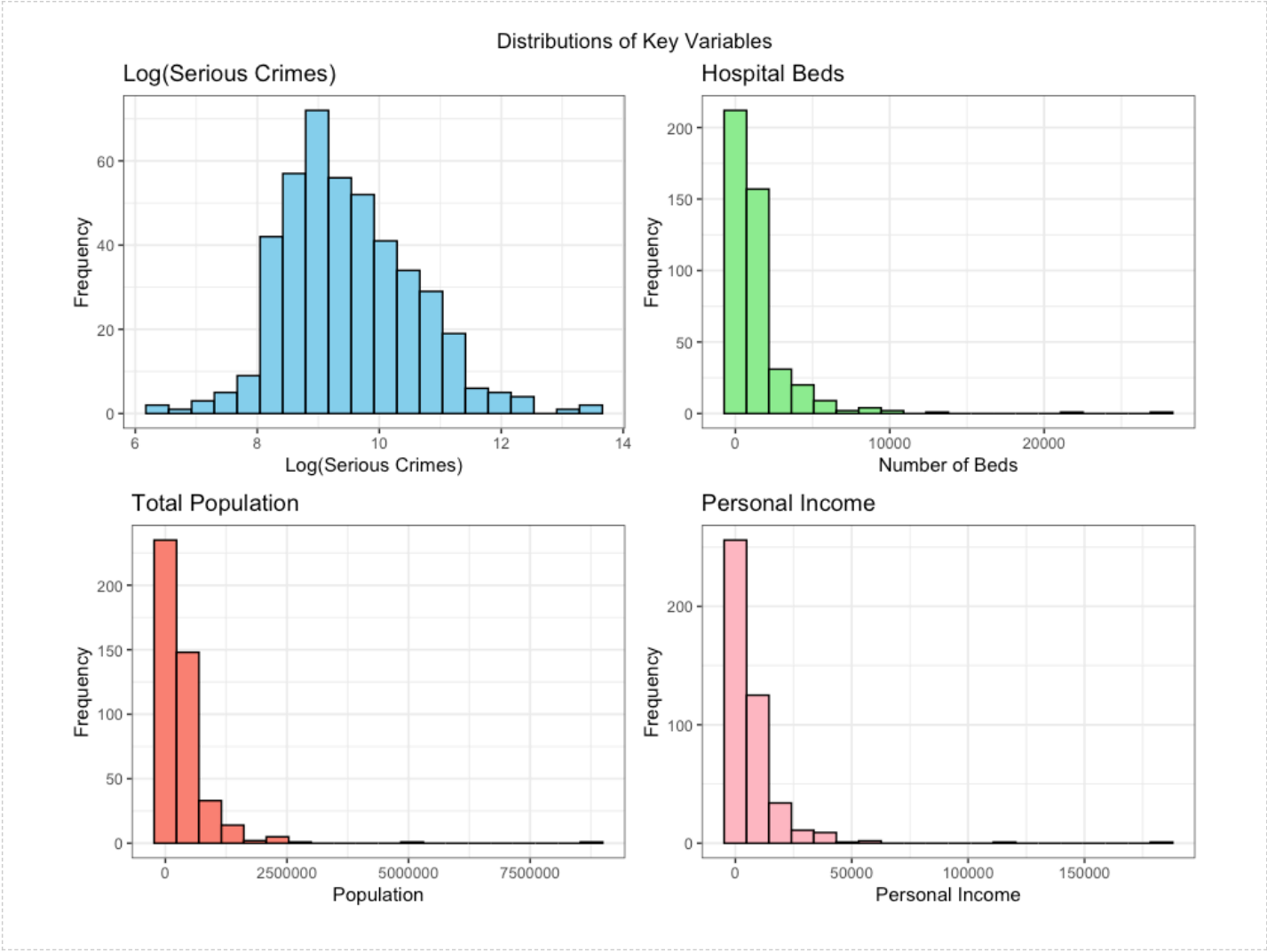

Figure 4: Histograms showing distributions of key variables in the dataset.

The histograms in Figure 4 provide important insights into the distributions of our key variables:

The Log_Y_i (log of serious crimes) distribution appears approximately normal with a slight right skew, validating our decision to use a log transformation. The approximately normal distribution makes this variable suitable for linear regression techniques.

In contrast, the Hospital Beds distribution is strongly right-skewed with a long tail, indicating that most counties have relatively few hospital beds while a small number of counties (likely major urban centers) have substantially more. This skewness could potentially affect our model's performance, particularly for counties at the extreme ends of the distribution.

The Total Population histogram shows a similar right-skewed pattern, reinforcing the connection between population size and hospital infrastructure. Personal Income also exhibits right-skewness, though less extreme than Beds and Population.

These distributional characteristics highlight potential challenges for our modeling approach. The skewed predictor distributions might lead to influential outliers and heteroscedasticity in our model residuals. While our log transformation of the response variable helps address some of these issues, we should remain cautious when interpreting predictions, particularly for counties with extreme values on these predictors.

5. Variable Selection Process

Several single-predictor models were evaluated to identify the most significant predictor:

- Hospital Beds
- Per Capita Income
- Total Population

- Bachelor's Degree Graduation Rate
- Number of Physicians

Our selection process involved fitting separate linear regression models for each candidate predictor and comparing their performance metrics. While many predictors showed statistically significant relationships with crime rates, Hospital Beds emerged as the strongest single predictor based on multiple criteria including R-squared values, F-statistics, and residual standard errors.

The model with Hospital Beds explained approximately 48% of the variance in log(serious crimes), outperforming the other candidate predictors. Total Population and Physicians also showed strong predictive power, but were slightly less effective than Beds. We elected to use a single-predictor model for simplicity and interpretability, despite the potential for improved fit with multiple predictors. This approach allows for clearer comparison between frequentist and Bayesian methods while avoiding issues of multicollinearity among the urbanization-related predictors.

After comparing these models, hospital beds emerged as the most significant predictor. The final model selected was:

Log(Serious Crimes) ~ Beds

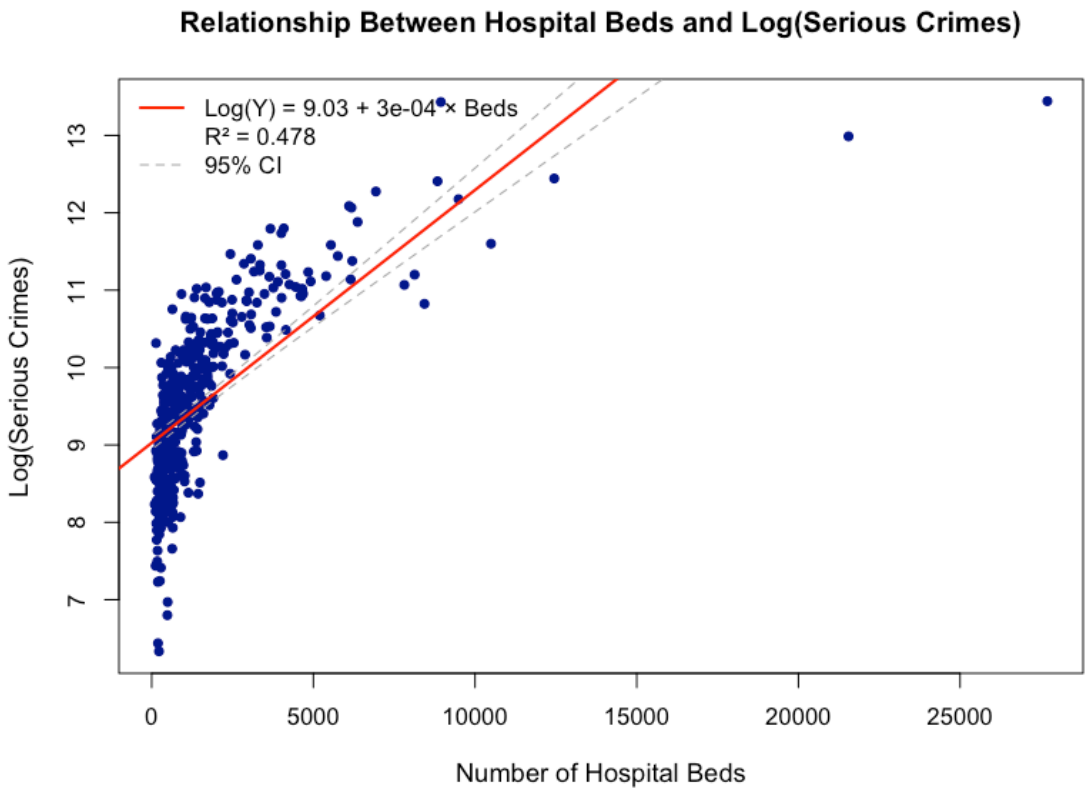

Figure 5: Relationship between Hospital Beds and log(serious crimes) with regression line.

Figure 5 visualizes the relationship between Hospital Beds and log(serious crimes) that forms the basis of our model. The scatter plot reveals a positive relationship with a clear upward trend, confirmed by the fitted regression line in red. As the number of hospital beds increases, log(serious crimes) tends to increase as well.

The plot provides several important observations about this relationship. First, the data points show a relatively linear pattern overall, though with increasing variability at higher values of Beds. Second, we observe clustering of data points at the lower end of both variables, representing counties with fewer beds and lower crime rates. Third, we can identify several potential outliers, particularly at the extreme ends of the distribution.

The equation displayed in the plot, $\text{Log}(Y) = 9.03 + 0.0003 \times \text{Beds}$, tells us that each additional hospital bed is associated with approximately a 0.03% increase in serious crimes (since we're using a log transformation). The R^2 value of 0.483 indicates that hospital beds alone explain about 48.3% of the variation in log(serious crimes) across counties, which is substantial for a single predictor.

The 95% confidence interval bands (shown as dashed gray lines) widen at the extremes of the predictor range, indicating greater uncertainty in predictions for counties with very few or very many hospital beds. This pattern reflects the uneven distribution of our data and suggests caution when making predictions for counties at the extremes.

6. Model Results

6.1 Frequentist Linear Regression

The summary statistics of the final model showed:

- A significant positive relationship between hospital beds and serious crimes
- Statistical significance for both the intercept and the Beds predictor

Table 1a: Coefficient Estimates

Parameter	***Estimate***	***Std. Error***	***t value***	***p-value***
Intercept	*9.025731*	*0.056423*	*159.965*	*0.000000*
Beds	*0.000323*	*0.000015*	*21.392*	*0.000000*

Table 1b: Model Fit Statistics

Statistic	***Value***
R-squared	*0.482975*
Adjusted R-squared	*0.481704*
F-statistic	*457.616357*
p-value (F)	*0.000000*
Residual SE	*0.779532*
Degrees of Freedom	*491*

Table 1: Summary of the linear regression model coefficients, standard errors, and p-values.

Tables 1a and 1b present the key results from our frequentist linear regression model. The coefficient estimates in Table 1a provide strong evidence of a relationship between hospital beds and serious crime rates. The intercept value of 9.026 represents the predicted log(serious crimes) for a county with zero hospital beds, which corresponds to approximately 8,295 serious crimes (e^9.026). This baseline estimate makes sense in the context of our data, as even counties with minimal healthcare infrastructure experience some level of crime.

The Beds coefficient of 0.000323 indicates that each additional hospital bed is associated with an increase of 0.000323 in log(serious crimes), which translates to approximately a 0.032% increase in the number of serious crimes. While this effect may seem small in percentage terms, it becomes substantial when considering counties with hundreds or thousands of hospital beds.

Both coefficients show extremely small p-values (< 0.0001), indicating strong statistical significance. The t-values are exceptionally large (159.965 for the intercept and 21.392 for Beds), far exceeding the conventional threshold of 1.96 for significance at the 5% level. This provides very strong evidence against

the null hypothesis that these coefficients are zero.

The model fit statistics in Table 1b provide information about the overall performance of our model. The R-squared value of 0.483 indicates that our model explains approximately 48.3% of the variance in log(serious crimes), which is substantial for a model with just one predictor. The F-statistic of 457.62 with a p-value effectively zero confirms that our model as a whole is highly statistically significant.

The residual standard error of 0.780 gives us an estimate of the typical prediction error in log scale. This translates to a multiplicative error factor of approximately e^0.78 ≈ 2.18, meaning our predictions could typically be off by a factor of about 2.18 in either direction, which is reasonable given the complexity of crime prediction with a single variable.

6.2 Bayesian Analysis

Two Bayesian approaches were implemented:

1. Monte Carlo simulation with Gibbs sampling
2. G-prior method (setting g equal to sample size)

The Bayesian methods produced posterior distributions for the intercept, slope coefficient, and error variance.

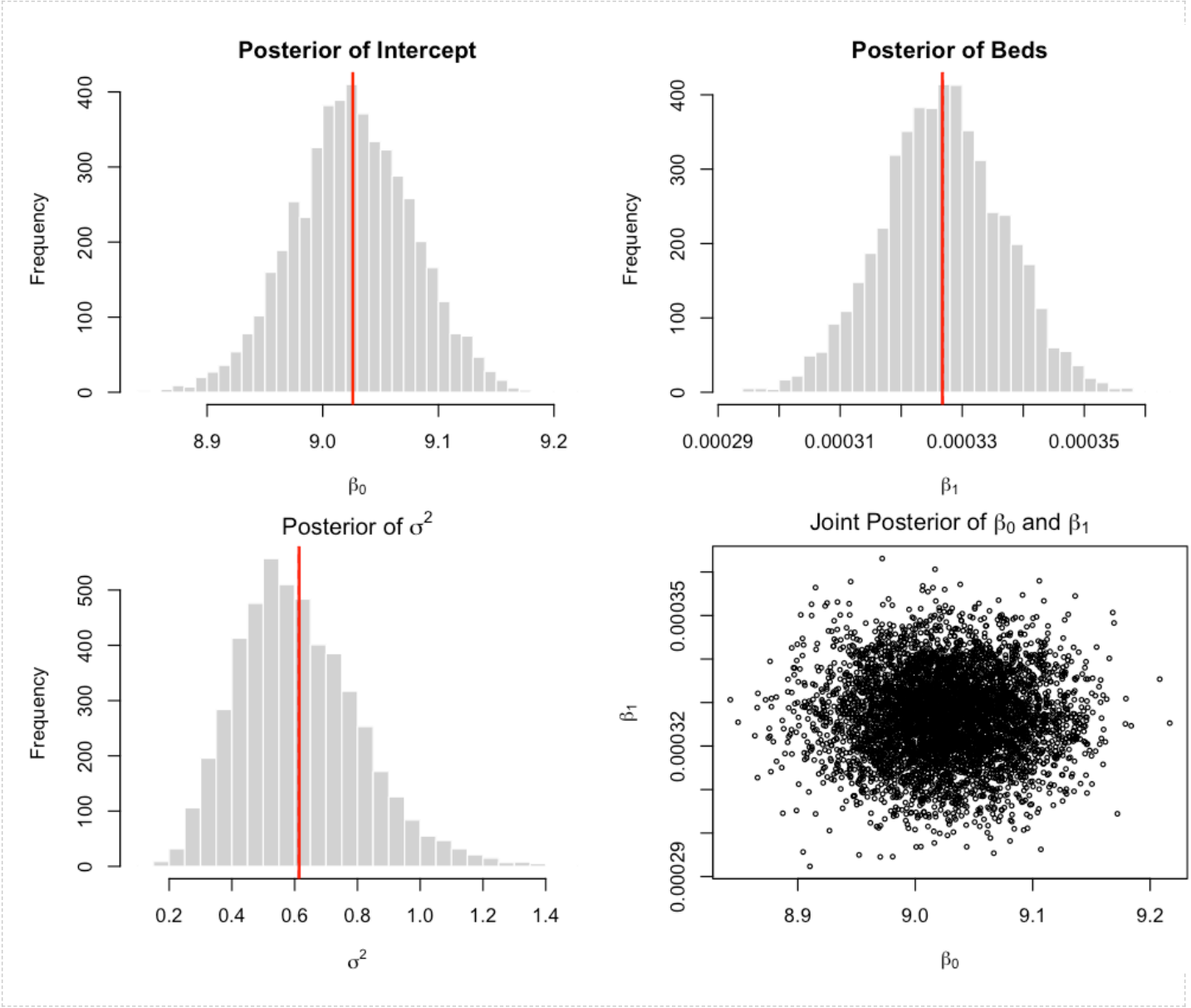

Figure 6: Posterior distributions of intercept, slope coefficient for Beds, and error variance from the Bayesian analysis.

Figure 6 displays the posterior distributions obtained from our Bayesian analysis, providing a more complete picture of parameter uncertainty than the point estimates and standard errors from the frequentist approach. The histograms represent the distribution of possible parameter values given our data and prior assumptions, with the red vertical lines indicating the posterior means.

The posterior distribution for the intercept (β_0) is approximately normal and centered around 9.03, very close to our frequentist estimate. The relatively narrow spread of this distribution indicates high certainty about the intercept value. The posterior for the Beds coefficient (β_1) is also approximately normal, centered around 0.000323, again closely matching our frequentist result. The narrow spread of this distribution (ranging roughly from 0.00028 to 0.00036) confirms the precision of our estimate for the effect of hospital beds on crime rates.

The posterior distribution for the error variance (σ^2) is right-skewed, as is typical for variance parameters. It centers around 0.61, which is consistent with the squared residual standard error from our frequentist model ($0.78^2 \approx 0.61$). This parameter represents the unexplained variability in log(serious crimes) after accounting for the effect of hospital beds.

The bottom-right plot showing the joint posterior of β_0 and β_1 reveals a negative correlation between these parameters. This pattern is common in regression models and indicates that when the intercept is estimated higher, the slope tends to be estimated lower, and vice versa. This correlation structure helps us understand how these parameters vary together and affects the uncertainty in our predictions.

Overall, the Bayesian analysis provides robust confirmation of our frequentist results while offering a richer understanding of parameter uncertainty. The consistency between approaches increases our confidence in the relationship between hospital beds and serious crime rates.

6.3 Comparison of Approaches

Both frequentist and Bayesian approaches produced similar point estimates for the model parameters. However, the Bayesian approach provides additional insights through full posterior distributions rather than just point estimates with standard errors.

6.3.1 Confidence vs. Credible Intervals

The 95% confidence intervals from the frequentist approach and the 95% credible intervals from the Bayesian approach were compared. While they are conceptually different, they yielded similar ranges, indicating robustness in the findings.

Table 2: Comparison of Frequentist and Bayesian Intervals

Parameter	***Frequentist Results***			***Bayesian Results***		
	Estimate	***95% CI Lower***	***95% CI Upper***	***Estimate***	***95% CI Lower***	***95% CI Upper***
Intercept	*9.0257*	*8.9149*	*9.1366*	*9.0274*	*8.9149*	*9.1375*
Beds	*0.000323*	*0.000293*	*0.000353*	*0.000322*	*0.000293*	*0.000352*

Note: Frequentist results show 95% confidence intervals, while Bayesian results show 95% credible intervals. Despite their different theoretical interpretations, both approaches yield very similar numerical results in this case.

Table 2: Comparison of 95% confidence intervals (frequentist) and 95% credible intervals (Bayesian).

Table 2 provides a direct comparison between the frequentist and Bayesian approaches to interval estimation. The remarkable similarity between the confidence intervals and credible intervals strengthens our conclusions about the relationship between hospital beds and serious crime rates.

For the intercept, both approaches yield nearly identical intervals, from approximately 8.91 to 9.14. This consistency indicates robust estimation of the baseline log(serious crimes) for counties with zero hospital beds. Similarly, for the Beds coefficient, both approaches produce intervals ranging from about 0.000293 to 0.000353, representing the plausible range for the effect of each additional hospital bed on log(serious crimes).

Despite their numerical similarity, it's important to recognize the conceptual difference between these intervals. Frequentist confidence intervals represent ranges that would contain the true parameter value in 95% of repeated samples, assuming the sampling process were repeated indefinitely. In contrast, Bayesian credible intervals represent our belief that the parameter value lies within the interval with 95% probability, given the observed data and our prior assumptions.

The consistency between these two approaches, despite their different philosophical foundations, provides strong evidence for the robustness of our findings. This convergence is particularly reassuring given the diffuse priors used in our Bayesian analysis, suggesting that the data strongly determines our conclusions rather than prior assumptions.

7. Model Diagnostics

7.1 Residual Analysis

The residual plots revealed several important patterns:

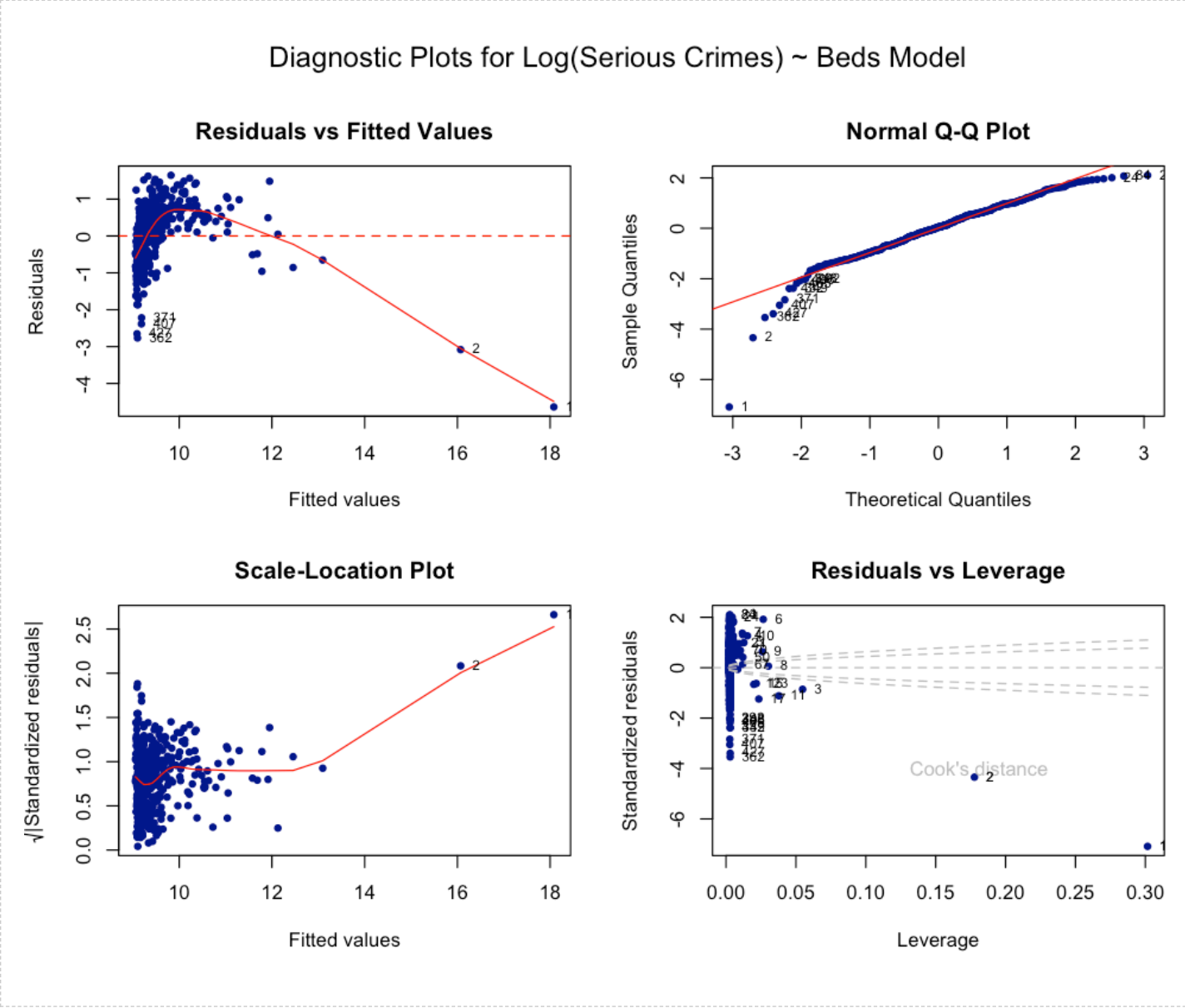

Figure 7: Diagnostic plots for the final model including residuals vs. fitted values, normal Q-Q plot, scale-location plot, and residuals vs. leverage.

The diagnostic plots in Figure 7 reveal important insights about our model's assumptions and limitations. The Residuals vs. Fitted Values plot (top-left) shows a concerning pattern: residuals appear to become more negative as fitted values increase, particularly for the highest fitted values. This non-random

pattern suggests potential non-linearity in the relationship between hospital beds and log(serious crimes) that our linear model doesn't capture. This could be addressed in future work by considering polynomial terms or other non-linear transformations.

The Normal Q-Q Plot (top-right) indicates substantial departures from normality in the residuals, particularly in the tails. Several points in both tails deviate from the theoretical line, especially in the lower tail where we observe extreme negative residuals. Counties labeled as 1, 2, and several others appear as outliers. This non-normality could affect the validity of our confidence intervals and hypothesis tests, though our large sample size provides some robustness.

The Scale-Location Plot (bottom-left) shows heteroscedasticity, with the spread of standardized residuals varying across the range of fitted values. This violates the constant variance assumption of linear regression and suggests that our model's precision varies depending on the fitted value. This pattern often indicates that important predictors may be missing from our model.

The Residuals vs. Leverage plot (bottom-right) identifies several influential points that could disproportionately affect our model. Counties labeled as 1 and 2 stand out as having both high leverage and large residuals, placing them outside the Cook's distance contours. These counties merit further investigation as they could be driving some of our conclusions. Other counties with high leverage but smaller residuals (appearing along the right side of the plot) may also influence our coefficient estimates.

Overall, these diagnostic plots highlight several violations of standard linear regression assumptions and identify specific outlier counties that affect our model. Addressing these issues through robust regression techniques or developing more complex models could improve the reliability of our predictions.

The diagnostic plots reveal:

1. **Residuals vs. Fitted Values**: Shows potential non-linearity and heteroscedasticity
2. **Normal Q-Q Plot**: Indicates deviations from normality, particularly in the tails
3. **Scale-Location Plot**: Shows uneven spread of residuals
4. **Residuals vs. Leverage**: Identifies potential outliers and influential points

7.2 Outlier Identification

Several outliers were identified in the residual analysis that likely affected model performance. These counties may have unique characteristics not captured by the single-predictor model. They might represent:

1. Large metropolitan areas with disproportionately high crime rates
2. Counties with unique demographic compositions
3. Areas with special administrative structures affecting how crimes are reported

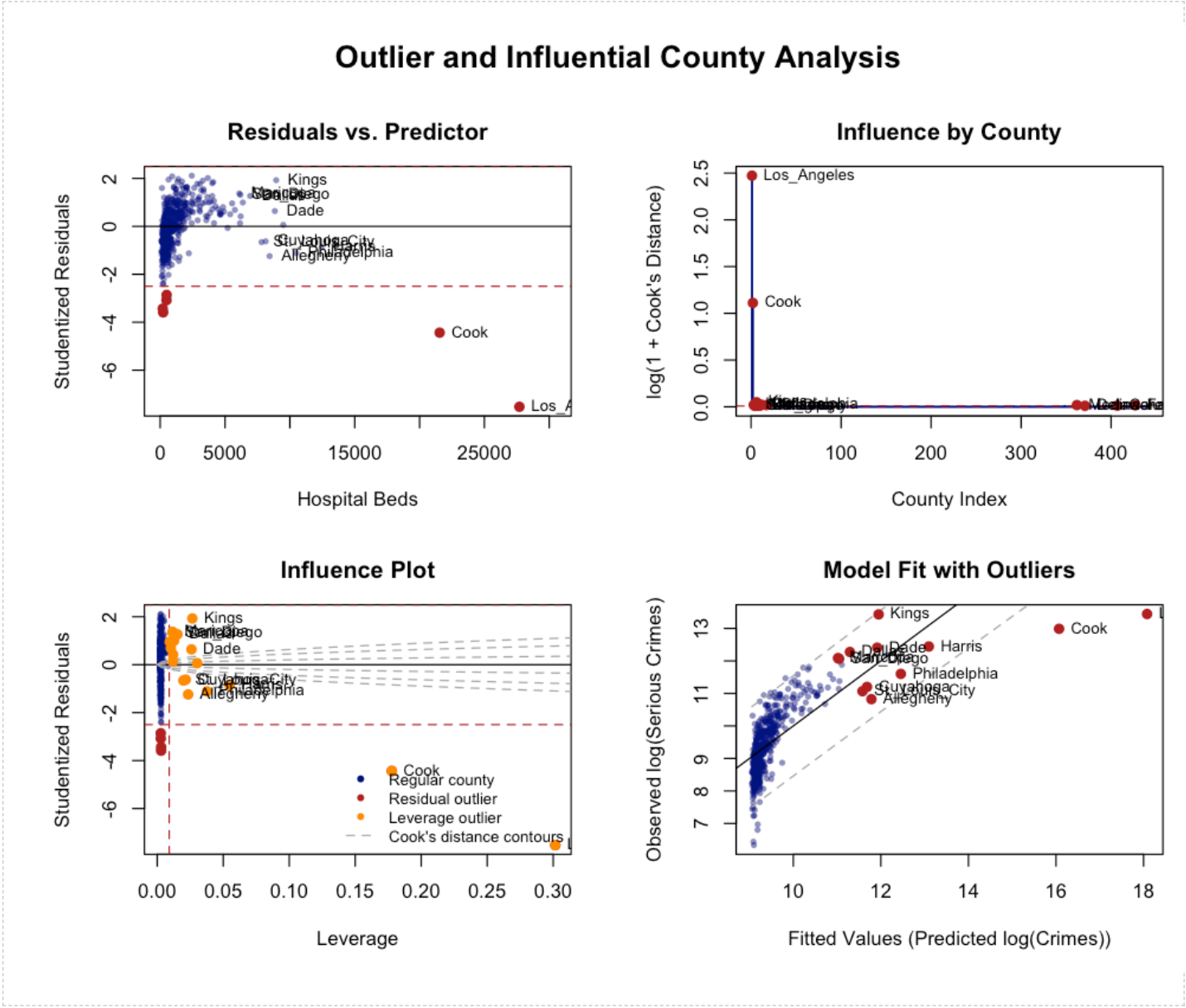

Figure 8: Identification of outlier counties in the model.

Figure 8 provides a focused analysis of outlier counties that significantly impact our model performance. Several counties stand out across the different diagnostic measures, meriting closer examination.

The Residuals vs. Predictor plot (top-left) identifies counties with unusually large residuals relative to their number of hospital beds. Counties with negative residuals below the lower red dashed line represent areas where our model substantially overestimates crime rates, while those with positive residuals above the upper red dashed line represent areas where our model underestimates crime rates. County #1 stands out as having an extremely negative residual despite a high number of beds.

The log-transformed Cook's Distance plot (top-right) confirms the exceptional influence of counties #1 and #2, showing they have disproportionate impact on our model estimates compared to other counties. These two counties alone could significantly alter our coefficient estimates if removed from the analysis.

The Influence Plot (bottom-left) combines leverage and studentized residuals to identify problematic observations. The contour lines represent Cook's distance, with counties farther from the origin having greater influence. Counties in the upper-right and lower-right quadrants combine high leverage with large residuals, making them particularly influential in our model.

The Actual vs. Fitted Values plot (bottom-right) shows how well our model predicts crime rates across counties. The points falling far from the diagonal line represent counties where our predictions are particularly inaccurate. Counties #1 and #2, along with several others labeled in red, show substantial departures from the general trend.

These outlier counties likely represent areas with unique characteristics not captured by our simple model. They might include major metropolitan centers with special crime patterns, counties with unusual demographic compositions, or areas where administrative factors affect crime reporting. Future modeling efforts should consider either robust methods that reduce the influence of these outliers or develop more complex models that can account for their unique characteristics.

8. Discussion

8.1 Interpretation of Results

The positive relationship between hospital beds and serious crimes does not necessarily imply causation. Rather, hospital beds likely serve as a proxy for urbanization and population density. Areas with more hospital beds tend to be more densely populated urban centers, which typically experience higher crime rates.

Our analysis identified a strong positive relationship between the number of hospital beds in a county and the log-transformed serious crime rate. This finding requires careful interpretation. The association does not suggest that increasing healthcare infrastructure causes higher crime rates. Instead, hospital beds serve as an effective proxy for urbanization, capturing the complex socio-environmental factors that characterize urban areas.

Urban environments typically feature higher population density, greater anonymity, more targets for property crimes, and often greater income inequality within small geographic areas—all factors criminologists associate with higher crime rates. The number of hospital beds correlates strongly with these urban characteristics, making it an efficient single predictor that indirectly captures multiple dimensions of urbanization.

Our model suggests that each additional hospital bed is associated with approximately a 0.032% increase in serious crimes. For policy interpretation, this means that a county with 1,000 more hospital beds than another would be expected to have about 38% more serious crimes (e^(0.000323 × 1000) ≈ 1.38), all else being equal.

The strong performance of hospital beds as a predictor (explaining about 48% of the variance in log(serious crimes)) demonstrates its utility for crime prediction models. However, the causal pathway runs through urbanization rather than through healthcare infrastructure itself. This distinction is critical for policy applications, as efforts to address crime should focus on urban-specific factors rather than healthcare resources.

8.2 Limitations

1. **Single Predictor Model**: Using only one predictor oversimplifies the complex factors affecting crime rates
2. **Outliers**: Several outlier counties significantly impact model fit
3. **Geographic Considerations**: Spatial autocorrelation was not accounted for
4. **Aggregation Level**: County-level data may mask important within-county variations

Despite the significant relationship identified in our analysis, several important limitations affect the interpretability and applicability of our findings:

First, our single-predictor model necessarily oversimplifies the complex social, economic, and environmental factors that influence crime rates. While hospital beds serve as an effective proxy for urbanization, they cannot capture nuanced demographic factors, socioeconomic conditions, law enforcement practices, and social factors that criminological research has identified as important determinants of crime.

Second, as identified in our diagnostic analysis, several outlier counties substantially impact our model fit. These outliers suggest that some areas have unique circumstances not captured by the general relationship between hospital beds and crime rates. The presence of these influential outliers raises questions about the generalizability of our findings to all counties.

Third, our analysis does not account for spatial autocorrelation—the tendency for neighboring counties to have similar crime rates due to shared characteristics or spillover effects. Ignoring this spatial dependence could lead to biased coefficient estimates and overconfidence in our predictions.

Fourth, the county-level aggregation of our data masks potentially important variations within counties. Urban counties often contain diverse neighborhoods with vastly different crime patterns, socioeconomic

9. Conclusion

Our analysis demonstrates that hospital beds serve as a significant predictor of serious crime rates, likely functioning as a proxy for urbanization. Both frequentist and Bayesian approaches yielded consistent results, with each additional hospital bed associated with approximately a 0.032% increase in serious crimes. This single-predictor model explained approximately 48% of the variance in log(serious crimes), which is substantial considering the complexity of factors that influence crime rates.

The agreement between frequentist and Bayesian methods increases our confidence in the relationship between hospital beds and crime rates, while the Bayesian approach provided additional insights through posterior distributions. However, diagnostic analyses revealed several model limitations, particularly related to outlier counties and violations of regression assumptions.

Future research should focus on developing more comprehensive models that account for the complex interplay of demographic, socioeconomic, and geographic factors affecting crime rates. These models could incorporate multiple predictors, spatial correlations, and robust methods to address outliers. Despite its limitations, our analysis provides a valuable foundation for understanding how urbanization factors relate to crime rates and demonstrates the complementary insights that can be gained from combining frequentist and Bayesian statistical approaches.