

Name: Esther Akinpelu (3195119)

Course Name and Number: Applied Regression Analysis (STAT 3103-001)

Date of Submission: 15th December 2025

PREDICTION OF CRIME RATE IN BOSTON

Introduction

The Boston Housing dataset provides information on housing and socioeconomic variables for different neighbourhoods in Boston. The primary response variable in this analysis is crime rate per capita (crim), which is known to have a right-skewed distribution. The dataset contains 506 observations with 14 variables, including those related to demographics, housing, and neighbourhood characteristics. The objective of this analysis is to examine which variables significantly predict crime rates in Boston and to develop a suitable prediction regression model. Multiple linear regression is employed, along with transformations, model selection procedures, and diagnostic checks, to ensure model validity.

Literature Review

The Boston Housing dataset has been widely used in applied regression and machine learning research to study relationships between neighbourhood characteristics and key socioeconomic outcomes. Ding (2024) employed this dataset to predict median housing value (medv) using multiple regression and advanced machine learning techniques, including Random Forest and XGBoost. The study focused primarily on predictive performance, evaluating models using R^2 , cross-validated R^2 , and root mean squared error (RMSE). By comparing traditional regression models with more flexible machine learning approaches, Ding demonstrated that nonlinear models can offer improved predictive accuracy when complex relationships exist among predictors.

While Ding's study emphasized prediction rather than inference, it addressed methodological challenges that are also relevant in classical regression analysis, including non-normality of the response variable, multicollinearity among predictors, and the need for model selection. Variables such as crime rate, pollution levels, and accessibility were treated as explanatory factors influencing housing value, highlighting the interconnected nature of neighbourhood characteristics within the Boston dataset.

In contrast, the present study models crime rate per capita (crim) as the response variable, with housing and socioeconomic variables, including median housing value (medv), serving as predictors. This shift in focus allows for direct examination of factors associated with crime rather than housing prices. Despite the difference in response variables, the methodological framework remains comparable, particularly in the use of transformations to address skewness, diagnostic checks to validate model assumptions, and selection procedures to balance model complexity with explanatory power. By extending the application of the Boston Housing dataset to crime rate prediction using multiple linear regression, this analysis contributes to existing literature by emphasizing interpretability and statistical inference alongside predictive accuracy.

Methodology

Multiple linear regression was employed in this analysis because the response variable, crime rate per capita (crim), is continuous. An initial full model containing all explanatory variables was fitted. Diagnostic checks were performed on the residuals to assess model assumptions. A normal Q–Q plot of the residuals indicated right-skewness and heavy upper-tail behaviour, motivating a logarithmic transformation of the response variable. The transformed response, $\log(\text{crim})$, resulted in improved residual normality.

An outlier analysis was conducted on the transformed model using studentized residuals. Observation 311 exhibited a large studentized residual ($r_{\text{student}} = 3.48$). Although the unadjusted p-value suggested that the observation is an outlier, the Bonferroni-adjusted p-value was not statistically significant at the 5% level. As a result, the observation was not automatically classified as a statistically significant outlier.

Multicollinearity among predictors was assessed using a correlation matrix, which revealed a strong correlation between rad and tax. To mitigate multicollinearity, models excluding each variable in turn were evaluated. The model excluding tax yielded a higher adjusted R^2 ; the tax variable was removed from the model for further procedures.

Model selection was conducted using forward selection, backward elimination, and stepwise selection, all based on the Akaike Information Criterion (AIC). Each procedure resulted in the same model comprising nine predictors: rad, nox, zn, lstat, age, black, medv, indus, and ptratio. In addition, subset selection was performed using the `regsubsets()` function from the *leaps* package, with Mallows' C_p , adjusted R^2 , BIC, and mean squared error used as selection criteria to confirm the robustness of the chosen model.

Results

The final regression model includes the seven predictors: zn, nox, age, rad, black, ptratio and lstat. Model selection was guided by Mallows' C_p , and models with C_p values slightly below 12 (k) were identified as candidates. The selected model achieves an adjusted R^2 of 0.8717, indicating that approximately 87% of the variation in crime rate (crim) is explained by these predictors. This adjusted R^2 is only 0.0006 lower than a model including all predictors except tax, showing that the selected model maintains explanatory power while remaining more parsimonious.

Examination of the regression coefficients indicates inverse relationships between the response variable and three predictors: zn (proportion of residential land), ptratio (pupil-teacher ration by town) and black (proportion of Black residents by town). Other predictors in the final regression model show positive associations with crime rate.

Conclusion

This study highlights the significant influence of neighbourhood characteristics on crime rates in Boston. By carefully applying transformations, diagnostic checks, and model selection procedures, a

reliable and interpretable regression model was developed. The analysis underscores the importance of environmental, demographic, and socioeconomic factors in shaping crime patterns, offering insights that could inform urban planning and policy interventions.

The methodological approach demonstrates the value of balancing model complexity with predictive accuracy, ensuring that key predictors are retained while minimizing overfitting. Future research could build on these findings by exploring interactions among predictors, incorporating spatial dependencies to account for geographic effects, or using robust regression methods to further mitigate the influence of outliers.

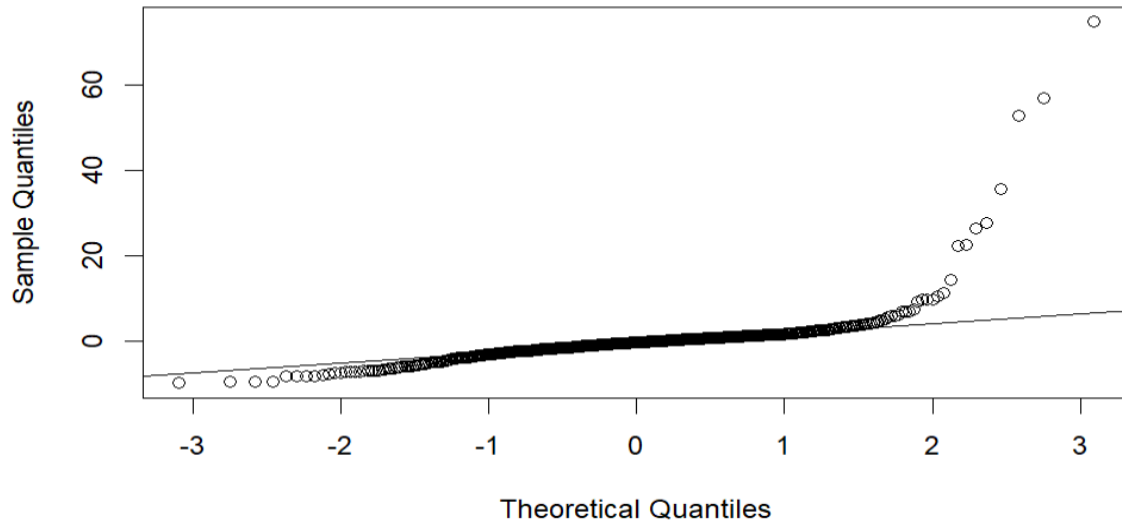
BIBLIOGRAPHY

- Ding, H. (2024). Predicting Boston housing price using machine learning models: Multiple Regression Model, Random Forest, and XGBoost. In B. Siuta-Tokarska (Ed.), Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024) (pp. 439-444). Atlantis Press.
- *Boston housing data*. CRAN. Retrieved December 14, 2025, from <https://search.r-project.org/CRAN/refmans/gausscov/html/boston.html>

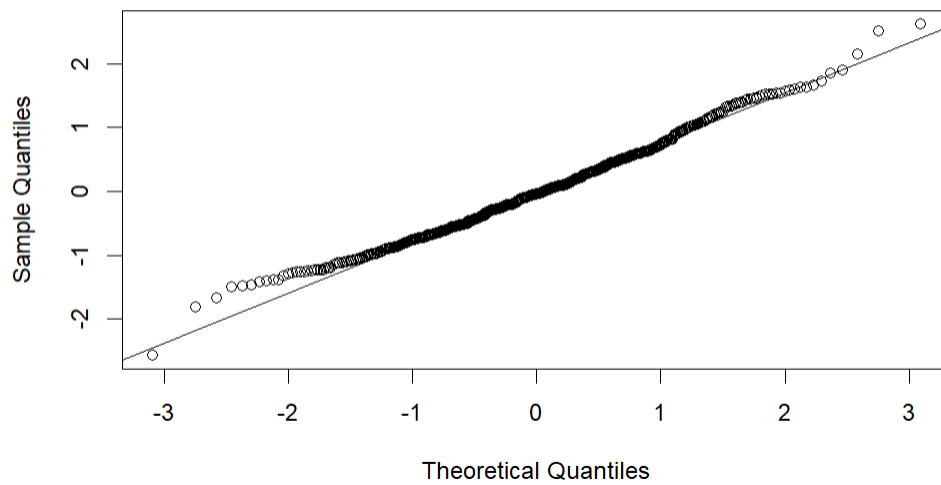
APPENDIX

Dataset: Boston Housing (available in R's MASS package)

Normal Q-Q plot before Transformation



Normal Q-Q plot after Log Transformation



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	1.0	-0.2	0.4	-0.1	0.4	-0.2	0.4	-0.4	0.6	0.6	0.3	-0.4	0.5	-0.4
zn	-0.2	1.0	-0.5	0.0	-0.5	0.3	-0.6	0.7	-0.3	-0.3	-0.4	0.2	-0.4	0.4
indus	0.4	-0.5	1.0	0.1	0.8	-0.4	0.6	-0.7	0.6	0.7	0.4	-0.4	0.6	-0.5
chas	-0.1	0.0	0.1	1.0	0.1	0.1	0.1	-0.1	0.0	0.0	-0.1	0.0	-0.1	0.2
nox	0.4	-0.5	0.8	0.1	1.0	-0.3	0.7	-0.8	0.6	0.7	0.2	-0.4	0.6	-0.4
rm	-0.2	0.3	-0.4	0.1	-0.3	1.0	-0.2	0.2	-0.2	-0.3	-0.4	0.1	-0.6	0.7
age	0.4	-0.6	0.6	0.1	0.7	-0.2	1.0	-0.7	0.5	0.5	0.3	-0.3	0.6	-0.4
dis	-0.4	0.7	-0.7	-0.1	-0.8	0.2	-0.7	1.0	-0.5	-0.5	-0.2	0.3	-0.5	0.2
rad	0.6	-0.3	0.6	0.0	0.6	-0.2	0.5	-0.5	1.0	0.9	0.5	-0.4	0.5	-0.4
tax	0.6	-0.3	0.7	0.0	0.7	-0.3	0.5	-0.5	0.9	1.0	0.5	-0.4	0.5	-0.5
ptratio	0.3	-0.4	0.4	-0.1	0.2	-0.4	0.3	-0.2	0.5	0.5	1.0	-0.2	0.4	-0.5
black	-0.4	0.2	-0.4	0.0	-0.4	0.1	-0.3	0.3	-0.4	-0.4	-0.2	1.0	-0.4	0.3
lstat	0.5	-0.4	0.6	-0.1	0.6	-0.6	0.6	-0.5	0.5	0.5	0.4	-0.4	1.0	-0.7
medv	-0.4	0.4	-0.5	0.2	-0.4	0.7	-0.4	0.2	-0.4	-0.5	-0.5	0.3	-0.7	1.0

Correlation matrix of the Boston dataset.

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
311	3.483374	0.00053942	0.27295
1 row			

The outlier discovered.

DECLARATION OF AUTHORSHIP

I declare that this submission is my own work and that all sources used have been properly acknowledged. This work has not been submitted, in whole or in part, for any other course.

Full Name: Esther Akinpelu

Date: December 15, 2025

Signature: ____ E.A. ____