Portfolio Project

Cassidy Halpin

MIS500-1

October 6, 2024

**Figure 1**

*Code for Linear Regression*

**A screenshot of a computer

Description automatically generated**

**Figure 2**

*Linear Regression Statistic results*

A screenshot of a computer

Description automatically generated

**Figure 3**

*Observed v. Predicted scatterplot*

A screenshot of a computer

Description automatically generated

**Figure 4**

*Residual Diagnostics*

A screenshot of a computer

Description automatically generated

**Figure 5**

*Residual by Regressor plots*

A screenshot of a computer

Description automatically generated

**Figure 6**

*Code for Stepwise Linear Regression*

A screenshot of a computer

Description automatically generated

**Figure 7**

*Stepwise selection summary*

A screenshot of a computer

Description automatically generated

**Figure 8**

*Fit Criteria for Crime in Florida.csv*

A screenshot of a computer

Description automatically generated

**Figure 9**

*ANOVA, R-squared, and parameter estimates*

A screenshot of a computer

Description automatically generated

**Figure 10**

*Observed versus predicted for stepwise linear regression*

A screenshot of a computer

Description automatically generated

**Figure 11**

*Residual Diagnostics for stepwise linear regression*

A screenshot of a computer

Description automatically generated

**Figure 12**

*Residuals by regressors plot for stepwise linear regression*

A screenshot of a computer

Description automatically generated

This analysis utilizes both linear and stepwise regression methods to investigate the relationship between **Income, High School Graduation Rates (HS)**, and **Urbanization (Urban)** on **Crime** using crime data from Florida counties. Linear regression was first used to model how these variables influence crime rates, followed by a stepwise regression to refine the model by selecting the most significant predictors.

In figure 2 the statistical summary of the linear regression can be seen. The F-value of 18.83 with a p-value < 0.0001 means the overall regression model is statistically significant. This is further justified with the R-squared value of 0.4728, which approximates 47.28% of the variance in Crime within Florida counties can be explained by the three predictor values. Each predictor however does not hold the same significance. Income with a p-value of 0.6852 is not a significant predictor of crime. High School Graduation rate with a p-value of 0.4025 is also not statistically significant. However, Urban has a p-value of <0.0001 indicating the percentage of urbanisation could be a strong predictor for crime. The observed vs. predicted plot indicates through the scatter around the line that there is a general relationship of these predictors towards crime, but there is some variance that leaves room for improvement. The same relationship can be seen in the residual diagnostics and the residual by regressor plots.

To refine the linear regression model of crime in Florida counties the stepwise selection method was used, and the selection stopped when the Schwarz Bayesian Criterion reached a local minimum of 413.7016 after including Urban as the predictor value. This suggest that Urban improved the model significantly compared to the other independent variables. Within the fit criteria plots in Figure 8, the four plots show that the Urban variable improved the model fit across all metrics. The model’s F-value of 55.11 has a p-value of < 0.0001, indicating the overall model is statistically significant. The total variance of the model is 52462 which means the model explains roughly half of the variation in crime. The Adjusted R-squared value is 0.4505, which accounts for the number of predictors. This value is very close to the regular R-squared of 0.4588 indicating that the model is well-specified without overfitting. The coefficient for Urban in the model is 0.562, this suggests that for every one unit increase in Urban, the crime rate increases by 0.562. The model has good fit and explains nearly half of the variance in Crime using Urban as a predictor. However, the model could potentially be improved by addressing some minor heteroscedasticity as seen in Figure 11 and looking into possible transformations or additional predictors that could explain the remaining unexplained variance.

While Urban is a highly significant predictor of crime in Florida counties, Income and High School Graduation rates are not. When transforming the model using Stepwise selection the Urban predictor was validated with a highly significant relationship indicating that as urbanization increases, crime rates tend to rise. The model explains approximately **45.88%** of the variation in Crime (R-squared), which indicates a moderate level of fit. While the model captures key relationships, there is still room for improvement, as nearly half of the variance in crime remains unexplained.