Executive Summary:

The initial step to find the best fit model for predicting crimes per capita was to create a baseline comparison using the overall model regarding all possible factors. The overall model with 15 predictors showed statistical significance with, $F(15, 424, .05) = 63.02, p = .00, R^2 = .69$ and $Adj R^2 = .679$. Thus, the model supported 69% of the variance; however, there were two major factors that needed to be considered before prescribing such a model for crime per capita prediction. First, this model had several nonstatistically significant factors within the model. And second, this model had several factors with inflated VIF values in addition to whether or not their $b$ weights were statistically significant. In order for appropriate model selection, an adjust $R$ squared command with the AIC and Cp statistic was run in SAS for the initial selection.

The model selected upon the highest adjusted $R$ squared with account of the lowest AIC score and a Cp closest to $p$ showed $area$, $population$, percentage of individuals ages 18 to 34, percentage of individuals over the age 65, number of hospital beds, number of crimes, percentage of bachelor degrees, $income per capita$, $total personal income$, $region$, and $population density$ as the best fit model. This model showed statistical significance with, $F(12, 427, .05) = 77.77, p = .00, R^2 = .68, Adj R^2 = .6773$. It should be noted that $total population$ and $total personal income$ had VIF values over 100. Such values suggest multicollinearity between the predictors in the model. To alleviate this issue, any predictor with a VIF value over 10 or with a nonsignificant $b$ weight was eliminated. Additionally, a log-log transformation was done to reduce VIF values and increase predicted explained variance. This final model was denoted as:

$$\text{Logcrimescap} = -8.04 - 0.0979 \text{Logland} + 0.403 \text{Log}\%18-34 + 0.273 \text{Log}\%65 + 0.505 \text{Logcrimes} + 0.279 \text{Logbeds} + 0.308 \text{Logpercentpoverty} + 0.170 \text{Logregion}$$

This model showed, $F(7, 432, .05) = 213.81, p = .00, R^2 = .776, Adj. R^2 = .772$. It must be noted that all predictors in the model were statistically significant, as LogLand, LogAge18-34, LogOver65, Logcrimes, Loghospitalbeds, Logpercentpoverty, and Logregion were $t(432) = 6.93, p = .00$, $t(432) = 3.44, p = .001$, $t(432) = 4.57, p = .00$, $t(432) = 21.99, p = .00$, $t(432) = 10.86, p = .00$, $t(432) = 6.93, p = .00$, and $t(432) = 5.70, p = .00$ respectively. It should further be noted that all VIF values were all below 10; thus, multicollinearity was not an issue between predictors. It should be addressed that the Mean Squared Error increased in comparison to the overall model; however, the Mean Squared Treatment increased to an even greater extent and thus created an $F$ ratio greater than the overall model. Finally for what it’s worth, the predicted $R$ squared value also went from $Predicted R^2 = .48$ in the overall model to $Predicted R^2 = .76$ in the fitted model.