Programming Exercise 7 - Assignment Sheet

Crime Analytics (CJUS 6106)

Instructions

For this exercise, you will be tasked with estimating linear regression models, creating figures describing bivariate and multivariate relationships between select variables, and evaluating some of the important assumptions of linear regression. You will be using the state_data file I have used in prior lectures. Note that this file must be knitted as a PDF. I strongly encourage you to use the .rmd file I provide for the exercise to begin your assignment.

Question 1

Create a variable called **total_crime** within the **state_data** data frame that is the sum of the seven individual crime rate variables. Provide a summary of this new variable using the **summary()** function and interpret its mean. Note that all of these variables are measured as crime rates per 100,000.

Question 2

Estimate an intercept-only linear regression model with the **total_crime** variable as the outcome. Summarize this model and explain why the intercept value is equivalent to the mean of the **total_crime** rate variable. Next, add the following independent variables to the model and provide an interpretation for each coefficient: the percent under the poverty line (**poverty**), the average state resident income in dollars (**avwage**), the percent income share for the top 0.1% of the population (**top_1**), the percent of the state population that is unemployed (**urate**), and the incarceration rate per 1,000 population (**inc_rate**). Also be sure to indicate when a coefficient is significant and at what probability value (e.g., $p<.05$, $p<.01$, $p<.001$). You should also store this regression model as an object because you will need to use these results in another question.

Question 3

Create an added variable plot using the stored regression model from the prior question and limit the graph output to just the **top_1** variable as I did in lecture.

Now, create a plot using ggplot that depicts the bivariate relationship between the top 0.1% income share and the total crime rate. Be sure to include a best fit line as I do in lecture.

Finally, comment on any differences you observe in these relationships. Do our inferences about the relationship between these variables change when we control for other variables? If so, how?

Question 4

Create a new variable called **incrate_quint** that is the incarceration rate per 1,000 divided into five equal quintiles. You will need to use the ntiles() function to accomplish this. Be sure to also make this a factor variable.

After you do so, estimate a new regression model predicting the total crime rate using only the quintile variable you just created. Confirm that the coefficients from the different quintiles reproduce the mean total crime rates for each category of this variable. Provide an interpretation for why these values should be the same.

Question 5

For this question, you will return to the multivariate regression model you estimated in Question 2 that you should have stored as an object. Using the results from this model, create a component plus residual plot for the **top_1** regressor.

Provide a comment about whether there is any evidence of nonlinearity in the relationship between **top_1** and **total_crime**.

Next, create a squared version of the **top_1** variable within the **state_data** data frame. Re-estimate the regression model from Question 2 including the new squared version of **top_1** (note - there should be six total independent variables in your model now).

Using the equation from the regression analysis lecture, compute the inflection point for the **top_1** variable where the slope turns to zero. Provide an interpretation of what this value means.

Question 6

Using the appropriate graphs and statistical tests (demonstrated in the regression analysis lecture), provide an indication for whether the regression model you estimated in Question 5 (with the squared term!) violates either of the assumptions that the residuals are 1) homoscedastic or 2) normally distributed.