**8.1**
Linear Regression can be used for forecasting voter turnout in a given district based on demographics and population. Predictors could be average age, percent of the population with college degrees, average income, gender breakdown, or unemployment rate.

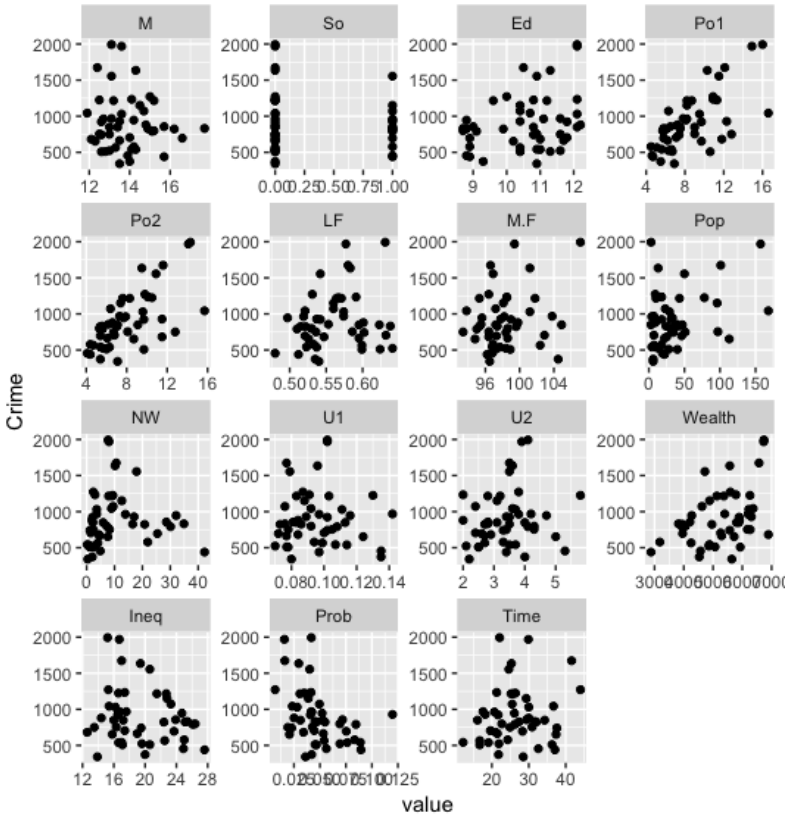
**8.2 - Please review the included 8.2.R file with this write up.**
**Summary**:
I used a linear regression model with the variables M + Ed + Po1 + U2 + Ineq + Prob. The predicted crime rate on the hypothetical city data with this model is 1304. The equation for this model is:

$$-5040.51 + (105.01957 * M) + (196.47 * Ed) + (115.02 * Po1) + (89.37 * U2) + (67.65 * Ineq) - (3801.84 * Prob)$$

**Detailed Explanation**:
First, I loaded and inspected the data. With 16 variables and only 47 observations, it's likely that a model using all of the variables will heavily over fit. With only 47 observations I decided not to split into train/test/val splits to keep as much data as possible to fit the regression. The issue of training and validating on the same model is alleviated by using AIC, BIC, and Adjusted-R Squared for model selection as these approaches can help control overfitting on training data by penalizing models that use unnecessary variables as predictors.

Because this is the same data set we used in week three for outlier detection, I considered removing the two data points that are potential outliers. As a reminder, there are two potential outliers in the data set. The grubbs test finds two outliers at a p-value of .1, but none at .05. Based on this finding I would likely not remove them, but it could possibly help with model accuracy. Ultimately, I decided not to remove the outliers after determining that removal did not improve the quality of fit of my models.

After deciding not to remove any potential outliers, I plotted each of the variables against crime rates in each city to get a general idea of which variables might have a good linear fit to the data.

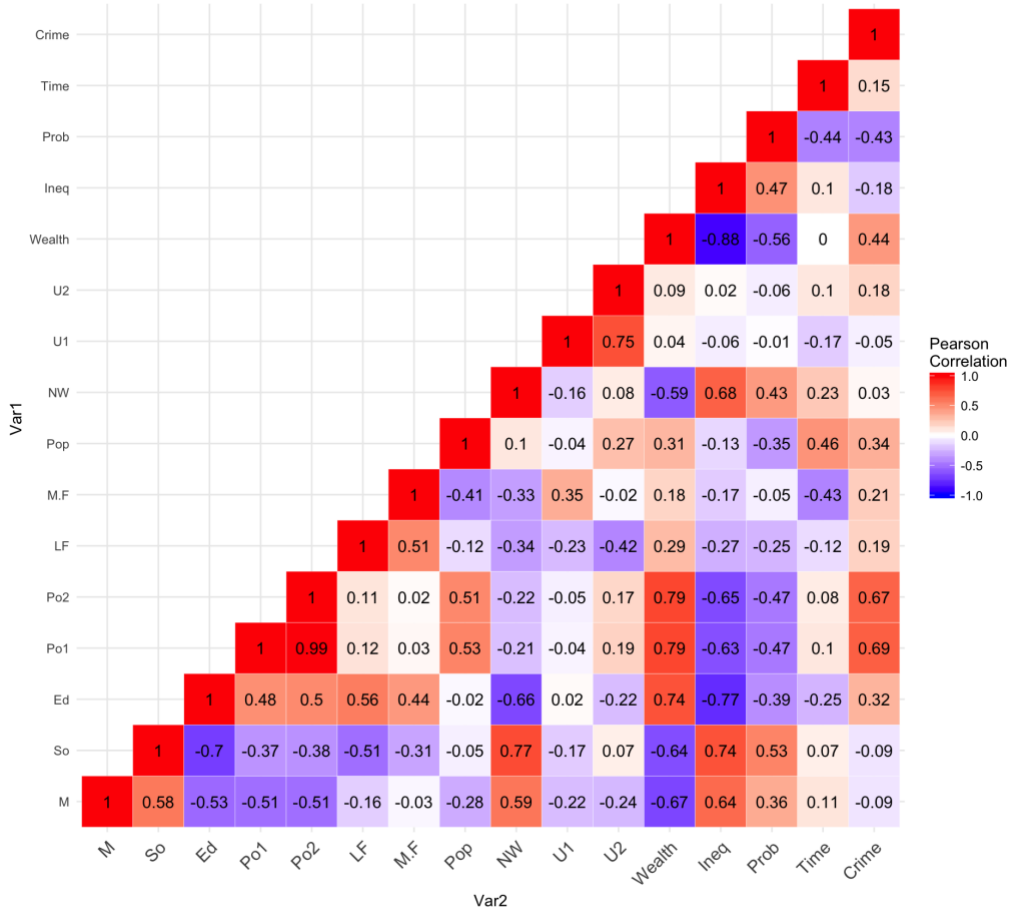


It looks like Education, Police Expenditure, and Wealth have positive correlations, while Inequality and Imprisonment have negative correlations, but we can figure out which are significant while building the model.

I built various models with a variety of different variables. First, I tried using all of the variables to see which were significant. M + Ed + Po1 + U2 + Ineq + Prob were all significant at a p-value of .1, while M + Ed + Ineq + Prob are significant at .05. I tried models using both of these sets of variables. I found that the model using the variables significant at .1 had the best balance of removing insignificant variables while still having strong predictive strength, with the lowest AIC and BIC, and the highest Adjusted R-squared. Quality statistics for the three initial models are below. At this point, it looks like the P .1 model is best, so I'll move forward with this model

| | Base Crime Model | Crime Model P .1 | Crime Model P .05 |
|---|---|---|---|
| AIC | 650.03 | **640.17** | 690.07 |
| BIC | 681.46 | **654.97** | 701.17 |
| Adj. R - Square | 0.7078 | **0.7307** | 0.1927 |

I decided to check for multicollinearity between variables to see if removing any variables from the P.1 model could improve it. A correlation matrix is below:




Of the variables in the p.1 model, the only variables with significant correlation are Education and Inequality. I tried a version of the p.1 model with the Education variable removed, but it had a negative impact on AIC, BIC, and adjusted R-squared. I also tried removing Inequality and had a similar result.

I also considered removing the Prob and U2 variables from the p.1 model, as they had the highest p values and were not significant at P.01. This also had a negative impact, so I decided to stick with the initial p.1 model.

I created a test city data frame using the values from the homework assignment and used the P.1model to predict the crime rate in the test city. The predicted crime rate was 1304. This would be the sixth highest crime rate in the data set if it was included in the initial data. However there is no reason to believe this is an incorrect prediction as the values for the predictive variables were also generally on the extreme ends of the range for each predictor.