Note: This represents only a single possible correct set of solutions. Homework 6 and 7 will not have a "correct" answer — rather, the proper analyses and critical thinking applied to the problem will be varying levels of "correct."

**Problem 1**: 100 points
Using the data file `UScrime.csv`, fit a model that can be used to predict the rate of offenses per 100,000 population in 1960. Find a first-order model that achieves $R_a^2 \geq 0.73$. You may use the entire dataset for training. Complete a regression analysis as detailed in the instructions above and report your results.

The data includes the following columns.

| Variable | Description |
|---|---|
| `M` | percentage of males aged 14–24 in total state population |
| `So` | indicator variable for a southern state |
| `Ed` | mean years of schooling of the population aged 25 years or over |
| `Po1` | per capita expenditure on police protection in 1960 |
| `Po2` | per capita expenditure on police protection in 1959 |
| `LF` | labour force participation rate of civilian urban males in the age-group 14-24 |
| `MF` | number of males per 100 females |
| `Pop` | state population in 1960 in hundred thousands |
| `NW` | percentage of nonwhites in the population |
| `U1` | unemployment rate of urban males 14–24 |
| `U2` | unemployment rate of urban males 35–39 |
| `Wealth` | wealth: median value of transferable assets or family income |
| `Ineq` | income inequality: percentage of families earning below half the median income |
| `Prob` | probability of imprisonment: ratio of number of imprisonments to number of offenses |
| `Time` | average time in months served by offenders in state prisons before their first release |
| `Crime` | crime rate: number of offenses per 100,000 population in 1960 |

Helpful hint: You'll only want one of Po1 and Po2, and only one of U1 and U2, to remain in the final model due to high multicollinearity.

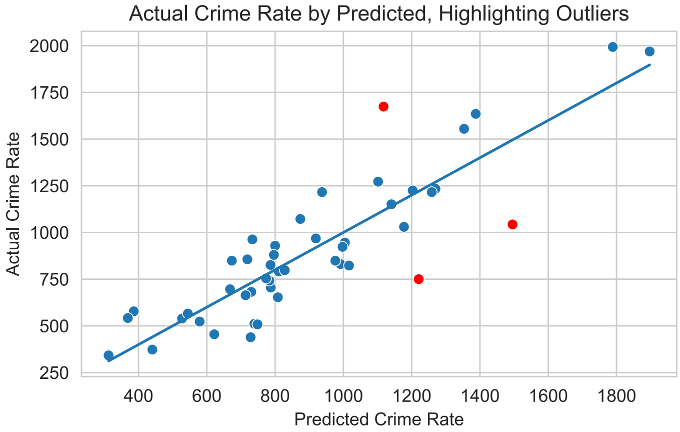
Once your model is complete, noting that this is not implying a causal relationship, comment on the level of association (slope) between each variable in your model with crime rates. What does a positive/negative slope mean?

**Executive Summary**: A linear regression model predicting Crime Rate in 1960 was able to explain 76.6% of the variability in Crime Rate with the percentage of young males in the state population, mean years of schooling of the adult population, per capita police expenditure in 1960, unemployment rate of urban males aged 35–39, income inequality, and the probability of imprisonment after committing a crime. Each factor was positively correlated with Crime Rate except for probability of imprisonment.

**Predicted model**

$$\hat{y} = -5040.5 + 105.0\texttt{M} + 196.5\texttt{Ed} + 115.0\texttt{Po1} + 89.37\texttt{U2} + 67.65\texttt{Ineq} - 3802\texttt{Prob}$$

A scatterplot of the hypothesized model — actual observed crime rate by predicted crime rate — is shown below. Three observations with high-magnitude residuals are highlighted. One (row 10) may provide a case study for an area with higher-than-expected crime. Two others (rows 18 and 28) may provide case studies for areas with lower-than-expected crime.

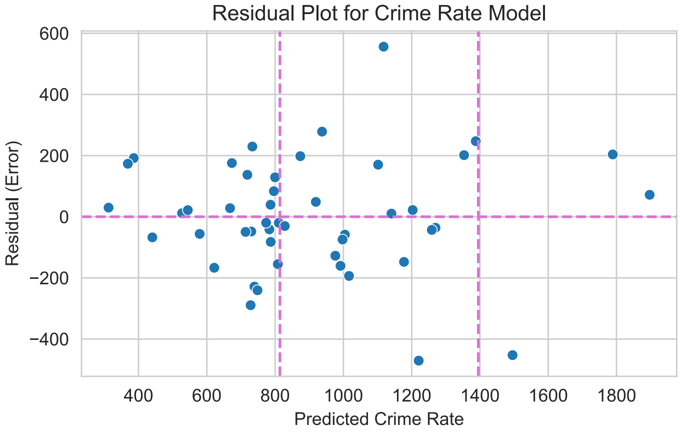


**Coefficient of determination**: $R^2 = 0.766$

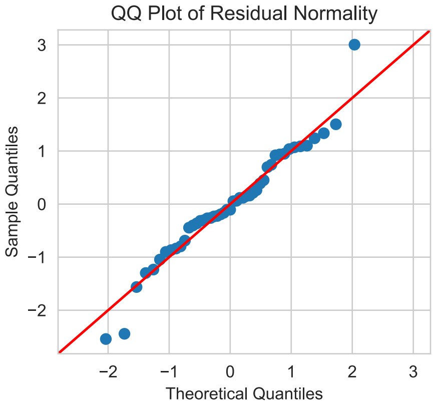
**Test for validity**: Testing the null hypothesis that the model has no predictive value against the alternative that it does results in a test statistic of $F = 21.81$ and a p-value $p < 0.0001$, allowing us to reject the null hypothesis. Each parameter of the model has an individual p-value $p < 0.05$, indicating significant evidence against and allowing us to reject the null hypotheses that each slope equals 0.

**Validation of Assumptions**

There is no apparent pattern in the mean or variance of residuals as shown in the residual plot below. Overall the predicted crime rate is right-skewed, giving us very little to assess the distribution for high crime rates.



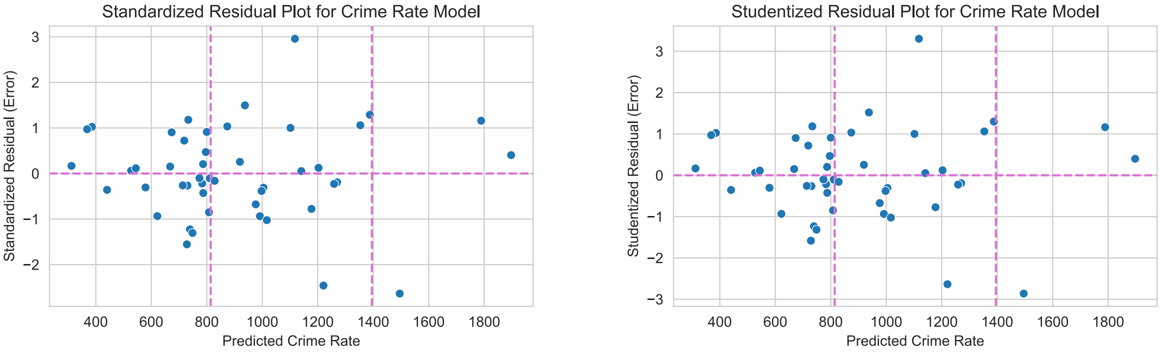
There does not appear to be strong deviation from normality in the residuals, although there are 3 extreme points which appear more extreme than expected. See the QQ plot below. The Shapiro-Wilk test returns a p-value of $p = 0.24$, which does not indicate significant deviation from normality.

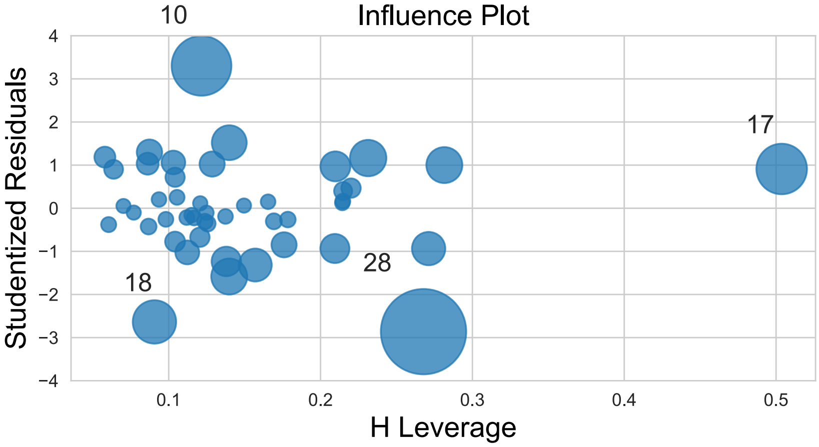


Without data on ‘date collected’ or ‘geographical coordinates,’ there is no apparent way to check for deviations from independence of errors.

**Other Diagnostics**

The standardized and Studentized residual plots further reinforce the value of further examining the highlighted outlier data points. However, the influence plot does not indicate that those values are generating undue influence upon the model — none of those points have both high-magnitude Studentized residuals and high leverage — so they are not a concern from a modeling perspective.





None of the Variance Inflation Factors indicate significant issues with multicollinearity.

| VIF | Model Term |
|---:|---|
| 945 | Intercept |
| 2.00 | M |
| 2.86 | Ed |
| 1.91 | Po1 |
| 1.36 | U2 |
| 3.53 | Ineq |
| 1.38 | Prob |

**Comments on levels of association**

The following relationships were included in the model. These correlations do not indicate causation but rather that they vary together in a predictable way.

*M: percentage of males aged 14–24 in total state population.* The predicted slope of the relationship between `M` and crime rate is 105, indicating that for every increase of 1% of the population that are males aged 14–24, 105 more offenses are expected per 100,000 population.

*Ed: mean years of schooling of the population aged 25 years or over.* The predicted slope of the relationship between `Ed` and crime rate is 196, indicating that for every increase of 1 mean year of schooling, 196 more offenses are expected per 100,000 population.

*Po1: per capita expenditure on police protection in 1960.* The predicted slope of the relationship between `Po1` and crime rate is 115, indicating that for every increase of $1 spent per capita on policing, 115 more offenses are expected per 100,000 population.

*U2: Unemployment rate of urban males aged 35–39.* The predicted slope of the relationship between `U2` and crime rate is 89.4, indicating that for every increase in unemployment rate of 1%, 89.4 more offenses are expected per 100,000 population.

*Ineq: percentage of families earning below half the median income.* The predicted slope of the relationship between `Ineq` and crime rate is 67.7, indicating that for every increase of 1% of families earning less than half the median income, 67.7 more offenses are expected per 100,000 population.

*Prob: ratio of number of commitments to number of offenses.* The predicted slope of the relationship between `Prob` and crime rate is -3802, indicating that for every increase of 1% in the percentage of offenses resulting in imprisonment, 38 fewer offenses are expected per 100,000 population.