1 Instructions

This will require a regression analysis. A regression analysis should include the following information.

- Non-technical summary of analysis
- Hypothesized Model: State your hypothesized model as $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots$ with numerical values for each β_i and defining each x_i.
- A scatterplot comparing your hypothesized model's predicted values (x-axis) to the observed data (y-axis).
- Parameter estimates and confidence intervals where requested (unless otherwise noted, use the t-distribution for tests/intervals)
- Coefficient of determination
- A test for validity in your model (slope(s) and/or coefficient of correlation). See Lesson 33. Remember all parts of a hypothesis test.
  - Hypotheses
  - Significance
  - Identify the test being used
  - Either p-value, or test statistic and rejection region
  - Technical conclusion
- Predictions and estimates (where requested) for your model. (See Lesson 33)
- Validation of your regression model's adequacy. (See Lesson 37)
  - Graphically confirm mean and variance is constant for all predicted values.
  - Graphically and analytically confirm normality of residuals.
  - If possible, confirm that errors are independent. If not, state that.
  - Identify any points of concern such as outliers and influence points.
  - Identify potential issues with multicollinearity by assessing Variance Inflation Factors.

2 Questions

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?