Problem 1
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.

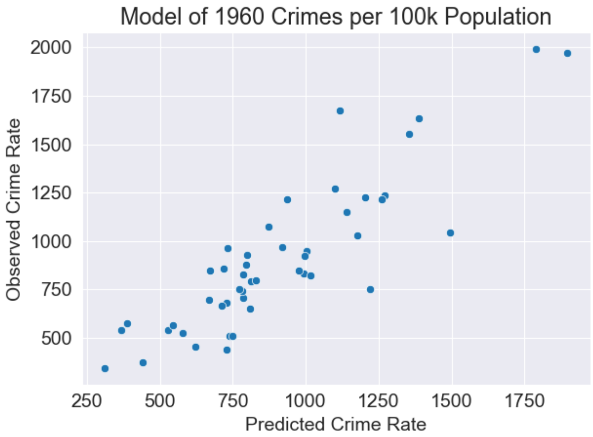
The data includes the following columns.

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?

- BLUF: There is enough evidence to determine that the model can be used to predict the crime rate for 1960.
- Hypothesized model: $\hat{y} = -5040.51 + 115.02x_1 + 67.65x_2 + 196.47x_3 + 105.02x_4 - 3801.84x_5 + 89.37x_6$
Where x_1 is Po1, x_2 is Ineq, x_3 is Ed, x_4 is M, x_5 is Prob, and x_6 is U2.
- Scatterplot of the hypothesized model's predicted values to the observed data.



Parameter	Coefficient	t	p-value	Lower CI	Upper CI
Intercept	$\hat{\beta}_0 = -5040.51$	-5.60	< 0.0001	-6859.16	-3221.85
Po1	$\hat{\beta}_1 = 115.02$	8.36	< 0.0001	87.23	142.82
Ineq	$\hat{\beta}_2 = 67.65$	4.85	< 0.0001	39.49	95.82
Ed	$\hat{\beta}_3 = 196.47$	4.39	< 0.0001	106.02	286.92
M	$\hat{\beta}_4 = 105.02$	3.15	0.0031	37.72	172.32
Prob	$\hat{\beta}_5 = -3801.84$	-2.49	0.0171	-6890.24	-713.44
U2	$\hat{\beta}_6 = 89.37$	2.18	0.0348	6.69	172.04

Table 1: Model results, with 95% confidence intervals.

- Parameter estimates: See the Coefficient column of Table 1.
- Coefficient of determination: $R^2 = 0.766$, $R_a^2 = 0.731$
- Hypothesis tests
  - Hypotheses:
    For the entire model: $H_0 : \beta_i = 0$ for all β_i. $H_a : \beta_i \neq 0$ for at least one β_i.
    For each parameter (considered separate individual tests): $H_0 : \beta_i = 0$, $H_a : \beta_i \neq 0$
  - Significance: $\alpha = 0.05$
  - Test Statistic: Model: $F = 21.81$. For parameters see t column of Table 1.
  - P-value: Model: $p = 3.42 \times 10^{-11}$. For parameters see p-value column of Table 1.
  - Technical Conclusion: Model: We reject the null hypothesis and conclude that at least one of the coefficients is not equal to zero. Looking at each parameter in Table 1 individually, we can also reject the null hypothesis for each one and conclude that the coefficient for each is not equal to zero.
- Validation of adequacy
  - Constant mean and variance.
    The mean of the residuals is approximately zero. Figure 1 shows the residuals plotted against the predicted values for crime rate. There is no apparent pattern to show that the mean or variance change over the predicted values.

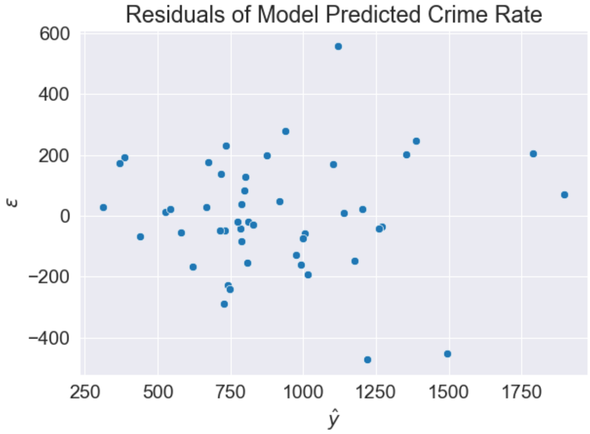


Figure 1

- Normality of residuals

  The histogram and QQ-plot of the residuals in Figure 2 shows the distribution is approximately normal. To confirm analytically I ran an Shapiro test since the data set is small ($n < 100$). For the test the null hypothesis is that the distribution is normal. The resulting p-value was 0.24. With such a high value we fail to reject the null hypothesis, supporting our assumptions determined graphically.

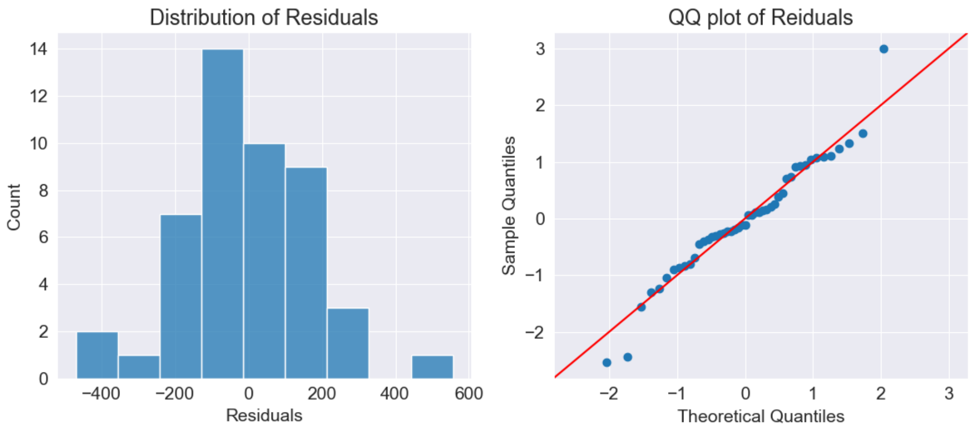


Figure 2

- Errors are independent

  There is no time series component that would cause dependency. There could be a dependency based on geographic regions of the US. However the only regional information is whether or not the state is a southern state. Figure 3 shows the residual plot again but with the southern states marked in red. This shows no dependency, and I could not think of any other way to test. So I assume the errors are independent.

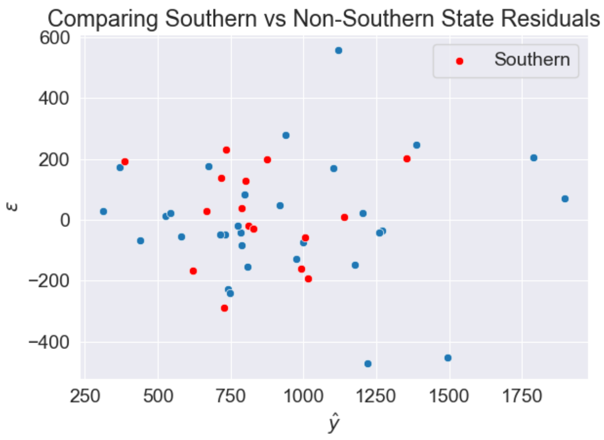


Figure 3

- Outliers and influence points

  The plot of the residuals (Figures 1 and 3) do show three potential outliers where $|\epsilon| > 400$. I used a influence plot to identify influence points and there were four. Figure 4 shows these points in orange on the residual plot. The three suspected outliers are influence points. The last one is surprising as it is around several other points, so I would not consider in an influence point.

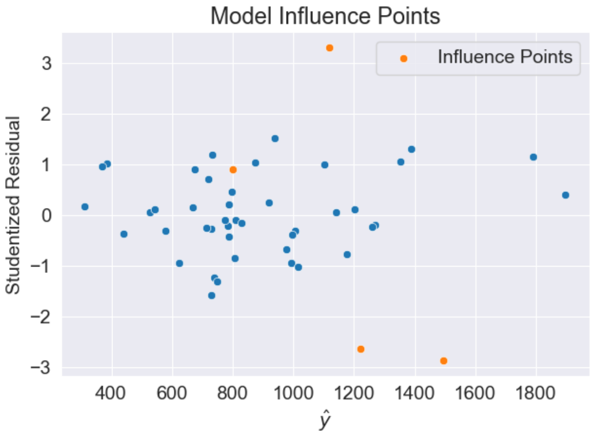


Figure 4

– Multicollinearity
  To check the multicollinearity of the variables in the model I used the Variance Inflation Factor (VIF). The results are shown in Table 2. Since all the VIFs are low, multicollinearity is not a concern for the model.

Variable	VIF
Po1	1.91
Ineq	3.53
Ed	2.86
M	2.00
Prob	1.38
U2	1.36

Table 2

Summary of Model Results: From Table 1, the parameters that have a positive association with the crime rate are Po1, Ineq, Ed, M, and U2. Meaning that when these increase the crime rate also increases. The only parameter with a negative association is Prob. This means that when Prob increases the crime rate decreases. Prob has the greatest absolute value of the coefficients, but that doesn't mean that it has the highest level of association.

If we normalized all the data first then the coefficients could be compared. The reason it can't be compared like that without normalization is because the values of the variables are significantly different. For example the Prob variable is a probability value between zero and one. The minimum value for all other variables in the model are greater than one. The coefficient has to be large for small valued coefficients in order to have an affect on the dependent variable.