Assignment 03

Regression Diagnostics

This goal of this assignment is to give you experience using regression diagnostics for detecting problematic observations. Turn in a printed version of your responses to each of the questions on this assignment.

In questions that ask you to "use matrix algebra" to solve the problem, you can either show your syntax and output from carrying out the matrix operations, or you can use Equation Editor to input the matrices involved in your calculations.

In addition, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 18 points.

Data Set

The data set you will use to answer the questions in this assignment contains measurements for 18 countries on: income inequality (`inequality`), democratic experience (`turnout`), economic development (`energy`), and socialist party strength (`socialist`). Specifically, the variables in the data set are:

- `country`: Country name
- `inequality`: Ratio of the share of income received by the most wealthy population quintile (richest 20%) to the share received by the poorest 40% of the population; Higher values indicate more income inequality
- `turnout`: Proportion of the adult population voting in the most recent national election prior to 1972
- `energy`: Energy consumption per capita (expressed in million metric tons of coal equivalents; higher values indicate more economic development
- `socialist`: Annual average proportion of seats held by socialist parties in the national legislature, over the first twenty postwar years

In particular, we are going to examine whether income inequality is related to the democratic experience and economic development of a country. We will examine this by regressing income inequality on voter turnout (democratic experience) and energy consumption (economic development).

Exploratory Analysis

1. Start by creating scatterplots to examine the relationship between each of the predictors and the outcome. Are there observations that look problematic in these plots? If so, identify the country(ies).

2. Fit the regression model specified earlier to the data. Report the fitted equation.

3. Create and include a set of plots that allow you to examine the assumptions for linear regression. Based on these plots, comment on the tenability of these assumptions.

Outliers, Leverage, and Influence

4. Compute the studentized residuals for the observations based on the fitted regression. Based on these values, identify any countries that you would consider as regression outliers. Explain why you identified these countries as regression outliers.

5. Fit a mean-shift model that will allow you to test whether the observation with the largest absolute studentized residual is statistically different from zero. Report the coefficient-level output (B, SE, t, and p) for this model.

6. Find (and report) the Bonferroni adjusted p-value for the observation with the largest absolute studentized residual. Based on this p-value, is there statistical evidence to call this observation a regression outlier? Explain.

7. Create and include an index plot of the leverage values. Include a line in this plot that displays the cutpoint for "high" leverage. Based on this plot, identify any countries with large leverage values.

8. Based on the evidence you have looked at in Questions #4–7, do you suspect that any of the countries might influence the regression coefficients? Explain.

Influence Measures

9. For each of the influence measures listed below, create and include an index plot of the influence measure. For each plot, also include a line that displays the cutpoint for "high" influence. **(3pts)**

    a. Scaled (standardized) DFBETA values
    b. Cook's D
    c. DFFITS
    d. COVRATIO

10. Show how the Cook's D value for the country with the largest Cook's D value is calculated using the country's leverage value and studentized residual.

11. Create and include the added-variable plots for each each of the coefficients. Based on these plots, identify any countries that you believe may be jointly influencing the regression coefficients.

Remove and Refit

12. Based on all of the evidence from the different influence measures you examined, identify and report the country(ies) that are influential. Explain how you decided on this set of observations.

13. Remove the observations you identified in Question #12 from the data and refit the regression model omitting these observations. Report the fitted equation.

14. Create and include a set of plots that allow you to examine the assumptions for linear regression. Based on these plots, comment on the tenability of these assumptions.

15. Compare and contrast the coefficient-level inferences from the model fitted with the full data and that fitted with the omitted observations.

16. Compare and contrast the model-level summaries, namely R^2 and the RMSE, from the model fitted with the full data and that fitted with the omitted observations.