Assignment 07

Using WLS to Model Data with Outliers

This goal of this assignment is to give you experience using methods for estimating regression results under violation of homoskedasticity. 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 20 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 socialist party strength of a country. (There is a hypothesis in political science that suggests that more socialist countries have less income inequality.)

Exploratory Analysis

1. Start by creating a scatterplot to examine the relationship between socialist party strength and income inequality (outcome).
2. Are there observations that look problematic in this plot? If so, identify the country(ies).
3. Fit a linear model regressing income inequality on socialist party strength. Examine and report a set of regression diagnostics that allow you to identify any observations that are regression outliers.

Weighted Least Squares Estimation

Rather than removing regression outliers from the data, we can instead fit a model that accounts for these observations. For example, fitting a model that allows for higher variance at x-values that have outliers. With higher variances, we would expect more extreme observations because of the increased variance. The WLS model allows for heteroskedasticity and can be used to model data that have extreme observations.

4. Compute the empirical weights that you will use in the WLS estimation. Report the weight for the United States. (Hint: We do not know the true variances in the population.)
5. Fit the WLS model. Report the fitted equation.
6. Based on the model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?
7. Create a scatterplot that shows the relationship between socialist party strength and income inequality. Include the country names as labels (or instead of the points). Include both the OLS and WLS regression lines on this plot.
8. Based on the plot, comment on how the residuals from the WLS model compare to the residuals from the OLS model.
9. Based on your response to Question #8, how will the model-level R^2 value from the WLS model compare to the model-level R^2 from the OLS model. Explain.
10. The mathematical formulaa for computing the studentized residuals for both the OLS and WLS models is given below. Compute and report the studentized residuals, using this formula, from both the OLS and WLS models for any regression outliers you identified in Question #2. (Hint: Remember that in an OLS regression the weight is 1 for each observation.)

$$e_i' = \frac{e_i}{s_{e(-i)}\sqrt{1-h_{ii}}} \times \sqrt{w_i}$$

11. Based on the values of the studentized residuals in the WLS model, are the observations you identified as regression outliers from the OLS model still regression outliers in the WLS model? Why or why not?
12. Explain why the is the case by referring to the formula.
13. Create and report residual plots of the studentized residuals versus the fitted values for the OLS and WLS models. Comment on which model better fits the assumptions.

Including Covariates

Now include the `energy` covariate into the model to examine the effect of socilist strength after controlling for economic development. Since the model has changed, we need to re-compute the weights and re-carry out the WLS analysis.

14. Use matrix algebra to compute the empirical weights based on the two-predictor model and report the weight for the United States.
15. Fit the two-predictor WLS model using matrix algebra. Report the fitted equation.
16. Compute and report the standard errors of the two-predictor WLS model using matrix algebra.

17. Using your results from Questions #14 and #15, compute and report the t-values and p-values. Show your work or syntax. While you can use the output of the `tidy()`, `summary()`, or other functions that automatically compute p-values to check your work, you can not use them to answer this question. (Hint: Use the `pt()` function.)

18. Based on the two-predictor WLS model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?

19. Based on the two-predictor OLS model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?

20. Which set of the model results should we trust. Explain by referring to the tenability of the assumptions.