PS405: Linear Models

Problem Set 7

Professor: Nicole Wilson TA: Artur Baranov

Due: Thursday, March 6, 2025

Please submit your write-up, including *any* code that you ran, via Canvas. We suggest using Quarto or `RMarkdown`, but any software that renders R code into a PDF document will work. I have provided an `RMarkdown` template if you prefer to use it (note that the process/format is very similar to Quarto, if you are familiar with that). This should be completed *before* class begins at 9:30 AM. Please save your write-up with easy-to-recognize file names (e.g., `ps1_wilson.pdf` or `ps1_baranov.pdf`). Late submissions will not be accepted without special permission from the course instructors. Where instructed to follow a specific coding procedure in `R`, such as the creation of a new function or a data generating process, please include the code in a code chunk in your write up. For example, you can use `echo = TRUE` in the chunk header in Quarto or `RMarkdown` to ensure your code is visible in the output. You do not need to submit the raw file you use to produce your PDF, but the course instructors may request it on a case-by-case basis.

Please post all questions to the Canvas anonymous discussion forum.

Problem 1

In the book, "The Economic Effects of Constitutions," Torsten Persson and Guido Tabellini examine possible institutional determinants of economic performance. A subset of the data they use ("tabellini.dta") is available on Canvas (remember that you will need the `foreign` package to load this). Below is the list of the variables.

- *rgdph* - real GDP per capita in constant dollars (chain index) expressed in international prices, base year 1985.
- *polityIV* - score for democracy, ranging from +10 (strongly democratic) to -10 (strongly autocratic).
- *gini_8090* - Gini index on income distribution, computed as the average of two data points: the observation closest to the 1980 and the observation closest to the 1990. When only one of the two years year is available, only that year is included.
- *trade* - sum of exports and imports of goods and services measured as a share of GDP.

a) First we are going to consider the normality of errors assumption.

i.) Regress GDP per capita on polity score, and check for non-normality of the errors using the studentized residuals. Note: Although `plot(model)` will give you a qqplot, is uses the standardized (but not the studentized) residuals and does not provide a confidence interval. Instead, use `qqPlot()` from the `car` package. If you have forgotten why studentized residuals are preferred, this is a good opportunity to look back at the slides. Although the conclusions may be similar.

ii.) Generate the log of GDP per capita and regress it on polity score. Check for non-normality using the studentized residuals. Do things look better than in (a)(i)? Why or why not?

iii.) Now, regress GDP (not logged) on the polity, gini and trade variables. Again, check for the normality of errors. Which approach do you prefer, transforming the dependent variable or adding predictors, to deal with the non-normality of errors?

b) Now, let's investigate the validity of homoskedasticty assuption.

i. Use your regression from part (a)(iii) (no log, 3 regressors) and produce a spread-location plot. Note: you can use the `which=` argument in the `plot()` function to specify which of the diagnostic plots you want. What do you notice? Do you think the errors are homoskedastic?

ii. Now check your evaluation more formally. Run a Breush-Pagan test on your regression. Report *and interpret* the result.

iii. Rerun the same regression with HC standard errors. You can use any package, but make sure to use one of the small sample corrections. Report the models with and without HC standard errors in a well-formatted, clearly-labeled table. What do you notice when you compare the two models?

c) Now we want to see if we have any influence points we should be concerned about.

i. Evaluate influence for your model in any way you please. For instance, you may look at a Cook's distance threshold (which you can find in the lecture slides), or evaluated using either `plot()` or `influencePlot()` from the `car` package. Which observations stand out as potentially concerning? When are are trying to figure out which country names map onto which observation number, make sure that none of your observations dropped out of your model due to missing data. You can confirm this by making sure that the number of rows in the model matrix `modname$model` is the same as the number of rows in your data. Do you have any guesses as to why these might be unusual observations? Feel free to look at the data to inform your guess.

ii. Since we are primarily interested in the coefficient for democracy, let's look at the DFbetaS for that coefficient instead to identify influence points of interest. If there were other observations you identified as influence points in part (c)(i) that do not show up here, check the DFbetaS for those observations directly, just to make sure they are not very close to the threshold and therefore potentially worth considering.

iii. Why do you think the particular observation(s) identified are so unusual? What would your next steps be to address this?