Problem set 8: Regression

Due November 20, 2023, at 9pm

(Your name here)

Note: Start with the file `ps8_2023_regression.qmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F23/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the `Render` button). Submit both the qmd file and the PDF via Canvas.

The dataset `brexit_data_gb_subset.csv` (available on the github under data) contains results of the 2016 UK Brexit referendum by local authority (cities, counties, etc), collected from the Electoral Commission website and 2011 census data by Claire Peacock.

Question 1

(1a) Having loaded the data, use `lm()` to regress `Percent_Leave` (the support for Brexit in the local authority) on `Region` (the region in which the local authority is located). Present a regression table using `modelsummary::modelsummary()` or a similar approach.

(1b) In the regression you just presented, what does the `(Intercept)` coefficient mean, i.e. what does that number signify?

(1c) Using the regression output above, report and interpret a 95% confidence interval for the coefficient on `RegionLondon`.

(1d) Using the regression above, what is the estimated average support for Brexit in London local authorities?

(1e) For this question you will produce several different estimates for the standard error of the estimated average support for Brexit in London local authorities.

(1e.1) Compute the standard error using the regression above (hint: use variance rule and `vcov()`)

(1e.2) Obtain the standard error from the output of the regression of `Percent_Leave` on `Region` with no intercept (hint: add `-1` to the `formula` argument)

(1e.3) Compute the standard error of the sample mean for London local authorities, i.e. with no regression

(1e.4) Compute the standard error as you did in (1e.1), except this time use `estimatr::lm_robust()` for the regression, which uses the "sandwich estimator" (Huber-White standard errors) and thus does not assume homoskedasticity

(1e.5) What does homoskedasticity mean in this case? Check whether it appears to be valid using the regions of `London`, `East`, and `Scotland` as examples.

Question 2

(2a) Regress `Percent_Leave` on `Bachelors_deg_percent` (the percent of local authority inhabitants who have at least a bachelors degree) and `Birth_UK_percent` (the percent of local authority inhabitants who were born in the UK). Interpret all of the coefficients.

(2b) Use the bootstrap (with $m = 1000$ resamples) to compute standard errors for your coefficients. You should store the m sets of coefficients in a matrix with m rows and 3 columns.

(2c) Compare the standard errors you compute above to the classical standard errors generated by `lm()` and the robust (Huber-White) standard errors generated by `estimatr::lm_robust()`.

(2d) Report the variance-covariance matrix of your bootstrap coefficient estimates. Interpret the bottom left entry of the matrix; what explains its sign? Confirm that it is similar to the variance covariance matrix you obtain by running the regression with `estimatr::lm_robust()`.