Introduction to Econometrics
Session 5 – Linear Regression: robust variance-covariance matrix

November 2025

1 Problem

We continue from the work done in the previous example. Load the dataset `EquationCitations` from the `AER` package. This dataset reports the number of citations received over the five years following publication for evolutionary biology articles published in 1998, as well as the number of equations in each article.

1. Plot a graph with the number of citations received by each article on the y-axis, and the number of pages on the x-axis.
2. For each possible number of pages, estimate the variance of the number of citations among articles of that length.
3. Plot these results on a graph. What can you conclude?
4. Estimate the regression of the number of citations on the number of pages, and store the result in the object `reg_cit_pages`.
5. Create a new column in the `EquationCitations` object corresponding to the residuals from the previous regression.
6. For each possible number of pages, compute (i) the square of the mean residual among articles of that length, and (ii) the mean of the squared residuals among articles of that length. How do these quantities relate to the conditional variance of the number of citations?
7. Plot the mean of the squared residuals against article length. What can you infer?
8. Recreate the object `reg_cit_pages` by adding the option `x = TRUE`. What does this option change?

9. Use the `x` component of `reg_cit_pages` to estimate the empirical counterpart of $\mathbb{E}[XX']$.

10. Create a diagonal matrix `diag_resid_sq`, of the same size as the data, whose diagonal entries correspond to the squared residuals for each article.

11. Using this matrix and the `x` component of `reg_cit_pages`, compute the empirical counterpart of the central term in the asymptotic variance-covariance matrix: $\mathbb{E}[XX'\varepsilon^2]$.

12. Use the results from Questions 9 and 11 to estimate the asymptotic variance-covariance matrix of the coefficient vector.

13. Compare this matrix to the one computed by R using the `vcovHC` function with the option `"HC0"`.

14. Compare this matrix to the one computed by R using the `vcov` function.

15. What standard errors are reported by default by the `coeftest` function in this case?

16. Using the variance-covariance matrix estimated in Question 11, compute heteroskedasticity-robust standard errors with the `coeftest` function.

17. What is the difference between the matrix estimated in Question 11 and the one computed by R with the `vcovHC` function and the option `"HC1"`?