Introduction to Econometrics
Session 5 – Linear Regression: Interpretation of Coefficients

September 2025

1 Problem

Load the dataset `EquationCitations` from the `AER` package. This dataset records the number of citations over the following 5 years for evolutionary biology papers published in 1998, as well as the number of equations in each paper.

1. Use the function `lm` to estimate the simple linear regression of `cites` on `pages`, and store the result in the object `regression_cit_pages`.
2. Create a variable representing, for each article, the average number of citations among papers with the same number of pages.
3. Create a plot showing:
   - the number of citations for each article;
   - the average number of citations for papers of the same length (in pages),

   as a function of the number of pages.
4. Display the regression coefficients from the object `regression_cit_pages`.
5. Compute the mean of the residuals, stored in the object `regression_cit_pages`, and the variable `pages`.
6. Compute the correlation between the residuals (stored in `regression_cit_pages`) and the variable `pages`.
7. Reproduce the plot from question 3 and add the predicted values from the regression.
8. Re-estimate the regression, replacing the dependent variable `cites` with the variable computed in question 2. What can you conclude?

9. Create a new dataset representing the Cartesian product that contains all possible pairs of articles that can be formed from the data.

10. For each pair of articles (i, j), compute the quantity $\texttt{p}_{ij} = \frac{\texttt{cites}_i - \texttt{cites}_j}{\texttt{pages}_i - \texttt{pages}_j}$. What does this quantity represent?

11. Compute the weighted mean of p_{ij} over all pairs (i, j), using weights proportional to the squared difference between the number of pages of the two articles. What does this quantity correspond to?

12. Compute the variance of the residuals (stored in `regression_cit_pages`) and the variance of the predicted values. How do they compare to the variance of the variable `cites`? How is the coefficient of determination constructed?

13. Use the variable `journal` to compute the average article length for each journal.

14. Use the function `lm` to estimate the regression of the number of pages on the variable `journal`. How can the coefficients be interpreted? Store the resulting object in `regression_pages_journal`.

15. Use the function `lm` to estimate the regression of the number of citations on the variable `journal`. How can the coefficients be interpreted? Store the resulting object in `regression_cit_journal`.

16. Use the function `lm` to run the regression of the residuals from `regression_cit_journal` on the residuals from `regression_pages_journal`.

17. Compare the coefficients from this regression to those from the regression of `cite` on `pages` and `journal`. What result does this illustrate?

18. For each journal, separately, estimate the regression of `cites` on `pages`. Compute the number of articles published in each journal, as well as the variance of the number of pages. How can you derive from this the coefficient on `pages` in the regression from question 17? What can you conclude more generally?

19. Compute the coefficient of determination for the regression fom question 17. How does it compare to the one from question 12?