Introduction to Econometrics
Session 8 – Linear Regression: Statistical Tests

November 2025

1 Problem 1: Review of Standard Tests

1. Load the `CPS1985` dataset from the `AER` package.
2. Estimate the regression of hourly wage `wage` on a female indicator and the level of education measured by years spent in the school system (`education`).
3. Construct the 95% confidence intervals for the regression coefficients, using the appropriate variance–covariance matrix.
4. Construct the 99% confidence intervals for the same coefficients.
5. Create two education variables: one specific to men, equal to `education` for men and 0 for women; and another equal to `education` for women and 0 for men.
6. Regress `wage` on the female indicator and these two new education variables.
7. Test the joint nullity of all coefficients except the intercept in this regression (F-test).
8. Use the `linearHypothesis` function from the `car` package to test the equality of the coefficients on these two education variables. How should one interpret the equality hypothesis being tested?
9. Is it possible to test the same hypothesis using a standard single-coefficient test in a regression?

2 Problem 2: Monte Carlo Simulation

1. Create a list containing 1000 independent generations of a table with 500 observations, with:
   - a variable `x` drawn uniformly;

- a normal variable `epsilon`, mean zero, with standard deviation equal to `x`;
- a variable `y1` defined by $\texttt{y1} = 3 + 2 * \texttt{x} + \texttt{epsilon}$;
- a variable `y2` defined by $\texttt{y2} = 4 + \texttt{epsilon}$.

2. For each of these simulations:
   - estimate the coefficients of the regression of `y1` on `x`, and store the estimated coefficient on `x` in the vector `hatbeta1`;
   - estimate the standard error of this coefficient under the homoskedasticity assumption, and store the values in the vector `hatsigma1homo`;
   - estimate the standard error of this coefficient under the heteroskedasticity assumption, and store the values in the vector `hatsigma1hetero`;
   - compute the 95% confidence intervals under each of the two assumptions, for all data generations.
3. Estimate the empirical mean of the values in `hatbeta1`.
4. Estimate the empirical standard deviation of the values in `hatbeta1`.
5. Compare this empirical standard deviation to the theoretical standard deviations under homoskedasticity and heteroskedasticity.
6. Estimate the share of confidence intervals (under homoskedasticity and heteroskedasticity) that contain the value 2.
7. For each of these simulations:
   - estimate the coefficients of the regression of `y2` on `x`, and store the estimated coefficient on `x` in the vector `hatbeta2`;
   - estimate the standard error of this coefficient under the homoskedasticity assumption, and store the values in the vector `hatsigma2homo`;
   - estimate the standard error of this coefficient under the heteroskedasticity assumption, and store the values in the vector `hatsigma2hetero`;
   - test the nullity of the coefficient on `x` under each of the two assumptions at the 5% level;
   - What proportion of simulations reject the null hypothesis in both cases?