Statistics for MFEs – Problem Set 6

Professor Martin Lettau

Due January 24, 2024, 2:00pm PST, to be submitted via bCourses

1. Download the data in `PS6_data.zip` from bCourses. Write a Jupyter notebook with functions specified below. Then apply the functions to the data in `PS6_data.zip`.

   Function `ols`: Compute OLS estimator and related variables

   (a) Input: $y, \mathbf{X}$ plus various options.

   (b) Return the OLS point estimates, the OLS and White standard errors, the R^2 and the value of the log likelihood function at the optimum.

   (c) Check whether a constant is included in $\mathbf{X}$.

   (d) If X does not include a constant, option to include a constant or not.

   (e) Return the projection matrices $\mathbf{P}$ and $\mathbf{M}$ and check whether $\mathbf{PM} = 0$.

   (f) Check whether the sample equivalent of the condition $\mathrm{E}(\mathbf{ex}) = 0$ is satisfied.

   (g) Check whether the sample equivalent of the second moment matrix $\mathbf{Q}_{xx} = \mathrm{E}(\mathbf{x}_i\,\mathbf{x}_i')$ is (close to) singular. The "condition number" of a matrix is a measure of how close a matrix is to being singular (see `http://mathworld.wolfram.com/ConditionNumber.html` and `numpy.linalg.cond`).

   Function `t_test`: Compute t-test whether $\hat{\beta}_i$ equal to b^0.

   (a) Input: The necessary output from `ols` and b^0.

   (b) Return the value of the t-test and flags whether the test rejects at the 1%, 5% and 10% levels.

   The file `PS6_data.zip` includes `Y.csv` and `X.csv`. The dependent variable(s) are in `Y.csv` and the independent variables are in `X.csv`. The regressions should include a constant.

   (a) Use the function `ols` to compute the OLS estimator(s) and print the results (nicely formatted).

   (b) Use the function `t_test` to test whether each individual point estimate is either 0 or 1 at the 1%, 5% and 10% levels.

   Note: This question is intentionally vague about how to exactly map the data into your functions. Use logic and common sense how to do so correctly.

2. Write a Jupyter notebook for the following simulation exercise. Consider the linear model $y_i = \alpha + \beta x_i + e_i$ where $x_i \sim f_x(\theta_x), e_i \sim f_e(\theta_e)$ and f_x, f_e are PDFs of some random variable. x_i and e_i are independent. For this question, assume that f_x and f_e are both standard normal distribution but we will consider different distributions below. Assume that $\alpha = \beta = 1$.

    (a) What are the theoretical properties of the OLS estimators $\widehat{\alpha}$ and $\widehat{\beta}$? What is their theoretical distribution given a sample size N?

    (b) Draw M i.i.d. random samples of length N for x_i and e_i. Start with $M = 5000$ and $N = 50$ but you should experiment with different values.

        For each $m = 1, ..., M$, compute the OLS estimators $\widehat{\alpha}_m$ and $\widehat{\beta}_m$ and their standard errors. Compute the means and standard deviations of the distribution of the M estimators $\widehat{\alpha}_m$ and $\widehat{\beta}_m$.

    (c) Plot the histograms of the M $\widehat{\alpha}_m$ and $\widehat{\beta}_m$.

    (d) Compare the distribution of the simulated $\widehat{\alpha}_m$ and $\widehat{\beta}_m$ to their theoretical distributions.

3. Repeat the exercise in question 1 but assume that x_i and e_i are correlated with a correlation coefficient of $\rho_{xe} = 0.5$.

4. Repeat the exercise in question 1 but assume that f_x, f_e are both t-distributions with 5 degrees of freedom.

5. Repeat the exercise in question 1 but assume that f_x, f_e are both uniform distributions $U[0, 1]$.

6. Experiment with different sample sizes, e.g. set N to 100, 500, ... You do not have to report all results in detail but describe the effect of N on the behavior of the OLS estimators in the questions above.