Tutorial 1 Simulation

Ben Deaner

January 20, 2025

For this tutorial, we will examine how the mean squared out-of-sample prediction error of the OLS estimator changes as the sample size n and dimension d grow. This will also be a helpful introduction to Monte Carlo simulation.

In the simulation we draw data from the following model:

$$\begin{pmatrix} X_i \\ U_i \end{pmatrix} \sim N(0, I)$$

$$Y_i = \beta_0' X_i + U_i$$

The OLS estimator in this case has the formula below:

$$\hat{\beta} = (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

We wish to estimate the mean squared error of the corresponding prediction model for various combinations of sample size and dimension of X_i. Given the model, we can re-write this as:

$$\hat{\beta} = \beta_0 + (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i U_i$$

A very useful practical point that allows us to speed up our simulations, is that the above only depends on the following matrix:

$$\frac{1}{n}\sum_{i=1}^{n} (X_i', U_i)'(X_i', U_i)$$

Rather than draw n separate observations for each simulation (which could make things very slow for large n), we will just draw the above directly from its distribution. Details of this distribution are given at the end of this notebook).

We can use the same trick when calculating the mean squared error on the test sample. Note that the mean squared error can be expanded as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i'\hat{\beta})^2 \\
=&\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} Y_i X_i'\hat{\beta} + \hat{\beta}'\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\hat{\beta}
\end{aligned}$$

Subtracting off the mean squared error from the prediction $X_i'\beta_0$ we get:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta_0)^2\\ =&-2\frac{1}{n}\sum_{i=1}^{n}Y_iX_i'(\hat{\beta}-\beta_0) + \hat{\beta}'\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\hat{\beta} - \beta_0'\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\beta_0\end{aligned}$$

Where $\frac{1}{n}\sum_{i=1}^{n}Y_iX_i' = \beta_0'\frac{1}{n}\sum_{i=1}^{n}X_iX_i' + \frac{1}{n}\sum_{i=1}^{n}U_iX_i'$ which can again be written in terms of blocks of the matrix $\frac{1}{n}\sum_{i=1}^{n}(X_i', U_i)'(X_i', U_i)$.

Finally a little detail on how we draw the matrix $\frac{1}{n}\sum_{i=1}^{n}(X_i', U_i)'(X_i', U_i)$ for the curious. A useful fact is that if an iid random vector Z_i has a zero mean normal distribution with variance-covariance matrix Σ, then $\frac{1}{n}\sum_{i=1}^{n}Z_iZ_i'$ has what is known as a Wishart distribution with n degrees of freedom and 'scale matrix' Σ/n. We write, $\frac{1}{n}\sum_{i=1}^{n}Z_iZ_i' \sim W_p(\Sigma/n, n)$, where p is the dimension of Z_i. Thus to reduce the computational intensity of the simulation, we draw $\frac{1}{n}\sum_{i=1}^{n}(X_i', U_i)'(X_i', U_i)$ from $W_p(I, n)$.