Pranav Madhukar
Mentor: Ronan Perry
Reference Book: Mostly Harmless Econometrics- An Empericist's Companion

Topics Discussed:

During the directed reading program, Ronan and I discussed the ideas and went over the mathematical proofs from the first three chapters of Mostly Harmless Econometrics: An Empiricist's Companion. This started off with a discussion of random variables and our interests in obtaining $E(\mathrm{Y})$ for a population.

At first, we considered the univariate case (i.e. not a conditional expectation) and showed that the mean is a commonly used estimator for the $E(Y)$ because it minimizes the mean squared error. We then extended this proof showing that mean is the best estimator under MSE to the multivariate case $E(Y|X)$ where X represents a vector of all the conditionalities of our analysis $X = X_1, X_2, X_3, \dots, X_i$. At this point we also discussed some properties of the conditional expectation related to decomposition, prediction and the analysis of variance (bias-variance tradeoff).

However, the population mean is usually inaccessible. So, we introduced the central limit theorem as a means to find some relationship between sample and population data. This was the focus of my presentation.

Central Limit Theorem:

$$\frac{\sqrt{n}}{\widehat{\sigma}}(\mu - \widehat{\mu}) \to N(0,1)$$

This is an incredibly powerful result because it allows us to take a sample from any arbitrary data distribution (say some population distribution) and note that it tends to a standard normal distribution as the number of samples increases. During the DRP, we tested the robustness of this theorem and generated plots for n=5, 10, 30, 100, 200 samples from various non-gaussian population distributions (ex. exponential, t-distribution df=1, geometric etc.) and showed that they follow the normal distribution even for quite small n.

For my DRP presentation, we simulated random values of political opinions (which in the real world too does not typically follow any clear distribution) and showed that in the bivariate case (Political Opinion ~ Age) under different distributions of the population, it still converged to a standard normal distribution as per the central limit theorem.

Regression:

From discussing the bivariate case of the central limit theorem, we also introduced the concept of a linear regression.

$$\mathrm{Y} = \beta \mathrm{X} + \varepsilon$$

We also briefly continued our discussion of the bias-variance tradeoff in the context of data fitting, ideal regression coefficients and estimated regression coefficients.