Guide to the Lecture: Lecture 3 - 5

Lecture *3 - 5* provides an in-depth coverage of classical linear regression model (CLM), which is the workhorse of regression models and is also the most used (and mis-used) statistical model. We begin with a one-explanatory-variable CLM to introduce (1) all (except one) of the statistical assumptions underlying CLM, (2) statistical properties of the Ordinary Least Square (OLS) estimators, algebraic properties of OLS statistics, important moment conditions, and important concepts such as fitted values, residuals, goodness-of-fit, R^2, total sum of squares, explained sum of squares, and residual sum of squares. Even with one explanatory variable, practical issues such as units of measurement, functional form, and incorporation of non-linear effects in CLM. It is important to note that classical linear model specifies the conditional expectation as a linear function of parameters, but it allows for nonlinear functions of both dependent and explanatory variables.

When studying statistical models, the underlying statistical assumptions are critically important. When applying a classical linear regression model to data, there is no guarantee that the CLM assumptions are satisfied. Therefore, as you read the texts, watch the video lectures, and attempt a few hands-on exercises, pay special attention to all of the assumptions made when a model is introduced; don't just blindly use the `lm()` `function` in R.

Starting in lecture 3, multiple linear regression models are introduced. Unlike simple linear regression models, multiple linear regression models include more than one explanatory variables, and the inclusion of more than one explanatory variables affects the interpretation of the model coefficients; that is, the "partialling out" interpretation of multiple regression. Although it is much more convenient and concise to use matrix notations when studying multiple linear regression, we stick with the scalar notations. Appendix E of Wooldridge covers matrix presentation of CLM.

Lecture 3 focuses on OLS estimation and introduces several of the key assumptions of CLM. Statistical properties, such as unbiasedness, consistency, efficiency, of the estimators are crucially dependent on these assumptions. Learn to interpreting the assumptions. I have found that many practitioners pay way too much attention to the assumption of no perfect collinearity. In practice, the violation of this assumption is very easy to detect, and most of the software "automatically" drop one of the variables that causes the perfect collinearity. However, we have to pay attention as to "which" variable(s) are being dropped "automatically" by the statistical software we use. In this lecture, we also examine the properties when (some of) these assumptions hold. Pay special attentions to the consequences on these properties when one or more of these assumptions is violated.

Lecture 4 focuses on statistical inference. The concept of minimum variance unbiased estimators is introduced? It is not a bad idea to review the concepts of statistical inference covered in lecture 2. The key difference of the statistical inference in this and lecture 2 is that inference in this lecture is done in the context of linear regression. Many practitioners are unclear about inference concerning one parameter, several parameters, and the whole regression itself. All of these concepts are covered in this lecture. Make sure you study chapter 4 very carefully.

Lecture 5 begins the study of model specification. The corresponding readings are chapter 6 and 7 in Wooldridge. The concepts and techniques covered in this lecture, together with lecture 3 and 4, have a lot of

practical importance in regression model building. Variable transformation (on both the dependent and independent variables) and extra variable creation (the so-called "feature engineering" in machine learning) are also covered extensively in this lecture. Many students in the past confused the materials in these three lectures with those taught in the "Experiment and Causal Inference" course. While there are some overlaps, the focus in regression modeling in these three lectures and the regression introduced in the "Experiment" course is different. Of course, the materials covered in "Mostly Harmless Econometrics", which I strongly recommend and is used in the "Experiment" course, can also be used in this course, but it does not mean that these three lectures and the "Experiment" course is the same.

Main Topics Covered in these Lectures:

- The concept of population model
- Bivariate and Multivariate Ordinary Least Square Regression Models
- The "partialling out" interpretation of multiple regressions
- Method of Moment Estimation
- Statistical Assumptions Underlying the Classical Linear Model
- The Gauss-Markov Theorem
- Diagnostics and responding to assumption violations
- Statistical Inference in the context of CLM
- OLS asymptotic
- Transformation and Interpretation of Dependent and Independent Variables:
    - Logarithmic transformations
    - Quadratics and higher-order polynomials
    - Indicator Variables
    - Interaction terms and their interpretation
- Regression Diagnostics (of the potential violations of OLS regression model assumptions
- Formal Statistical test of OLS regression model assumptions

Learning Outcomes:

After successfully completing this lecture, you shall be able to

- understand and apply the Neyman-Pearson approach to statistical inference (in the context of linear regression model building)
- explain the subjective and objective definitions of probability
- list the axioms of probability
- define concepts such as null and alternative hypotheses, rejection region, size of the test, Type I error, Type II error, and the beta and power of a test
- describe multiple comparisons, post-hoc tests, and planned tests
- understand the meaning of statistical significance
- define objective probability and its relationship (if at all) to hypothesis
- apply Bayes' Theorem
- describe the Bayesian approach to statistical inference