ECON 3740: INTRODUCTION TO ECONOMETRICS

Instructor: Chaoyi Chen
Department of Economics and Finance, University of Guelph

Lecture 11

Lecture outline

Last lecture, we learned the measurement of goodness of fit of the MLR, assumptions of the OLS estimators for the Unbiasedness. Today, we will

- study the variance of the OLS estimators
  - More Assumptions for the MLR Model
  - Sampling Variances of the OLS Slope Estimators
  - The Components of OLS Variances
  - Multicollinearity

MLR: The variance of OLS estimators, Assumptions

- Assumption MLR.5 (Homoskedasticity): $Var(\mu_i|x_{i1}, .., x_{ik}) = \sigma^2$.
  - Intuitively, this means that the value of the explanatory variables must contain no information about the variance of the unobserved factors.
  - A **short** hand notation: $Var(\mu_i|\boldsymbol{x}_i) = \sigma^2$, where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})$, i.e., all explanatory variables are collected in a random vector.
- Example: In the wage equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$$

- The homoskedasticity assumption:

$$Var(\mu_i|educ_i, exper_i, tenure_i) = \sigma^2$$

may also be hard to justify in many cases.

MLR: The variance of OLS estimators, Theorem

- Theorem (Sampling Variances of the OLS Slope Estimators): Under assumptions MLR.1-MLR.5,

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)},\ j = 1, ..., k,$$

where σ^2 is the variance of error term, $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in explanatory variable x_j, and R_j^2 is the R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant), i.e., the R^2 in the regression

$$x_j \sim 1, x_1, .., x_{j-1}, x_{j+1}, ..., x_k.$$

- Note that

$$SST_j(1-R_j^2) = SSR_j = \sum_{i=1}^{n} \widehat{r}_{ij}^2,$$

where $\widehat{r}_{ij}$ is the residual in the regression (1).

- Compare with the SLR case ($Var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$), the MLR case replace $x_i - \bar{x}$ with $\widehat{r}_{ij}$.

MLR: The Components of OLS Variances

- The error variance, σ^2:
  - A high error variance increases the sampling variance because there is more "noise" in the equation.
  - A large error variance necessarily makes estimates imprecise.
  - The error variance does not decrease with sample size.
- The total sample variation in the explanatory variable x_j, SST_j:
  - More sample variation leads to more precise estimates.
  - Total sample variation automatically increases with the sample size. why?
    - For any $n+1$ values $x_i : i = 1, ..., n+1$,

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \leq \sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^2 \leq \sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2$$

    unless $x_{n+1} = \bar{x}_n$
  - Increasing the sample size n is thus a way to get more precise estimates.
  - These two components are similar as in the SLR model.

MLR: The Components of OLS Variances Continue

- The linear relationships among the independent variables, R_j^2:
  - In the regression of x_j on all other independent variables (including a constant), the R^2 will be the higher the better x_j can be linearly explained by the other independent variables.
  - Sampling variance of $\widehat{\beta}_j$ will be the higher the better explanatory variable x_j can be linearly explained by other independent variables.
  - The problem of almost linearly dependent explanatory variables is called multicollinearity (i.e., $R_j^2 \longrightarrow 1$ for some j).
  - If $R_j^2 = 1$, i.e., there is perfect collinearity between x_j and other regressors, then β_j cannot be identified. This is why $Var(\widehat{\beta}_j) = \infty$ now.
  - Multicollinearity plays a similar role as the sample size n in $Var(\widehat{\beta}_j)$: both work to increase $Var(\widehat{\beta}_j)$
  - Like perfect collinarity, multicollinearity is a **small-sample** problem. As larger and larger data sets are available nowadays, i.e., n is much larger than k, it is seldom a problem in current econometric practice.

MLR: An Example for Multicollinearity

- Consider the following MLR model

$$avgscore = \beta_0 + \beta_1 teacherexp + \beta_2 mateexp + \beta_3 otherexp + ...,$$

where

avgscore is average standardized test score of school
teacgerexp is expenditures for teachers
mateexp is expenditures for instructional materials
otherexp is other expenditures

- The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.
- It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.
- As a consequence, sampling variance of the estimated effects will be large.

MLR: Discussion of Multicollinearity

- In the above example, it would probably be better to lump all expenditure categories together because effects cannot be disentangled.
- In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias).
- Only the sampling variance of the variables involved in multicollinearity will be inflated; the estimates of other effects may be very precise.
- Note that multicollinearity is not a violation of MLR.3 in the strict sense.
- Multicollinearity may be detected through "variance inflation factors":

$$VIF_j = \frac{1}{1 - R_j^2}$$

- As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10 (or R_j^2 should not be larger than 0.9).