ECON 3740: INTRODUCTION TO ECONOMETRICS

INSTRUCTOR: CHAOYI CHEN
Department of Economics and Finance, University of Guelph

Lecture 10

Lecture outline

Last lecture, we learned the model, motivation, OLS estimates of the MLR as well as the FWL theorem. Today, we will

- continue multiple linear regression analysis: estimation
  - Goodness-of-fit
  - The Expected Value of the OLS Estimators
    - Standard Assumptions for the MLR Model
    - Unbiasedness of OLS
    - Including Irrelevant Variables in a Regression Model
    - Omitted Variable Bias

MLR: algebraic properties of OLS regression

- $\sum_{i=1}^{n} \widehat{\mu}_i = 0$: deviations from the fitted regression line sum up to zero.
- $\sum_{i=1}^{n} x_{ij} \widehat{\mu}_i = 0$, $j = 1, ..., k$: correlations between deviations and regressors are zero.
- $\bar{y} = \widehat{\beta}_0 + \bar{x}_1 \widehat{\beta}_1 + ... + \bar{x}_k \widehat{\beta}_k$: sample averages of y and of the regressors lie on the fitted regression plane.
- These properties are corollaries of the FOCs for the OLS estimates.

MLR: goodness-of-fit

- Similar to the SLR, we can decompose total variation as

$$SST = SSE + SSR$$

- R-squared:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Alternative expression for R-squared [*Proof not required*]

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})]\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\widehat{y}_i - \bar{\widehat{y}})^2\right)} = \frac{\widehat{Cov}(y, \widehat{y})^2}{\widehat{Var}(y)\widehat{Var}(\widehat{y})} = \widehat{Corr}(y, \widehat{y})^2$$

- R-squared is equal to the squared correlation coefficient between the actual and the predicted value of the dependent variable.
- R-Squared Cannot Decrease When One More Regressor Is Added

MLR: Standard Assumptions for the MLR Model

- Assumption MLR.1 (Linear in Parameters):

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \mu$$

  - In the population, the relationship between y and x is linear.
  - The "linear" in linear regression means "linear in parameter".

- Assumption MLR.2 (Random Sampling): The data $\{(x_{i1}, ..., x_{ik}, y_i) : i = 1, .., n\}$ is a random sample drawn from the population, i.e., each data point follows the population equation,

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \mu_i$$

MLR: Standard Assumptions for the MLR Model continue

- Assumption MLR.3 (No **Perfect** Collinearity): In the sample (and therefore in the population), none of the independent variables is constant and there are no exact relationships among the independent variables.

  - Remark 1: The assumption only rules out perfect collinearity/ correlation between explanatory variables; imperfect correlation [1] is allowed.
  - Remark 2: If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated.
  - Remark 3: Constant variables are also ruled out (This is the case that the independent variable is collinear with intercept).
  - Remark 4: In essence, this is an extension of the assumption $Var(x)$ is positive in the SLR model.

[1] This is referred to as multicollinearity. We will formally discuss it later

MLR: Standard Assumptions for the MLR Model: an example for perfect collinearity

- Recall the average score model,

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + \mu$$

  - Remark 1: we fully expect *expend* and *avginc* to be correlated
  - Remark 2: Assumption MLR.3 only rules out perfect correlation between *expend* and *avginc* in our sample.
  - Remark: However, in a small sample, *avginc* may accidentally be an exact multiple of expend; it will not be possible to disentangle their separate effects because there is exact covariation (Cause we cannot move one while fixing another). As a result, we will have perfect collinearity problem.

- Assumption MLR.4 (Zero Conditional Mean):

$$E[\mu_i|x_{i1}, .., x_{ik}] = 0$$

  - Remark 1: The value of the explanatory variables must contain no information about the mean of the unobserved factors.

  - Remark 2: In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error.

- An Example: Reconsider

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + \mu$$

  - Remark: If *avginc* was not included in the regression, it would end up in the error term; it would then be hard to defend that *expend* is uncorrelated with the error.

MLR: Standard Assumptions for the MLR Model - Discussion of Assumption 4

- Explanatory/independent variables that are **correlated** with the error term are called endogenous; endogeneity is a violation of assumption MLR.4.
- Explanatory variables that are uncorrelated with the error term are called exogenous; MLR.4 holds if all explanatory variables are exogenous.
- Exogeneity is the **key** assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators.

MLR: Unbiasedness of OLS

- Theorem (Unbiasedness of OLS): Under assumptions MLR.1-MLR.4,

$$E[\widehat{\beta}_j] = \beta_j,\ j = 0, 1, ..., k,$$

for every β_j.

- Unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values.

MLR: Including Irrelevant Variables in a Regression Model

- Suppose

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu$$

  where $\beta_3 = 0$ (eg. x_3 is irrelevant to y).

- Remark 1: This will not influence the unbiasedness property: $E[\widehat{\beta}_3] = \beta_3 = 0$

- Remark 2: However, including irrelevant variables may increase sampling variance. (More noise and no information data used to estimate the model)

MLR: Omitted Variable Bias: the Simple Case

- Suppose the **true** model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mu$$

  i.e., the true model contains both x_1 and x_2 ($\beta_1 \neq 0, \;\beta_2 \neq 0$), while the estimated model is

$$y = \alpha_0 + \alpha_1 + e$$

  i.e., x_2 is **omitted**.
- If x_1 and x_2 are correlated, assume a linear regression relationship between them:

$$x_2 = \delta_0 + \delta_1 x_1 + v.$$

- Then

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + \mu \\ &= \beta_0 + \beta_2\delta_0 + (\beta_1 + \beta_2\delta_1)x_1 + (\mu + \beta_2 v). \end{aligned}$$

- If y is only regressed on x_1, the estimated intercept is $\beta_0 + beta_2\delta_0 = \alpha_0$ and the estimated slope is $\beta_1 + \beta_2\delta_1 = \alpha_1$.
- why? The new error term $e = \mu + \beta_2 v$ satisfies the zero conditional mean assumption: $E[\mu + \beta_2 v|x_1] = E[\mu|x_1] + \beta_2 E[v|x_1] = 0$.
- That is, all estimated coefficients will be **biased**.