ECON 3740: INTRODUCTION TO ECONOMETRICS

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

Lecture 9

Lecture outline

Last lecture, we finished topic five. Today, we will

- study topic 6: multiple linear regression analysis: estimation
    - The model and motivation
    - Mechanics and intepretation of OLS
        - The OLS estimates
        - Interpret OLS estimates
        - A "Partialling Out" interpretation of multiple regression
        - FWL theorem

MLR: the model and motivation

- The multiple linear regression (MLR) model is defined as

$$y_i = \beta_0 + \beta_1 x_{1i} + .. + \beta_k x_{ki} + \mu_i$$

which tries to explain variable y in terms of variables $x_1, x_2, ..., x_k$.

- Terminologies for $y, (x_1, x_2, ..., x_k), \mu, \beta_0, (\beta_1, .., \beta_k)$ are the same as in the SLR model.

- Motivations:
  - 1. Incorporate more explanatory factors into the model
  - 2. Explicitly hold fixed other factors that otherwise would be in μ
  - Allow for more flexible functional forms

- Motivation 1 is easy to understand. We will provide four examples to illustrate the other two motivations: Examples *i* and *ii* for motivation 2 and Examples *iii* and *iv* for motivation 3.

MLR: motivation 2 - example *i*

- Example *i*: Suppose we have the augmented education-wage model

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \mu,$$

where

$$wage = \text{hourly wage}$$
$$educ = \text{ years of education}$$
$$exper = \text{ years of labor market experience}$$
$$\mu = \text{all other factors affecting } wage$$

- Now, β_1 measures effect of education EXPLICITLY HOLDING EXPERIENCE FIXED
- If omitting *exper*, then $E[\mu|educ] \neq 0$ given that *educ* and *exper* are correlated. This implies $\hat{\beta}_1$ is biased.

MLR: motivation 2 - example *ii*

- Example *ii*: Suppose, we would like to investigate the effect of per student spending on average score. Therefore, we propose the following model

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

where

avgscore= average standardized test score of school
expend=per student spending at this school
avginc=average family income of students at this school
μ= all other factors affecting *avgscore*

- Per student spending is likely to be **correlated** with average family income at a given high school because of school financing.
- Omitting average family income in regression would lead to **biased** estimate of the effect of spending on average test scores.
- In a simple regression model, effect of per student spending would partly include the effect of family income on test scores. (what is the intuition behind? direct effect and indirect effect)

MLR: motivation 3 - example *iii*

- Example *iii*: Suppose, we would like to investigate the effect of family income on family consumption. Therefore, we propose the following model

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \mu$$

where

$$\begin{aligned} cons &= \text{ family consumption} \\ inc &= \text{ family income} \\ inc^2 &= \text{family income squared} \\ \mu &= \text{all other factors affecting } cons \end{aligned}$$

- Model has two explanatory variables: income and income squared.
- Consumption is explained as a **quadratic** function of income.
- One has to be very careful when interpreting the coefficients:

$$\frac{\partial cons}{\partial inc} = \beta_1 + 2\beta_2 inc$$

which depends on how much income is already there. (Note that $\frac{\partial cons}{\partial inc}$ is the marginal propensity to consume)

MLR: motivation 3 - example *iv*

- Example *iv*: To investigate the effect of CEO Tenure on CEO salary. One can propose:

$$log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + \mu$$

where

$$log(salary) = \text{log of CEO salary}$$
$$log(sales) = \text{ log sales}$$
$$ceoten = \text{CEO tenure with the firm}$$

- Model assumes a **constant** elasticity relationship between CEO salary an the sales of her or his firm. (why?)
- Model assumes a **quadratic** relationship between CEO salary and his or her tenure with the firm.
- Note that the model is still linear model since the "linear" in linear regression means linear in parameter, not "linear in the variables".

MLR: mechanics and interpretation of OLS - obtain OLS estimates

- Suppose we have a random sample $\{(x_{i,1}, ..., x_{ik}, y_i) : i = 1, .., n\}$, where the first subscript of x_{ij}, i, refer to the observation number, and the second subscript j, $j = 1, .., k$, refer to different independent variables.
- Define the residuals at arbitrary $\beta = (\beta_0, \beta_1, .., \beta_k)$ as

$$\hat{\mu}_i(\beta) = y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik}$$

- Minimize the sum of squared residuals:

$$\underset{\beta}{min} SSR(\beta) = \underset{\beta}{min} \sum_{i=1}^{n} \hat{\mu}_i(\beta)^2 = \underset{\beta_0, \beta_1, .., \beta_k}{min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik})^2$$

MLR: mechanics and interpretation of OLS - obtain OLS estimates

- Differentiate the $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik})^2$ w.r.t $\beta_0, \beta_1, .., \beta_k$, we have the FOCs

$$
\begin{aligned}
\sum_{i}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik}) &= 0 \\
\sum_{i}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik}) &= 0 \\
&\ldots \\
\sum_{i}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - .. - \beta_k x_{ik}) &= 0
\end{aligned}
$$

Therefore, we have $k+1$ equations and $k+1$ unknown variables $(\beta_0, \beta_1, .., \beta_k)$ to solve. One can calculate the OLS estomates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ through R.

MLR: mechanics and interpretation of OLS - interpret OLS estimates

- In the MLR model, $y_i = \beta_0 + \beta_1 x_{1i} + .. + \beta_k x_{ki} + \mu_i$. Hence

$$\beta_j = \frac{\partial y}{\partial x_j}.$$

This helps us analyze the *ceteris paribus* effect with the meaning - "by how much does the dependent variable change if the j^{th} independent variable is increased by one unit, holding all other independent variables and the error term **constant**".

- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration. The multiple linear regression (MLR) model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration.
- It has still to be assumed that unobserved factors do not change if the explanatory variables are changed.

MLR: mechanics and interpretation of OLS - interpret OLS estimates, an example

- Example: Determinants of College GPA. After the estimation, the fitted regression is:

$$\widehat{colGPA}_i = 1.29 + 0.453 hsGPA + 0.0094 ACT$$

$colGPA$=grade point average at college
$hsGPA$= high school grade point average
ACT= achievement test score [1]

- Holding ACT fixed, another point on high school GPA is associated with another 0.453 points college GPA
- Or: If we compare two students with the **same** ACT, but the $hsGPA$ of student A is one point higher, we predict student A to have a $colGPA$ that is 0.453 higher than that of student B.
- Holding high school GPA fixed, another 10 points on ACT are associated with less than one-tenth point on college GPA.

[1] Examples of achievement test are SAT, AP, and the national university entrance exam in some countries

MLR: mechanics and interpretation of OLS - a "Partialling Out" interpretation of multiple regression

- One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:
  - Step 1. Regress the explanatory variable on *all other explanatory variables.*
  - Step 2. Regress y on the residuals from this regression
- Mathematically, suppose we regress y on the constant 1, x_1, and x_2 (denoted as $y \sim 1, x_1, x_2$), and want to get $\hat{\beta}_1$. We implement following regressions
  - $x_{i1} \sim 1, x_{i2} \Longrightarrow \widehat{r}_{i1}$ [2]
  - $y_i \sim \widehat{r}_{i1} \Longrightarrow$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \widehat{r}_{i1} y_i}{\sum_{i=1}^n \widehat{r}_{i1}^2}$$

- In Step 1, the constant regressor is not required since the mean of $\widehat{r}_{i1}$ is equal to zeros from Step 1. If the constant regressor is added in, the formula of $\hat{\beta}_1$ is the same since $\bar{\widehat{r}}_1 = \frac{1}{n}\sum_{i=1}^n \widehat{r}_{i1} = 0$

[2] we use $\widehat{r}$ to denote the corresponding residuals for each regression

- Recall, from the SLR, step 1 gives: $x_{i1} = \widehat{x}_{i1} + \widehat{r}_{i1}$ with $\widehat{x}_{i1} = \widehat{\delta}_0 + \widehat{\delta}_1 x_{i2}$, $\sum_{i=1}^n \widehat{r}_{i1} = 0$, $\sum_{i=1}^n x_{i2}\widehat{r}_{i1} = 0$ and $\sum_{i=1}^n \widehat{x}_{i1}\widehat{r}_{i1} = 0$.
- Recall that the FOCs of OLS when $k = 2$ are

$$\begin{gathered}\sum_i^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{i2}) = 0 \\ \sum_i^n x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{i2}) = 0 \\ \sum_i^n x_{i2}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{i2}) = 0\end{gathered}$$

- From the second FOC,

$$\begin{gathered}\sum_i^n (\widehat{\delta}_0 + \widehat{\delta}_1 x_{i2} + \widehat{r}_{i1})(y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{i2}) \\ = \widehat{\delta}_0 \sum_{i=1}^n \widehat{\mu}_i + \widehat{\delta}_1 \sum_{i=1}^n x_{i2}\widehat{\mu}_i + \sum_{i=1}^n \widehat{r}_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_k x_{i2}) \\ = -\widehat{\beta}_0 \sum_{i=1}^n \widehat{r}_{i1} - \widehat{\beta}_2 \sum_{i=1}^n x_{i2}\widehat{r}_{i1} + \sum_{i=1}^n \widehat{r}_i \Big(y_i - \widehat{\beta}_1 (\widehat{x}_{i1} + \widehat{r}_{i1}) \Big)\end{gathered}$$

where $\widehat{\mu}_{i1} = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2}$, the second iequality is from the first and third FOCs, and the third equality is from the properties of $\widehat{r}_{i1}$ above.

- Solving the last equality, $\sum_{i=1}^n \widehat{r}_{i1} y_i = \widehat{\beta}_1 \sum_{i=1}^n \widehat{r}_{i1}^2$, we can obtain the $\widehat{\beta}_1$ formula.

MLR: FWL theorem and the connection with SLR

- Why does this procedure work? This procedure is usually called the FWL theorem and was proposed in the following two papers:
  - Frisch, R. And F. Waugh, 1933, Partial Time Regressions as Compared with Individual Trends, *Econometrica*, 1, 387-401.
  - Lovell, M.C., 1963, Seasonal Adjustment of Economic Time Series, *Journal of the American Statistical Association*, 58, 993-1010.
- The residuals from the first regression is the part of the explanatory variable that is **uncorrelated** with the other explanatory variables.
- The slope coefficient of the second regression therefore represents the **isolated** (or pure) effect of the explanatory variable on the dependent variable.
- Recall that in the SLR,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- So in the MLR, we replace $x_i - \bar{x}$ by $\widehat{r}_{i1}$. Intuitively, $x_i - \bar{x}$ us the residual in the regression of x_i on all other explanatory variables [3],