

Multiple Regression Analysis (Estimation)

1 Motivation

The SLR4. assumption, $E(u_i|X_i) = 0$, is unrealistic. This implies that that u_i is uncorrelated with X_i because no matter what the value of X_i is, the expected value of u_i would still be 0! Thus, when this assumption does not hold, OLS estimates (β_0 and β_1 will be biased). (Why? – note: $Cov(X_i, \hat{u}_i) = 0$ is always true by the OLS calculation. This does not mean that $Cov(X_i, u_i) = 0$ in reality.)

In this case, the multiple regression analysis is introduced in order to achieve the condition $E(u_i|X_i) = 0$. It also enables us to explain the dependent variable better and to conduct the "ceteris paribus" or "holding all other things constant" analysis.

Example: If we want to find the relation between wage and education in the simple linear regression, would our β_0 and β_1 be biased? Probably!

- Consider a simple linear regression

1.1 Assumption SLR 4 ($E(u|X) = 0$) in the Multiple Regression Context

Consider the *wage* equation.

- In the case of simple regression, the assumption SLR4 ($E(u|X) = 0$) has to be satisfied in order to achieve an unbiased estimator of β_0 and β_1 .
 - In this two-variable regression of assumption SLR4 ($E(u|X) = 0$) becomes $E(u|X_1, X_2) = 0$.
 - Therefore, the OLS estimator of β_0, β_1 and β_2 would be unbiased if $E(u|educ, inc) = 0$.
 - For example, "innate ability" is not included in eq. ???. Thus, if "innate ability" can explain *wage*, it would be in u .
 - If it is true that $E(\text{innate ability}|educ, inc) = 0$, or the expected value of *innate ability* is the same and (equal to zero) for all education and income levels, then the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ would be unbiased. (Is this likely?)
-

2 The Model with k Independent Variables

- The "population" version of the multiple linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u,$$

where

β_0 is the intercept.

β_1 is the parameter associated with X_1 .

β_2 is the parameter associated with X_2 , and so on.

u is the error term

- Y, X_1, X_2, \dots, X_k are variables
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are parameters

TABLE 2.1

Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

2.1 Accounting Nonlinearity

- The regression model requires linearity in parameters.
- Similar to the Simple Regression Model, the Multiple Regression Model can also take into account the nonlinear relationships between variables.

Example: In the CEO Salary example, we could write the relations between CEO salary (*salary*), firm sales (*sales*) and CEO age (*age*) as follows:

$$\log(\textit{salary}) = \beta_0 + \beta_1 \log(\textit{sales}) + \beta_2 \textit{age} + \beta_3 \textit{age}^2 + u$$

- - This model has $k = 3$ because there are 3 regressors. $Y = \log(\textit{salary})$, $X_1 = \log(\textit{sales})$, $X_2 = \textit{age}$, $X_3 = \textit{age}^2$.
 - β_1 measures the change in $\log(\textit{salary})$ with respect to $\log(\textit{sales})$, holding other factors fixed. β_1 is the sales elasticity of CEO salary.
 - How do we measure the change in $\log(\textit{salary})$ with respect to age, holding other factors fixed?
 - How do we measure the change in *salary* with respect to age, holding other factors fixed?
- In any case, the OLS estimates of β would be unbiased if

$$E(u|X_1, X_2, \dots, X_k) = 0.$$

This is the Multiple Regression version of assumption SLR 4—all factors in the unobserved error term should be uncorrelated with the explanatory variables.

3 Estimation of parameters and properties of estimators

3.1 Deriving OLS Estimators

- We begin with a multiple regression with 2 regressors. Regressions with more regressors can be analyzed in the exact same fashion. Let the population regression model be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where the estimated OLS equation (sample version) of the above regression can be written as

- As before, the OLS estimators are the ones that minimize the sum of residual squared given the observations $i = 1, 2, \dots, n$ in the sample.

$$\arg \min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n (\hat{u}_i)^2 = \arg \min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

First Order Condition (F.O.C):

$$w.r.t. \hat{\beta}_0 \Rightarrow 0 = -2(\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})) \quad (4.1)$$

$$w.r.t. \hat{\beta}_1 \Rightarrow 0 = \quad (4.2)$$

$$w.r.t. \hat{\beta}_2 \Rightarrow 0 = \quad (4.3)$$

- Solving equations 4.1, 4.2 and 4.3 simultaneously, we can derive the solution for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ (the calculation is beyond the scope of this course).

3.2 Algebraic Properties of OLS estimators

1.

2.

3.

4.

3.3 How could the multiple regression analysis enable *ceteris paribus* analysis?

- Consider a multiple regression function of *wage*

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + u \quad (4.4)$$

Here,

- β_0 is the intercept.
 β_1 measures the change in *wage* with respect to *educ*, holding other factors fixed.
 β_2 measures the change in *wage* with respect to *inc*, holding other factors fixed.
- What if the function of *wage* is, instead written as

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + \beta_3 educ^2 + u$$

Then,

β_0 is the intercept.

The change in *wage* with respect to *educ* (holding other factors fixed) is measured by:

The change in *wage* with respect to *inc* (holding other factors fixed) is measured by:

4 Expected Value of the OLS Estimators

- Under assumptions MLR 1 to 4 (see Wooldridge), $\hat{\beta}_{OLS}$ are unbiased.
- 2 issues should be considered regarding the biasedness of $\hat{\beta}_{OLS}$

4.1 Issue #1: Including Irrelevant Variable (Overspecifying the Model)

- Suppose we specify the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (4.5)$$

and this model satisfies the multiple regression assumptions 1 to 4

4.2 *Issue #2: Excluding Relevant Variable (Underspecifying the Model → omitted variable bias. This is a serious problem!)*

- Suppose we the **TRUE** model is actually

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where none of the β is zero **and this model satisfies the multiple regression assumptions 1 to 4.**

- But we omit variable X_2 and estimate the following equation using OLS

5 Variance of the OLS Estimators

- The $\hat{\beta}_{OLS}$ would be the most efficient among the linear unbiased estimators if assumption 5 is satisfied
- Multiple Linear Regression (MLR) assumption 5: Homoskedasticity

The error term u has the same variance given any values of the explanatory variables.

$$Var(u|X_1, X_2, \dots, X_k) = \sigma^2$$

- Example:

- If the MLR assumption 5 is true, then

6 Estimator of the OLS Variance

- Since we don't know what σ^2 is (population concept), we need to find an estimator of it.

- Thus, STATA's calculation of the std.err. of $\hat{\beta}_j$ is

$$\widehat{std.deviation}.\hat{\beta}_j = std.err.\hat{\beta}_j = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^N (X_{ij} - \bar{X}_j)^2 (1 - R_j^2)}}.$$

Comments: