Lecture-4: Multiple Linear Regression-Estimation

In Today's Class

- Recap
- Simple regression model estimation
- Gauss-Markov Theorem
- Hand calculation of regression estimates

Multiple Regression Analysis (MLR)

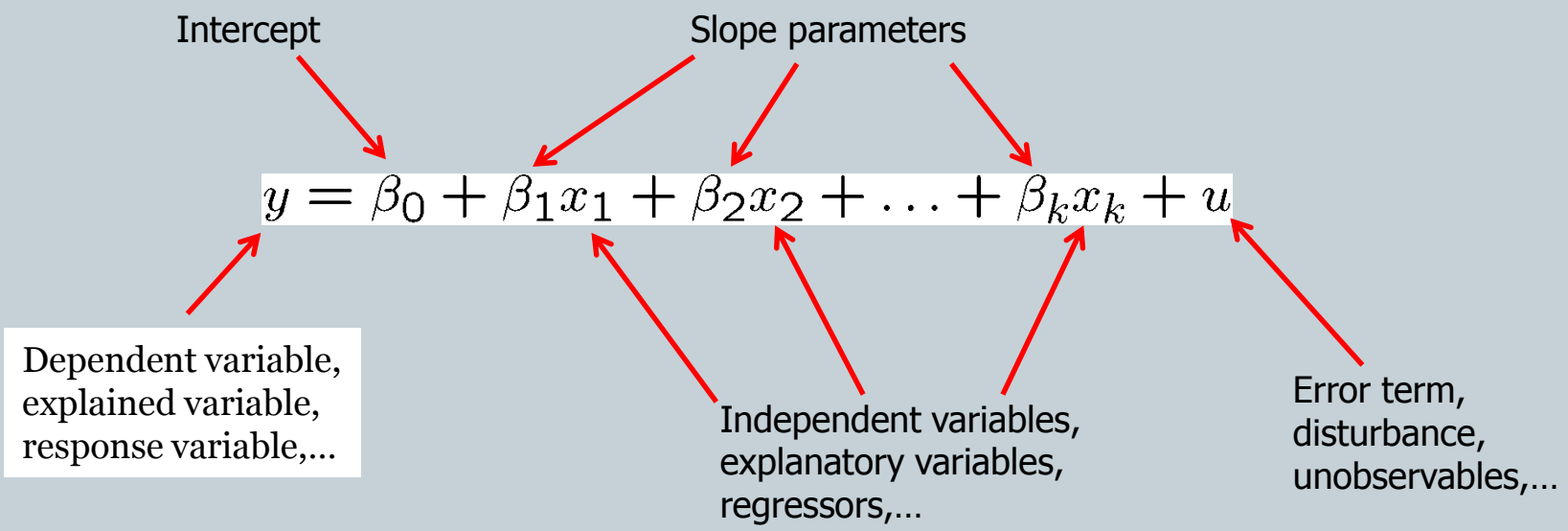
- Allows us to explicitly control for many factors those simultaneously affect the dependent variable
- This is important for
  - examining theories
  - assessing various policies of independent variables
- MLR can can accommodate many independent variables that may be correlated with the dependent variable we can infer causality.
  - In such instances simple regression analysis may be misleading or underestimate the model strength

MLR Motivation

- Incorporate more explanatory factors into the model
- Explicitly hold fixed other factors that otherwise would be in u
- Allow for more flexible functional forms
- Can take as many as independent variables

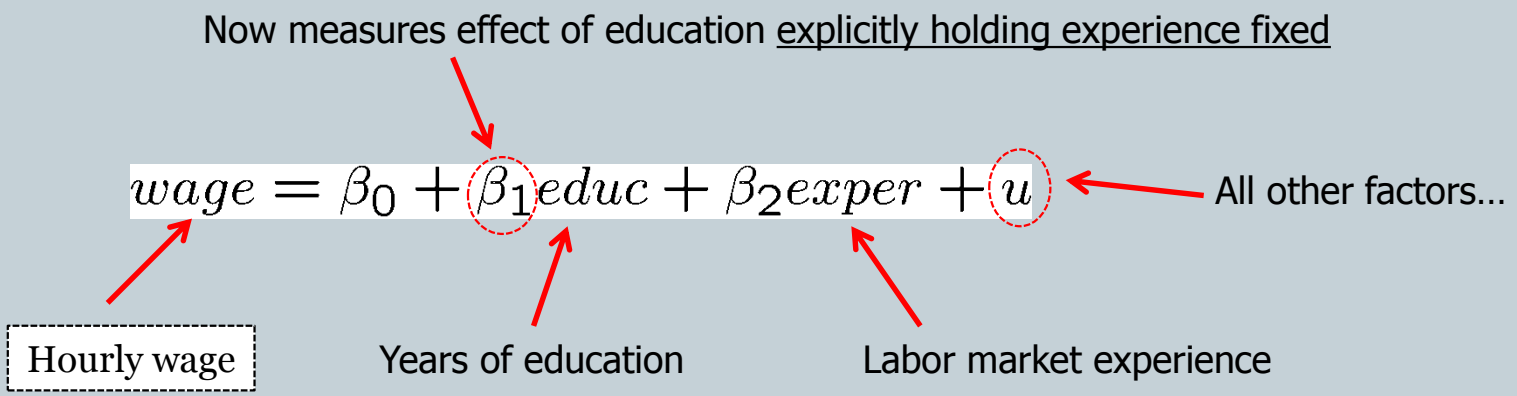
MLR Notation

- Explains "y" in terms of $x_1, x_2, \ldots, x_k$



MLR Example-1

- Wage equation



MLR Example-2

- Average test score, student spending, and income

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

Other factors

Average standardized test score of school

Per student spending at this school

Average family income of students at this school

MLR Example-3

- **Example: Family income and family consumption**

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

Family consumption

Family income

Family income squared

Other factors

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

By how much does consumption increase if income is increased by one unit?

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

Depends on how much income is already there

MLR Example-4

- **Example: CEO salary, sales and CEO tenure**

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

Log of CEO salary

Log sales

Quadratic function of CEO tenure with firm

  - Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm
  - Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm
- **Meaning of linear regression**
  - The model has to be linear in the parameters (not in the variables)

Parallels with Simple Regression

- β_0 is still the intercept
- β_1 to β_k all called slope parameters
- u is still the error term (or disturbance)
- Still need to make a zero conditional mean assumption, so now assume that
- $E(u|x_1,x_2, \ldots,x_k) = 0$
- Still minimizing the sum of squared residuals, so have k+1 first order conditions

Interpreting Multiple Regression (1)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k, \text{ so}$$

$$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1 + \Delta\hat{\beta}_2 x_2 + \ldots + \Delta\hat{\beta}_k x_k,$$

so holding $x_2, \ldots, x_k$ fixed implies that

$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1$, that is each β has

a *ceteris paribus* interpretation

Interpreting Multiple Regression *(2)*

- **Interpretation of the multiple regression model**

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

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
  - Ceteris paribus-interpretation
  - It has still to be assumed that unobserved factors do not change if the explanatory variables are changed

MLR Estimation *(1)*

OLS Estimation of the multiple regression model

- **Random sample**

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \; i = 1, \ldots n\}$$

- **Regression residuals**

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}$$

- **Minimize sum of squared residuals**

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

Minimization will be carried out by computer

MLR Estimation *(2)*

- Estimates can be derived from the first order conditions
  - **Properties of OLS on any sample of data**
  - **Fitted values and residuals**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

Fitted or predicted values

$$\hat{u}_i = y_i - \hat{y}_i$$

Residuals

  - **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

Deviations from regression line sum up to zero

$$\sum_{i=1}^{n} x_{ij} \hat{u}_i = 0$$

Correlations between deviations and regressors are zero

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_k \bar{x}_k$$

Sample averages of y and of the regressors lie on regression line

Goodness-of-fit (1)

We can think of each observation as being made up of an explained part, and an unexplained part,

$y_i = \hat{y}_i + \hat{u}_i$ We then define the following :

$\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

Then SST = SSE + SSR

Goodness-of-fit (2)

- How do we think about how well our sample regression line fits our sample data?
- Can compute the fraction of the total sum of squares (SST) that is explained by the model, call this the R-squared of regression
- $R^2 = SSE/SST = 1 - SSR/SST$

Goodness-of-Fit (3)

We can also think of R^2 as being equal to the squared correlation coefficient between the actual y_i and the values $\hat{y}_i$

$$R^2 = \frac{\left(\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum (y_i - \bar{y})^2\right)\left(\sum (\hat{y}_i - \bar{\hat{y}})^2\right)}$$

More about *R*-squared

- R^2 can never decrease when another independent variable is added to a regression, and usually will increase

- Because R^2 will usually increase with the number of independent variables, it is not a good way to compare models

Assumptions on MLR (1)

- **Standard assumptions for the multiple regression model**
- **Assumption MLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

In the population, the relationship between y and the explanatory variables is linear

- **Assumption MLR.2 (Random sampling)**

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \; i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

Each data point therefore follows the population equation

Assumptions on MLR (2)

- **Standard assumptions for the multiple regression model (cont.)**
- **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"

- **Remarks on MLR.3**
  - The assumption only rules out perfect collinearity/correlation between explanatory variables; imperfect correlation is allowed
  - If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated
  - Constant variables are also ruled out (collinear with intercept)

Assumptions on MLR (3)

- **<u>Standard assumptions for the multiple regression model (cont.)</u>**
- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

The value of the explanatory variables must contain no information about the mean of the unobserved factors

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

Assumptions on MLR (4)

- **Discussion of the zero mean conditional assumption**
  - Explanatory 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 explanat. var. are exogenous
  - Exogeneity is the key assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators
- **Theorem 3.1 (Unbiasedness of OLS)**

$$MLR.1-MLR.4 \quad \Rightarrow \quad E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

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

MLR Unbiasedness

- Population model is linear in parameters: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- We can use a random sample of size n, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i=1, 2, \ldots, n\}$, from the population model, so that the sample model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$
- $E(u|x_1, x_2, \ldots x_k) = 0$, implying that all of the explanatory variables are exogenous
- None of the x's is constant, and there are no exact linear relationships among them

A "Partialling Out" Interpretation

Consider the case where $k = 2$, i.e.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, then

$\hat{\beta}_1 = \left(\sum \hat{r}_{i1} y_i\right) / \sum \hat{r}_{i1}^2$, where $\hat{r}_{i1}$ are the residuals from the estimated regression $\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 \hat{x}_2$

"Partialling Out" continued

- Previous equation implies that regressing y on x_1 and x_2 gives same effect of x_1 as regressing y on residuals from a regression of x_1 on x_2
- This means only the part of x_{i1} that is uncorrelated with x_{i2} is being related to y_i so we're estimating the effect of x_1 on y after x_2 has been "partialled out"

Simple vs Multiple Reg Estimate

Compare the simple regression $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

with the multiple regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Generally, $\tilde{\beta}_1 \neq \hat{\beta}_1$ unless:

$\hat{\beta}_2 = 0$ (i.e. no partial effect of x_2) OR

x_1 and x_2 are uncorrelated in the sample

Too Many or Too Few Variables

- What happens if we include variables in our specification that don't belong?
- There is no effect on our parameter estimate, and OLS remains unbiased

- What if we exclude a variable from our specification that does belong?
- OLS will usually be biased

Omitted Variable Bias

Suppose the true model is given as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, but we estimate $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$, then

$$\tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2}$$

Omitted Variable Bias (cont)

Recall the true model, so that

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, so the numerator becomes

$$\sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) =$$

$$\beta_1 \sum (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i$$

Omitted Variable Bias (cont)

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

since $\mathrm{E}(u_i) = 0$, taking expectations we have

$$E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

Omitted Variable Bias (cont)

Consider the regression of x_2 on x_1

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \text{ then } \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

$$\text{so } E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

Summary of Direction Bias

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Omitted Variable Bias Summary

- Two cases where bias is equal to zero
  - $\beta_2 = 0$, that is x_2 doesn't really belong in model
  - x_1 and x_2 are uncorrelated in the sample

- If correlation between x_2 , x_1 and x_2 , y is the same direction, bias will be positive
- If correlation between x_2 , x_1 and x_2 , y is the opposite direction, bias will be negative

The More General Case

- Technically, can only sign the bias for the more general case if all of the included *x*'s are uncorrelated

- Typically, then, we work through the bias assuming the *x*'s are uncorrelated, as a useful guide even if this assumption is not strictly true

Variance of the OLS Estimators

- Now we know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so
- Assume $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$ (Homoskedasticity)

Variance of OLS (cont)

- Let $\boldsymbol{x}$ stand for $(x_1, x_2, \ldots x_k)$
- Assuming that $\text{Var}(u|\boldsymbol{x}) = \sigma^2$ also implies that $\text{Var}(y|\boldsymbol{x}) = \sigma^2$

- The 4 assumptions for unbiasedness, plus this homoskedasticity assumption are known as the Gauss-Markov assumptions

Variance of OLS (cont)

Given the Gauss-Markov Assumptions

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}, \text{ where}$$

$$SST_j = \sum\left(x_{ij} - \bar{x}_j\right)^2 \text{ and } R_j^2 \text{ is the } R^2$$

from regressing x_j on all other x's

Components of OLS Variances

- The error variance: a larger σ^2 implies a larger variance for the OLS estimators
- The total sample variation: a larger SST_j implies a smaller variance for the estimators
- Linear relationships among the independent variables: a larger R_j^2 implies a larger variance for the estimators

Misspecified Models

Consider again the misspecified model

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1, \text{ so that } Var\left(\tilde{\beta}_1\right) = \frac{\sigma^2}{SST_1}$$

Thus, $Var\left(\tilde{\beta}_1\right) < Var\left(\hat{\beta}_1\right)$ unless x_1 and x_2 are uncorrelated, then they're the same

Misspecified Models (cont)

- While the variance of the estimator is smaller for the misspecified model, unless $\beta_2 = 0$ the misspecified model is biased

- As the sample size grows, the variance of each estimator shrinks to zero, making the variance difference less important

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, because we don't observe the errors, u_i
- What we observe are the residuals, $\hat{u}_i$
- We can use the residuals to form an estimate of the error variance

Error Variance Estimate (cont)

$$\hat{\sigma}^2 = \left(\sum \hat{u}_i^2\right) / (n-k-1) \equiv SSR/df$$

$$\text{thus, } se\left(\hat{\beta}_j\right) = \hat{\sigma} / \left[SST_j\left(1-R_j^2\right)\right]^{1/2}$$

- $df = n - (k + 1)$, or $df = n - k - 1$
- df (i.e. degrees of freedom) is the (number of observations) – (number of estimated parameters)

The Gauss-Markov Theorem

- Given our 5 Gauss-Markov Assumptions it can be shown that OLS is "BLUE"
- Best
- Linear
- Unbiased
- Estimator
- Thus, if the assumptions hold, use OLS