Multiple Linear Regression

CIVL 7012/8012

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 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, …, x_k

Intercept

Slope parameters

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

Dependent variable, explained variable, response variable,…

Independent variables, explanatory variables, regressors,…

Error term, disturbance, unobservables,…

MLR Example-1

- Wage equation

Now measures effect of education explicitly holding experience fixed

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

All other factors…

Hourly wage

Years of education

Labor market experience

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 → $cons$
Family income → inc
Family income squared → inc^2
Other factors → u

  - 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* interpreta tion

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

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

- Minimize sum of squared residuals

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

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

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

Fitted or predicted values

Residuals

  - Algebraic properties of OLS regression

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} x_{ij}\hat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \ldots + \widehat{\beta}_k \bar{x}_k$$

Deviations from regression line sum up to zero

Correlations between deviations and regressors are

Sample averages of y and of the regressors lie on regression li...

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 $\text{SST} = \text{SSE} + \text{SSR}$

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
- $R^2 = SSE/SST = 1 – SSR/SST$

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)

- <u>Standard assumptions for the multiple regression model</u>
- **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)**

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

  - 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

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

Including /Omitting Irrelevant Variables

- **Including irrelevant variables in a regression model**

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

= 0 in the population

No problem because $E(\widehat{\beta}_3) = \beta_3 = 0$.

However, including irrevelant variables may increase sampling variance.

- **Omitting relevant variables: the simple case**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$ True model (contains x_1 and x_2)

$$y = \alpha_0 + \alpha_1 x_1 + w$$ Estimated model (x_2 is omitted)

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

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
  - β_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

Goodness-of-fit: Adjusted R-square (1)

- **More on goodness-of-fit and selection of regressors**
- **General remarks on R-squared**
  - A high R-squared does not imply that there is a causal interpretation
  - A low R-squared does not preclude precise estimation of partial effects
- **Adjusted R-squared**
  - What is the ordinary R-squared supposed to measure?

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)}$$

Goodness-of-fit: Adjusted R-square (2)

- **Adjusted R-squared (cont.)**
  - A better estimate taking into account degrees of freedom would be

$$\bar{R}^2 = 1 - \frac{(SSR/(n-k-1))}{(SST/(n-1))} = adjusted\ R^2$$

Correct degrees of freedom of nominator and denominator

  - The adjusted R-squared imposes a penalty for adding new regressors
  - The adjusted R-squared increases if, and only if, the t-statistic of a newly added regressor is greater than one in absolute value
- **Relationship between R-squared and adjusted R-squared**

$$\bar{R}^2 = 1 - (1 - R^2)(n-1)/(n-k-1)$$

The adjusted R-squared may even get negative

Goodness-of-fit: Adjusted R-square (3)

- **Using adjusted R-squared to choose between nonnested models**
  - Models are nonnested if neither model is a special case of the other

$$rdintens = \beta_0 + \beta_1 \log(sales) + u \quad \leftarrow \quad R^2 = .061, \bar{R}^2 = .030$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u \quad \leftarrow \quad R^2 = .148, \bar{R}^2 = .090$$

  - A comparison between the R-squared of both models would be unfair to the first model because the first model contains fewer parameters
  - In the given example, even after adjusting for the difference in degrees of freedom, the quadratic model is preferred

Incorporating Non-linearities in SLR

- Incorporating nonlinearities: Semi-logarithmic form
- **Regression of log wages on years of eduction**

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Natural logarithm of wage

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{1}{wage} \cdot \frac{\partial wage}{\partial educ} = \frac{\frac{\partial wage}{wage}}{\partial educ}$$

Percentage change of wage

... if years of education are increased by one year

Incorporating Non-linearities in SLR

- **Fitted regression**

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

The wage increases by 8.3 % for every additional year of education (= return to education)

For example:

$$\frac{\frac{\partial wage}{wage}}{\partial educ} = \frac{\frac{+0.83\$}{10\$}}{+1\ \text{year}} = 0.083 = +8.3\%$$

Growth rate of wage is 8.3 % per year of education

Incorporating Non-linearities in SLR

- Incorporating nonlinearities: Log-logarithmic form
- CEO salary and firm sales

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary — Natural logarithm of his/her firm's sales

- This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\partial \log(salary)}{\partial \log(sales)} = \frac{\frac{\partial salary}{salary}}{\frac{\partial sales}{sales}}$$

Percentage change of salary

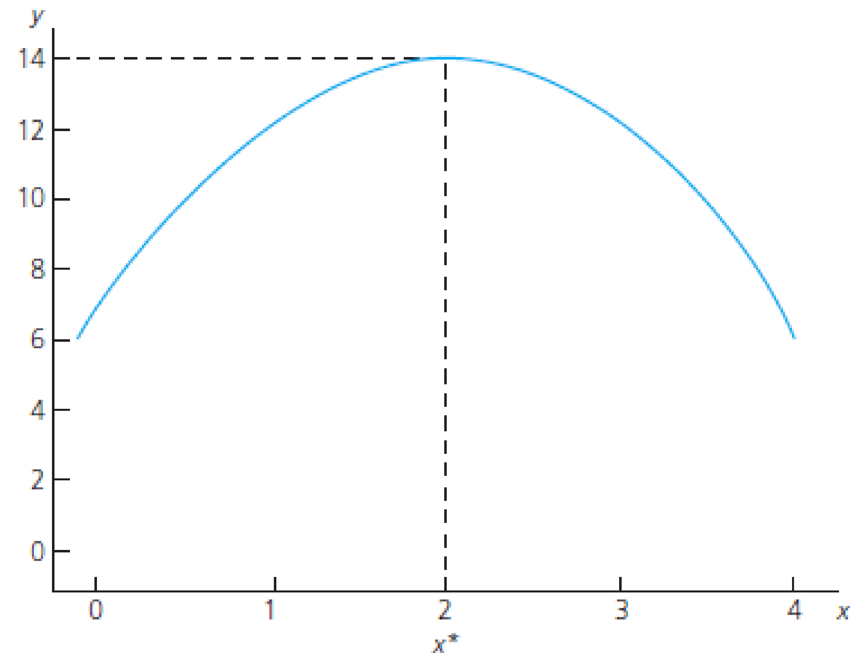
... if sales increase by 1 %

Logarithmic changes are always percentage changes

Introducing Quadratic Forms

- $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- When beta1 >0; and beta2<0, then
- $x^* = \frac{\beta_1}{-2\beta_2}$

FIGURE A.3 Graph of $y = 6 + 8x - 2x^2$.



Source: Woolridge, Introductory Econometrics: A Modern Approach