Econometrics

	Course Calendar
Week	Main Content
Week 7	Extension of Simple Regression: Functional Forms I
Week 8	Extension of Simple Regression: Functional Forms II
Week 9	Extension of Simple Regression: Functional Forms III
Week 10	Multiple Regression
Week 11	Multiple Regression: Problem of Inference
Week 12	Multiple Regression: Functional Forms
Week 13	Introduction to Dummy Variables
Week 14	Introduction to Dummy Variables and Regression Methods
Week 15	Regression with Dummy Variables: Hands-on-Exercise
Week 16	Application of Regression

Econometrics

Lecture 13. Multiple Regression Analysis: Inference (continuation) & Functional Forms; Introduction to Dummy Variables

Geetha Rani Prakasam, Ph.D

Professor,

Recap

- Multiple Linear Regression: Testing of Hypothesis
- Steps involved
- The “Incremental” or “Marginal” contribution of an explanatory variable: using the example of Child Mortality as the regressand
- Testing the equality of two regression coefficients **Examples of cubic cost function**
- **Restricted least square – Cobb-Douglas production function**
- Comparing two regressions:
- Testing the stability of the estimated regression model over time – using the savings and income data of US as an example

Outline

- **Testing of Functional Forms**
- **Functional Forms in Multiple Regression**
  - **The Cobb–Douglas Production Function**
  - **Polynomial Regression Models**
- **Introduction to Dummy Variables**

8.11.Testing the functional form of regression:

- Choosing between Linear and Log–linear Regression Models
- The choice between a linear regression model (the regressand is a linear function of the regressors) or a log–linear regression model (the log of the regressand is a function of the logs of the regressors) is a perennial question in empirical analysis.
- Choosing between linear and log-linear regression (the log of the regressand is a function of the logs of the regressors) models: MWD Test *(MacKinnon, White and Davidson)*

8.11.Testing the functional form of regression

- To illustrate this test, assume the following
- **H_0: Linear Model** Y is a linear function of regressors, the X_s;
- **H_1: Log-linear Model** Y is a linear function of logs of regressors, the $\ln X_s$;
- where, as usual, H0 and H1 denote the null and alternative hypotheses.

8-11. Testing the functional form of regression:

- Step 1: Estimate the linear model and obtain the estimated Y values. Call them Yf (i.e.,Y^). Take lnYf.
- Step 2: Estimate the log-linear model and obtain the estimated lnY values, call them lnf (i.e., ln^Y)
- Step 3: Obtain Z_1 = (lnYf – lnf)
- Step 4: Regress Y on X_s and Z_1. Reject H_0 if the coefficient of Z_1 is statistically significant, by the usual t - test
- **Step 5: Obtain Z_2 = antilog of (lnf – Yf)**
- **Step 6: Regress lnY on $\ln X_s$ and Z_2. Reject H_1 if the coefficient of Z_2 is statistically significant, by the usual t-test**

8.11.Testing the functional form of regression

- Although the MWD test seems involved, the logic of the test is quite simple.
- If the linear model is in fact the correct model, the constructed variable Z_1 should not be statistically significant in Step IV, for in that case the estimated Y values from the linear model and those estimated from the log–linear model (after taking their antilog values for comparative purposes) should not be different.
- The same comment applies to the alternative hypothesis H_1.

EX. 8.5 The Demand for Roses

Quarter	Y	X2	X3	X4	X5
1971–III	11,484	2.26	3.49	158.11	1
–IV	9,348	2.54	2.85	173.36	2
1972–I	8,429	3.07	4.06	165.26	3
–II	10,079	2.91	3.64	172.92	4
–III	9,240	2.73	3.21	178.46	5
–IV	8,862	2.77	3.66	198.62	6
1973–I	6,216	3.59	3.76	186.28	7
–II	8,253	3.23	3.49	188.98	8
–III	8,038	2.6	3.13	180.49	9
–IV	7,476	2.89	3.2	183.33	10
1974–I	5,911	3.77	3.65	181.87	11
–II	7,950	3.64	3.6	185	12
–III	6,134	2.82	2.94	184	13
–IV	5,868	2.96	3.12	188.2	14
1975–I	3,160	4.24	3.58	175.67	15
–II	5,872	3.69	3.53	188	16

EX. 8.5 The Demand for Roses

- Exercise 7.16 - data on the demand for roses in the Detroit metropolitan area for the period 1971–III to 1975–II.
- For illustrative purposes, we will consider the demand for roses as a function only of the prices of roses and carnations, leaving out the income variable for the time being.
- Now we consider the following models:
- Linear model: $Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + u_t$ ***(8.10.1)***
- Log–linear model:

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2\tau} + \beta_3 \ln X_{3\tau} + u_\tau \qquad (8.10.2)$$

EX. 8.5 The Demand for Roses

- where Y is the quantity of roses in dozens,
- X_2 is the average wholesale price of roses (\$/dozen), and
- X_3 is the average wholesale price of carnations (\$/dozen).
- A priori, α_2 and β_2 are expected to be negative (why?), and α_3 and β_3 are expected to be positive (why?).
- Slope coefficients in the log–linear model are elasticity coefficients.

EX. 8.5 The Demand for Roses

- The regression results are as follows
- $\hat{Y}_\tau = 9734.2176 - 3782.1956X_{2\tau} + 2815.2515X_{3\tau}$
  t = (3.3705) (−6.6069) (2.9712) **(8.10.3)**
  F = 21.84 $R^2 = 0.77096$
- $\ln Y_\tau = 9.2278 - 1.7607 \ln X_{2\tau} + 1.3398 \ln X_{3\tau}$
  t = (16.2349) (−5.9044) (2.5407) **(8.10.4)**
  F = 17.50 $R^2 = 0.7292$
- As these results show, both the linear and the log–linear models seem to fit the data reasonably well:
- **The parameters have the expected signs and the *t and R^2 values are statistically* significant.**

EX. 8.5 The Demand for Roses

- To decide between these models on the basis of the **MWD test, we first test the hypothesis** that the true model is linear.
- Then, following Step IV of the test, we obtain the following regression:

$$\hat{Y}_t = 9727.5685 - 3783.0623X_{2t} + 2817.7157X_{3t} + 85.2319Z_{1t}$$

$$t = \quad (3.2178) \quad (-6.3337) \quad (2.8366) \quad (0.0207) \qquad \textbf{(8.11.5)}$$

$$F = 13.44 \qquad R^2 = 0.7707$$

- Since the coefficient of Z_1 is not statistically significant (the p value of the estimated t is 0.98), we do not reject the hypothesis that the true model is linear.

EX. 8.5 The Demand for Roses

- Suppose we switch gears and assume that the true model is log–linear.
- Following step VI of the MWD test, we obtain the following regression results:

$$\widehat{\ln Y_t} = 9.1486 - 1.9699 \ln X_t + 1.5891 \ln X_{2t} - 0.0013 Z_{2t}$$

$$t = (17.0825) \quad (-6.4189) \quad (3.0728) \quad (-1.6612) \qquad \textbf{(8.11.6)}$$

$$F = 14.17 \qquad R^2 = 0.7798$$

- The coefficient of Z_2 is statistically significant at about the 12 percent level (p value is 0.1225).

EX. 8.5 The Demand for Roses

- Therefore, we can reject the hypothesis that the true model is log–linear at this level of significance.
- Of course, if one sticks to the conventional 1 or 5 percent significance levels, then one cannot reject the hypothesis that the true model is log–linear.
- As this example shows, it is quite possible that in a given situation we cannot reject either of the specifications.

EX. 8.5 The Demand for Roses

- $\hat{Y}_\tau = 9727.5685 - 3783.0623X_{2\tau} + 2817.7157X_{3\tau} + 85.2319Z_{1\tau}$
  t = (3.2178) (−6.3337) (2.8366) (0.0207) **(8.10.5)**
  $F = 13.44$ $R^2 = 0.7707$
- Since the coefficient of Z_1 is not statistically *significant (the p value of the estimated t is* 0.98), we do not reject the hypothesis that the true model is linear.
- Suppose we switch gears and assume that the true model is log–linear. Following step VI of the MWD test, we obtain the following regression results:
- $\ln \hat{Y}_\tau = 9.1486 - 1.9699 \ln X_\tau + 1.5891 \ln X_{2\tau} - 0.0013Z_{2\tau}$
  t = (17.0825) (−6.4189) (3.0728) (−1.6612) **(8.10.6)**
  $F = 14.17$ $R^2 = 0.7798$

EX. 8.5 The Demand for Roses

- The coefficient of Z_2 *is statistically significant at about the 12 percent level (p value is* 0.1225). Therefore, we can reject the hypothesis that the true model is log–linear at this level of significance.
- Of course, if one sticks to the conventional 1 or 5 percent significance levels, then one cannot reject the hypothesis that the true model is log–linear.
- As this example shows, it is quite possible that in a given situation we cannot reject either of the specifications.

Functional Forms: The Cobb–Douglas Production Function

- Appropriate transformations can convert nonlinear relationships into linear ones so that we can work within the framework of the classical linear regression model.
- The various transformations discussed there in the context of the two-variable case can be easily extended to multiple regression models.
- We demonstrate transformations in this section by taking up the multivariable extension of the two variable log–linear model; others can be found in the exercises and in the illustrative examples discussed throughout the rest of this book.

The Cobb–Douglas production function:

- The specific example we discuss is the celebrated Cobb–Douglas production function of production theory.
- The Cobb–Douglas prod. function, in its stochastic form, may be expressed as
- $\mathbf{Y_i = \beta_1 X^{\beta 2}_{2i} X^{\beta 3}_{3i} e^{U}_{i}}$ **(7.9.1)**
- where Y = output
- X_2 = labor input; X_3 = capital input
- u = stochastic disturbance term; e = base of natural logarithm

The Cobb–Douglas production function

- From Eq. (7.9.1) it is clear that the relationship between output and the two inputs is nonlinear.
- However, if we log-transform this model, we obtain:

By log-transform of this model:

$$\ln Y_i = \ln\beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + U_i$$

$$= \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + U_i \qquad \textbf{(7.9.2)}$$

- where $\beta_0 = \ln \beta_1$.
- Thus written, the model is linear in the parameters β_0, β_2, and β_3 and is therefore a linear regression model.

The Cobb–Douglas production function:

- Notice, though, it is nonlinear in the variables Y and X but linear in the logs of these variables.
- In short, (7.9.2) is a log-log, double-log, or log-linear model, the multiple regression counterpart of the two-variable log-linear model (6.5.3).
- The properties of the Cobb–Douglas production function are quite well known:
- 1. β_2 is the (partial) elasticity of output with respect to the labor input, that is, it measures the percentage change in output for, say, a 1 percent change in the labor input, holding the capital input constant (see exercise 7.9).

Properties of the Cobb–Douglas production function

- 2. Likewise, β_3 is the (partial) elasticity of output with respect to the capital input, holding the labor input constant.
- 3. The sum ($\beta_2 + \beta_3$) gives information about the returns to scale, that is, the response of output to a proportionate change in the inputs.
- If this sum is 1, then there are constant returns to scale, that is, doubling the inputs will double the output, tripling the inputs will triple the output, and so on.
- If the sum is less than 1, there are decreasing returns to scale—doubling the inputs will less than double the output. Finally, if the sum is greater than 1, there are increasing returns to scale—doubling the inputs will more than double the output.

Properties of the Cobb–Douglas production function

- Before proceeding further, note that whenever we have a log–linear regression model involving any number of variables the coefficient of each of the X variables measures the (partial) elasticity of the dependent variable Y with respect to that variable. Thus, if we have a k-variable log-linear model:
- $\ln Yi = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \cdots + \beta_k \ln X_{ki} + u_i$ (7.9.3)
- each of the (partial) regression coefficients, β_2 through β_k, is the (partial) elasticity of Y with respect to variables X_2 through X_k.

Properties of the Cobb–Douglas production function

- To see this, differentiate (7.9.3) partially with respect to the log of each X variable.
- Therefore, $\partial \ln Y/\partial \ln X_2 = (\partial Y/\partial X_2)(X_2/Y) = \beta_2$,
- which, by definition, is the elasticity of Y with respect to X_2 and
- $\partial \ln Y/\partial \ln X_3 = (\partial Y/\partial X_3)(X_3/Y) = \beta_3$,
- which is the elasticity of Y with respect to X_3, and so on.
- To illustrate the Cobb–Douglas production function, we obtained the data shown in Table 7.3; these data are for the agricultural sector of Taiwan for 1958–1972.

TABLE 7.3 REAL GROSS PRODUCT, LABOR DAYS, AND REAL CAPITAL INPUT IN THE AGRICULTURAL SECTOR OF TAIWAN, 1958–1972

Year	Real gross product (millions of NT $)*, Y	Labor days (millions of days), X2	Real capital input (millions of NT $), X3
1958	16,607.70	275.5	17,803.70
1959	17,511.30	274.4	18,096.80
1960	20,171.20	269.7	18,271.80
1961	20,932.90	267.0	19,167.30
1962	20,406.00	267.8	19,647.60
1963	20,831.60	275.0	20,803.50
1964	24,806.30	283.0	22,076.60
1965	26,465.80	300.7	23,445.20
1966	27,403.00	307.5	24,939.00
1967	28,628.70	303.7	26,713.70
1968	29,904.50	304.7	29,957.80
1969	27,508.20	298.6	31,585.90
1970	29,035.50	295.5	33,474.50
1971	29,281.50	299.0	34,821.80
1972	31,535.80	288.1	41,794.30

Source: Thomas Pei-Fan Chen, "Economic Growth and Structural Change in Taiwan—1952–1972, A Production Function Approach," unpublished Ph.D. thesis, Dept. of Economics, Graduate Center, City University of New York, June 1976, Table II. *New Taiwan dollars.

Basic Econometrics, Damodar Gujarati, Page, 225

Example 7.3: The Cobb-Douglas Production function

- Assuming that the model (7.9.2) satisfies the assumptions of the classical linear regression model, we obtain the following regression by the OLS method.

$$
\begin{aligned}
\widehat{\ln Y_i} &= -3.3384 + 1.4988 \ln X_{2i} + 0.4899 \ln X_{3i} \\
&\quad\; (2.4495) \quad (0.5398) \qquad\quad (0.1020) \\
t &= (-1.3629) \quad (2.7765) \qquad\quad (4.8005) \\
&\qquad R^2 = 0.8890 \qquad \text{df} = 12 \\
&\qquad \bar{R}^2 = 0.8705
\end{aligned}
\tag{7.9.4}
$$

- From Eq. (7.9.4) we see that in the Taiwanese agricultural sector for the period 1958–1972 the output elasticities of labor and capital were 1.4988 and 0.4899, respectively.

Example 7.3: The Cobb-Douglas Production function

- In other words, over the period of study, holding the capital input constant, a 1 percent increase in the labor input led on the average to about a 1.5 percent increase in the output.
- Similarly, holding the labor input constant, a 1 percent increase in the capital input led on the average to about a 0.5 percent increase in the output.
- Adding the two output elasticities, we obtain 1.9887, which gives the value of the returns to scale parameter.
- As is evident, over the period of the study, the Taiwanese agricultural sector was characterized by increasing returns to scale.

Example 7.3: The Cobb-Douglas Production function

- From a purely statistical viewpoint, the estimated regression line fits the data quite well.
- The R^2 value of 0.8890 means that about 89 percent of the variation in the (log of) output is explained by the (logs of) labor and capital.
- Note: the estimated standard errors can be used to test hypotheses about the "true" values of the parameters of the Cobb– Douglas production function for the Taiwanese economy.

7.10 Polynomial Regression Models

- We now consider a class of multiple regression models, the polynomial regression models, that have found extensive use in econometric research relating to cost and production functions.
- In introducing these models, we further extend the range of models to which the classical linear regression model can easily be applied.
- To fix the ideas, consider Figure 7.1, which relates the short-run marginal cost (MC) of production (Y) of a commodity to the level of its output (X).
- The visually-drawn MC curve in the figure, the textbook U-shaped curve, shows that the relationship between MC and output is nonlinear.

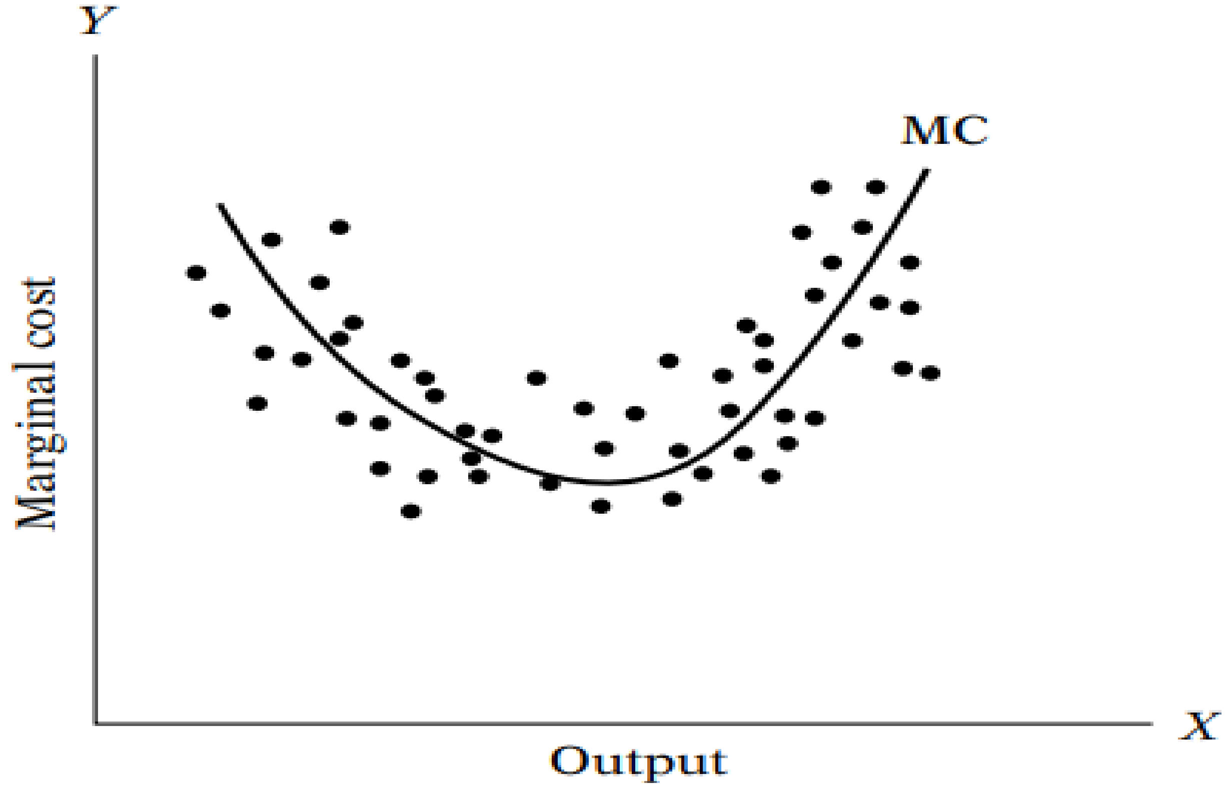
7.10 Polynomial Regression Models

- If we were to quantify this relationship from the given scatter points, how would we go about it?
- In other words, what type of econometric model would capture first the declining and then the increasing nature of marginal cost?
- Geometrically, the MC curve depicted in Figure 7.1 represents a parabola. Mathematically, the parabola is represented by the following equation:
- $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \ldots + \beta_k X^k_i + U_i \qquad (7.10.1)$$

7.10 Polynomial Regression Models

- $\mathbf{Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + U_i}$
- is called a quadratic function, or more generally, a second-degree polynomial in the variable X—the highest power of X represents the degree of the polynomial (if X_3 were added to the preceding function, it would be a third-degree polynomial, and so on).

FIG 7.1 The U-shaped marginal cost curve



Source: Basic Econometrics, Damodar Gujarati, Page, 226

7.10. Polynomial Regression Models

- The stochastic version of (7.10.1) may be written as
- **$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + U_i$ (7.10.2)**
- which is called a second-degree polynomial regression. The general kth degree polynomial regression may be written as
- **$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \ldots + \beta_k X^k_i + U_i$ (7.10.3)**
- Notice that in these types of polynomial regressions there is only one explanatory variable on the right-hand side but it appears with various powers, thus making them multiple regression models.
- Incidentally, note that if X_i is assumed to be fixed or nonstochastic, the powered terms of X_i also become fixed or nonstochastic.

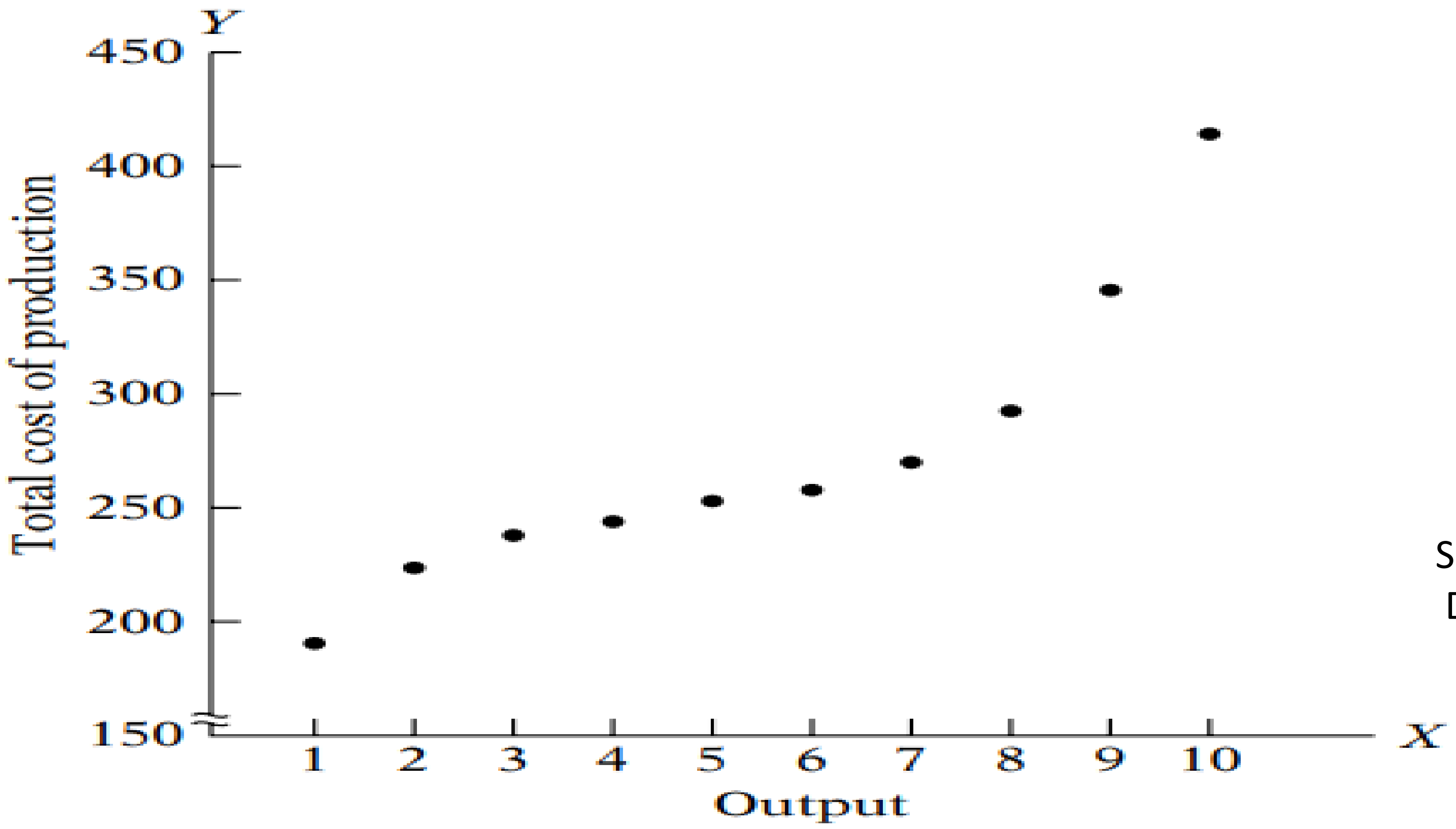
7.10. Polynomial Regression Models

- Do these models present any special estimation problems?
- Since the second-degree polynomial (7.10.2) or the kth degree polynomial (7.10.13) is linear in the parameters, the β's, they can be estimated by the usual OLS or ML methodology.
- But what about the collinearity problem?
- Aren't the various X's highly correlated since they are all powers of X?
- Yes, but remember that terms like X_2, X_3, X_4, etc., are all nonlinear functions of X and hence, strictly speaking, do not violate the no multicollinearity assumption.
- In short, polynomial regression models can be estimated by the techniques presented in this lecture and present no new estimation problems.

Ex. 7.4 Estimating the Total Cost Function

- As an example of the polynomial regression, consider the data on output and total cost of production of a commodity in the short run given in Table 7.4.
- What type of regression model will fit these data?
- For this purpose, let us first draw the scattergram, which is shown in Figure 7.2.

FIG 7.2: The total cost curve



Source: Basic Econometrics,
Damodar Gujarati, Page. 228

Ex. 7.4 Estimating the Total Cost Function

- From this figure it is clear that the relationship between total cost and output resembles the elongated S curve; notice how the total cost curve first increases gradually and then rapidly, as predicted by the celebrated law of diminishing returns.
- This S shape of the total cost curve can be captured by the following cubic or third-degree polynomial:
- $\mathbf{Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \beta_3 X^3_i + U_i}$ **(7.10.4)**
- where Y = total cost and X = output

TABLE 7.4 TOTAL COST (Y) AND OUTPUT (X)

Output	Total cost, $
1	193
2	226
3	240
4	244
5	257
6	260
7	274
8	297
9	350
10	420

Source: Basic Econometrics, Damodar Gujarati, Page. 227

Ex. 7.4 Estimating the Total Cost Function

- Given the data of Table 7.4, we can apply the OLS method to estimate the parameters of (7.10.4).
- But before we do that, let us find out what economic theory has to say about the short-run cubic cost function (7.10.4).
- Elementary price theory shows that in the short run the marginal cost (MC) and average cost (AC) curves of production are typically U-shaped—initially, as output increases both MC and AC decline, but after a certain level of output they both turn upward, again the consequence of the law of diminishing return. This can be seen in Figure 7.3 (see also Figure 7.1).

Ex. 7.4 Estimating the Total Cost Function

- And since the MC and AC curves are derived from the total cost curve, the U-shaped nature of these curves puts some restrictions on the parameters of the total cost curve (7.10.4).
- As a matter of fact, it can be shown that the parameters of (7.10.4) must satisfy the following restrictions if one is to observe the typical U-shaped short-run marginal and average cost curves:

1. β_0, β_1, and $\beta_3 > 0$

2. $\beta_2 < 0$ **(7.10.5)**

3. $\beta_2^2 < 3\beta_1\beta_3$

Ex. 7.4 Estimating the Total Cost Function

- All this theoretical discussion might seem a bit tedious.
- But this knowledge is extremely useful when we examine the empirical results, for if the empirical results do not agree with prior expectations, then, assuming we have not committed a specification error (i.e., chosen the wrong model), we will have to modify our theory or look for a new theory and start the empirical enquiry all over again.
- But as noted in the Introduction, this is the nature of any empirical investigation.

Empirical Results

- When the third-degree polynomial regression was fitted to the data of Table 7.4, we obtained the following results:

$$\hat{Y}_i = \underset{(6.3753)}{141.7667} + \underset{(4.7786)}{63.4776X_i} - \underset{(0.9857)}{12.9615X_i^2} + \underset{(0.0591)}{0.9396X_i^3}$$

$$R^2 = 0.9983 \qquad \textbf{(7.10.6)}$$

- (Note: The figures in parentheses are the estimated standard errors.)

Example 7.5 GDP Growth Rate, 1960–1985 & Relative Per Capita GDP, In 119 Developing Countries

- As an additional economic example of the polynomial regression model, consider the following regression results:

$$\widehat{GDPG}_i = 0.013 + 0.062\,RGDP - 0.061\,RGDP^2$$
$$se = (0.004) \quad (0.027) \quad (0.033) \qquad \textbf{(7.10.7)}$$
$$R^2 = 0.053 \qquad adj\,R^2 = 0.036$$

- per capita GDP, 1960 (percentage of U.S. GDP per capita, 1960).

Ex. 7.5 GDP Growth Rate, 1960–1985 & Relative Per Capita GDP, In 119 Developing Countries

- The R^2 (adj R^2) tells us that, after taking into account the number of regressors, the model explains only about 3.6 percent of the variation in GDPG.
- Even the unadjusted R^2 of 0.053 seems low.
- This might sound a disappointing value but, as we shall show in the next lecture, such low R^2's are frequently encountered in cross-sectional data with a large number of observations.
- Besides, even an apparently low R^2 value can be statistically significant (i.e., different from zero).
- As this regression shows, GDPG in developing countries increased as RGDP increased, but at a decreasing rate; that is, developing economies were not catching up with advanced economies.

Ex. 7.5 GDP Growth Rate, 1960–1985 & Relative Per Capita GDP, In 119 Developing Countries

- If we take the derivative of (7.10.7), we will obtain
- dGDPG / dRGDP = 0.062 − 0.122 RGDP
- showing that the rate of change of GDPG with respect to RGDP is declining. If we set this derivative to zero, we will get RGDP ≈ 0.5082.
- Thus, if a country’s GDP reaches about 51 percent of the U.S. GDP, the rate of growth of GDPG will crawl to zero.
- This example shows how relatively simple econometric models can be used to shed light on important economic phenomena.

Summary and Conclusion

- Testing of Hypothesis: Choosing between Linear and Log–linear Regression Models
- Different functional forms
- Total cost cubic function
- Cobb-Douglas Production Function

Introduction to Dummy Variables

- Dummy variables are qualitative variables
- Whether the student will get admission into an university or not ?
- Whether it is going to rain today or not ?
- Whether it is going to be sunny day or not ?
- Before that a short introduction to what is data and data analysis

Reference

Basic Econometrics by Damodar Gujarati, **Chapter 8:MULTIPLE REGRESSION ANALYSIS: The Problem of Inference**

Basic Econometrics by Damodar Gujarati, **Chapter 7:MULTIPLE REGRESSION ANALYSIS: The Problem of Estimation**

Basics of Production Function and Cobb_Douglas Production Function

- A short video from the following sources are taken on the
- **Basics of Production Function and**
- **https://www.youtube.com/watch?v=_fM2TjqArc4**
- **Cobb_Douglas Production Function**
- **https://www.youtube.com/watch?v=cBCFKiaMKkw**

What Next?

- Dummy Variables before that
- What is data and data analysis?