Multiple Regression Analysis: The Problem of Estimation

Multiple Regression Analysis: The Problem of Estimation

- The Three-Variable Model
- Interpretation of Multiple regression
- The Meaning of Partial Regression Coefficients
- OLS Estimation of the Partial Regression Coefficients
- The Multiple Coefficient of Determinant and the Multiple Coefficient of Correlation
- R-Squared and the Adjusted R-Squared
- Partial Correlation Coefficients

The Three-Variable Model

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

The coefficients β_2 and β_3 are called the **partial regression coefficients**

Assumptions

- Linear regression model, or linear in the parameters
- Fixed X values or X values independent of the error term. We require zero covariance between u_i and each X variables

$$\text{cov}(u_i, X_{2i}) = \text{cov}(u_i, X_{3i}) = 0$$

- Zero mean value of disturbance u_i

$$E(u_i \mid X_{2i}, X_{3i}) = 0, \; for \; each \; i$$

- Homoscedasticity or constant variance of u_i

$$\text{var}(u_i) = \sigma^2$$

- No autocorrelation, or serial correlation, between the disturbances

$$\text{cov}(u_i, u_j) = 0, \quad i \neq j$$

- The numbers of observations n must be greater than the number of parameters to be estimated
- There must be variation in the values of the X variables
- No exact collinearity between the X variables
- There is no specification bias

Interpretation of Multiple Regression Equation

$$E(Y_i \mid X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

The conditional mean or expected value of Y conditional upon the given or fixed values of X_2 *and* X_3

The Meaning of Partial Regression Coefficients

The coefficients β_2 and β_3 are called the **partial regression coefficients**

The meaning of partial regression coefficient is as follows

β_2 measures the change in the mean value of, Y, E(Y), per unit change in X_2, holding the value of X_3 constant.

β_3 measures the change in the mean value of, Y, E(Y), per unit change in X_3, holding the value of X_2 constant.

OLS Estimation of the Partial Regression Coefficients

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

The OLS procedure consists of choosing the values of the unknown parameters so that the residual sum of squares (RSS) $\sum \hat{u}_i^2$ is as small as possible

$$\min \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i})^2$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Variance and Standard Errors of OLS Estimators

In all these formulas variance σ^2 is the homoscedastic) variance of the population disturbances

An unbiased estimator of u_i for the three variable model is given by σ^2

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$$

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 + 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i} \right)^2} \right] \cdot \sigma^2$$

$$se(\hat{\beta}_1) = +\sqrt{\text{var}(\hat{\beta}_1)}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}, \qquad r_{23}^2 = \frac{(\sum x_{2i} x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2}$$

$$se(\hat{\beta}_2) = +\sqrt{\text{var}(\hat{\beta}_2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

$$se(\hat{\beta}_3) = +\sqrt{\text{var}(\hat{\beta}_3)}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

Properties of OLS Estimators

- The three-variable regression line passed through the means. Thus in the k-variable linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

$$\hat{\beta}_1 = \bar{Y} - \beta_2 \bar{X}_2 - \beta_3 \bar{X}_3 - \ldots - \beta_k \bar{X}_k$$

- The mean value of the estimated $Y_i = \hat{Y}_i$ is equal to the mean value of the actual Y_i
- $\sum \hat{u}_i = \bar{\hat{u}} = 0$
- The residuals $\hat{u}_i$ are uncorrelated with X_{2i} and X_{3i}, that is $\sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = 0$
- The residuals $\hat{u}_i$ are uncorrelated with $\hat{Y}_i$; that is $\sum \hat{u}_i \hat{Y}_i = 0$

- From the variance equation, it is evident that as r_{23}, the correlation coefficient between X_2 and X_3, increases towards 1, the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$ increase for given values of σ^2 and $\sum x_{2i}^2$ or $\sum x_{3i}^2$ In the limit, when $r_{23} = 1$ (i.e., perfect collinearity), these variances become infinite

- For given values of r_{23} and $\sum x_{2i}^2$ or $\sum x_{3i}^2$, the variances of the OLS estimators are directly proportional to σ^2 increases. Similarly, for given values of σ^2 and r_{23}, the variance of $\hat{\beta}_2$ is inversely proportional to $\sum x_{2i}^2$; that is, the greater the variation in the sample values of X_2, the smaller the variance of $\hat{\beta}_2$ and therefore β_2 can be estimated more precisely

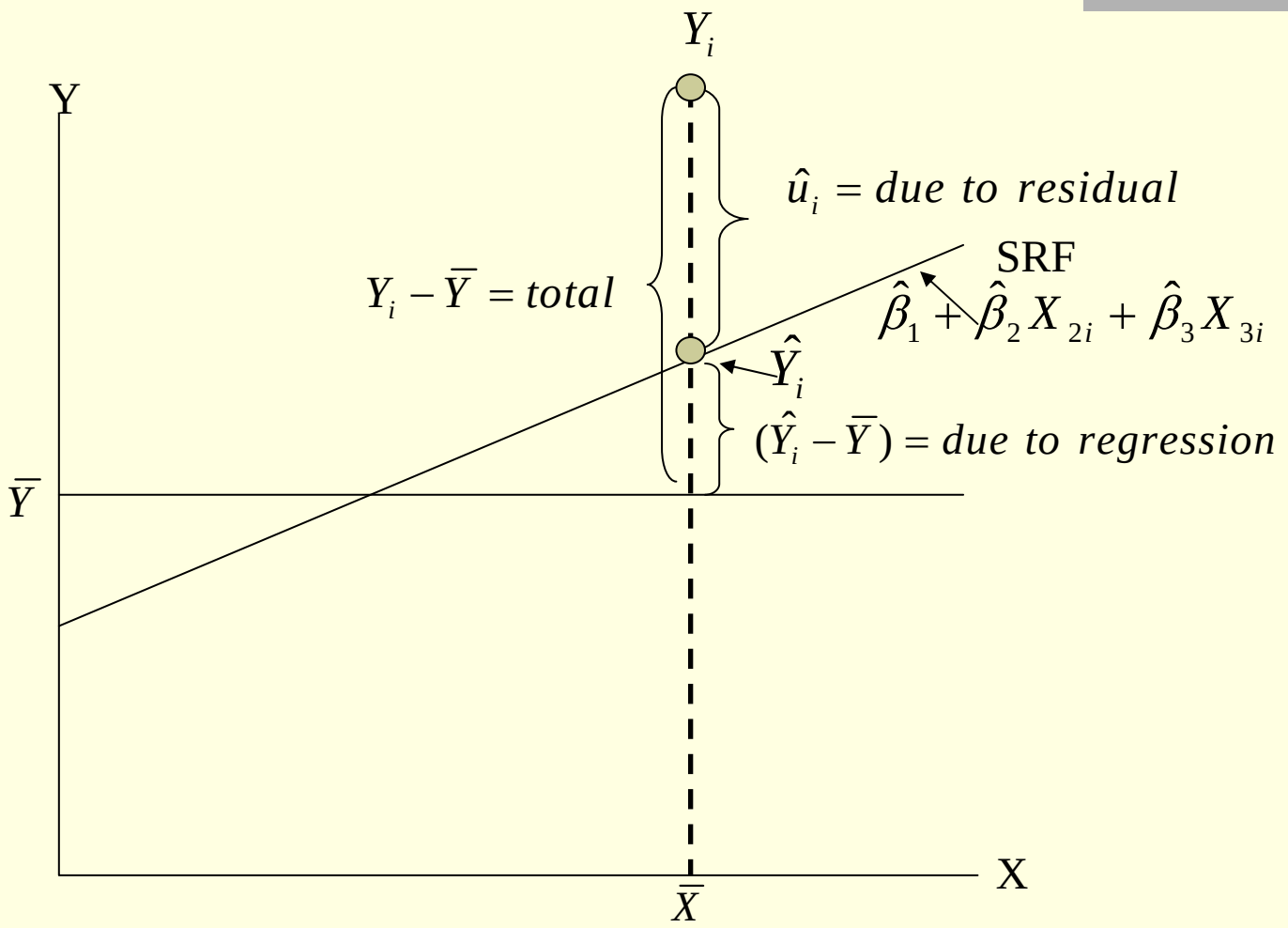
- Given the assumptions of the classical linear regression model, one can prove that the OLS estimators on the partial regression coefficients not only are linear and unbiased but also have minimum variance in the class of all linear unbiased estimators. (BLUE)

The Multiple Coefficient of Determinant (R^2)

- Measure the goodness of fit of the regression equation
- We would like to know the proportion of the variation of Y explained by the variables jointly.
- The quantity that gives this information is known as the **multiple coefficient of determinant**

The Multiple Coefficient of Correlation (R)

- Measure of the degree of association between Y and all the explanatory variables jointly


Y_i
Y
$\hat{u}_i = due\ to\ residual$
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
$Y_i - \bar{Y} = total$
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y}) = due\ to\ regression$
$\bar{Y}$
X
$\bar{X}$

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$$

Total variation of the actual Y values about their sample mean

(Total Sum of Squares, TSS)

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \hat{Y})^2$$

Variation of the estimated Y values about their mean (Explained Sum of Squares, ESS)

$$\sum \hat{u}_i^2$$

Residual or unexpected variation of the Y values about the regression line (Residual Sum of Squares, RSS)

$$TSS = ESS + RSS$$

R-squared

$$R^2 = \frac{ESS}{TSS}$$

$$= 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

R-squared

- Lies between 0 and 1
- Nondecreasing function of the number of explanatory variables

Relationship between R-Squared and the variance of a partial regression coefficient in the k variable multiple regression model

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right)$$

Where $\hat{\beta}_j$ is the partial regression coefficient of regressor X_j and R_j^2 is the R^2 in the regression of X_j on the remaining (k-2) regressors.

Example- Child mortality

$$Y = Child\ \ mortality\ \ (CM)$$

$$X_2 = per\ \ capita\ GNP\ \ (PGNP)$$

$$X_3 = Female\ literacy\ \ rate(FLR)$$

TABLE 6.4 Fertility and Other Data for 64 Countries

Observation	CM	FLFP	PGNP	TFR	Observation	CM	FLFP	PGNP	TFR
1	128	37	1870	6.66	33	142	50	8640	7.17
2	204	22	130	6.15	34	104	62	350	6.60
3	202	16	310	7.00	35	287	31	230	7.00
4	197	65	570	6.25	36	41	66	1620	3.91
5	96	76	2050	3.81	37	312	11	190	6.70
6	209	26	200	6.44	38	77	88	2090	4.20
7	170	45	670	6.19	39	142	22	900	5.43
8	240	29	300	5.89	40	262	22	230	6.50
9	241	11	120	5.89	41	215	12	140	6.25
10	55	55	290	2.36	42	246	9	330	7.10
11	75	87	1180	3.93	43	191	31	1010	7.10
12	129	55	900	5.99	44	182	19	300	7.00
13	24	93	1730	3.50	45	37	88	1730	3.46
14	165	31	1150	7.41	46	103	35	780	5.66
15	94	77	1160	4.21	47	67	85	1300	4.82
16	96	80	1270	5.00	48	143	78	930	5.00
17	148	30	580	5.27	49	83	85	690	4.74
18	98	69	660	5.21	50	223	33	200	8.49
19	161	43	420	6.50	51	240	19	450	6.50
20	118	47	1080	6.12	52	312	21	280	6.50
21	269	17	290	6.19	53	12	79	4430	1.69
22	189	35	270	5.05	54	52	83	270	3.25
23	126	58	560	6.16	55	79	43	1340	7.17
24	12	81	4240	1.80	56	61	88	670	3.52
25	167	29	240	4.75	57	168	28	410	6.09
26	135	65	430	4.10	58	28	95	4370	2.86
27	107	87	3020	6.66	59	121	41	1310	4.88
28	72	63	1420	7.28	60	115	62	1470	3.89
29	128	49	420	8.12	61	186	45	300	6.90
30	27	63	19830	5.23	62	47	85	3630	4.10
31	152	84	420	5.79	63	178	45	220	6.09
32	224	23	530	6.50	64	142	67	560	7.20

Note: CM = Child mortality, the number of deaths of children under age 5 in a year per 1000 live births.
FLFP = Female literacy rate, percent.
PGNP = per capita GNP in 1980.
TFR = total fertility rate, 1980–1985, the average number of children born to a woman, using age-specific fertility rates for a given year.

Source: Chandan Mukherjee, Howard White, and Marc Whyte, *Econometrics and Data Analysis for Developing Countries,* Routledge, London, 1998, p. 456.

CM is the number of deaths of children under five per 1000 live births

PGNP is per capita GNP in 1980

FLR is measured in percent

$$CM_i = \beta_1 + \beta_2 PGNP_i + \beta_3 FLR_i + u_i$$

$$\widehat{CM}_i = 263.6416 - 0.0056 PGNP_i - 2.2316 FLR_i$$

$$se = (11.5932) \quad (0.0019) \quad (0.2099)$$

$$R^2 = 0.7077$$

As PGNP increases by one thousand dollars, on average, the number of deaths of children under age 5 goes down by about 5.6 per thousand live births.

Holding the influence of PGNP constant, on average, the number of deaths of children under age 5 goes down by about 2.23 per thousand live births as the female literacy rate increases by one percentage point

If the values of PGNP and FLR rate were fixed at zero, the mean child mortality rate would be about 263 deaths per thousand live births

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2 / (n-k)}{\sum y_i^2 / (n-1)}$$

k = the number of parameters in the model including the intercept term

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

- For k>1, Adjusted R-Squared < R-Squared implies that as the number of X variables increases, the adjusted R-Squared increases less than the unadjusted R-Squared
- Adjusted R-Squared can be negative

Other criteria are often used to judge the adequacy of a regression model

- Akaike's Information criterion
- Amemiya's Prediction criteria

Comparing Two R-Squared values

- The Sample Size n must be the same
- The dependent variable must be the same

Functional Form

- The Cobb-Douglas Production Function
- Polynomial Regression Models

The Cobb-Douglas Production Function

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{\mu_i}$$

$Y = output$

$X_2 = labor\ \ input$

$X_3 = capital\ \ input$

$u = stochastic\ \ disturbance\ term$

$e = base\ of\ \ natural\ \ \log$

$$\beta_2 + \beta_3 = 1 \quad CONSTANT\ \ RETURN\ \ TO\ \ SCALE$$

$$\beta_2 + \beta_3 < 1 \quad DECREASING\ \ RETURN\ \ TO\ \ SCALE$$

$$\beta_2 + \beta_3 > 1 \quad INCREASING\ \ RETURN\ \ TO\ \ SCALE$$

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

$$= \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

Example

TABLE 7.3
Value Added, Labor Hours, and Capital Input in the Manufacturing Sector of the U.S., 2005

Source: *2005 Annual Survey of Manufacturers*, Sector 31: Supplemental Statistics for U.S.

Area	Output Value Added (thousand of $) *Y*	Labor Input Worker Hrs (thousands) *X2*	Capital Input Capital Expenditure (thousands of $) *X3*
Alabama	38,372,840	424,471	2,689,076
Alaska	1,805,427	19,895	57,997
Arizona	23,736,129	206,893	2,308,272
Arkansas	26,981,983	304,055	1,376,235
California	217,546,032	1,809,756	13,554,116
Colorado	19,462,751	180,366	1,790,751
Connecticut	28,972,772	224,267	1,210,229
Delaware	14,313,157	54,455	421,064
District of Columbia	159,921	2,029	7,188
Florida	47,289,846	471,211	2,761,281
Georgia	63,015,125	659,379	3,540,475
Hawaii	1,809,052	17,528	146,371
Idaho	10,511,786	75,414	848,220
Illinois	105,324,866	963,156	5,870,409
Indiana	90,120,459	835,083	5,832,503
Iowa	39,079,550	336,159	1,795,976
Kansas	22,826,760	246,144	1,595,118
Kentucky	38,686,340	384,484	2,503,693
Louisiana	69,910,555	216,149	4,726,625

Maine	7,856,947	82,021	415,131
Maryland	21,352,966	174,855	1,729,116
Massachusetts	46,044,292	355,701	2,706,065
Michigan	92,335,528	943,298	5,294,356
Minnesota	48,304,274	456,553	2,833,525
Mississippi	17,207,903	267,806	1,212,281
Missouri	47,340,157	439,427	2,404,122
Montana	2,644,567	24,167	334,008
Nebraska	14,650,080	163,637	627,806
Nevada	7,290,360	59,737	522,335
New Hampshire	9,188,322	96,106	507,488
New Jersey	51,298,516	407,076	3,295,056
New Mexico	20,401,410	43,079	404,749
New York	87,756,129	727,177	4,260,353
North Carolina	101,268,432	820,013	4,086,558
North Dakota	3,556,025	34,723	184,700
Ohio	124,986,166	1,174,540	6,301,421
Oklahoma	20,451,196	201,284	1,327,353
Oregon	34,808,109	257,820	1,456,683
Pennsylvania	104,858,322	944,998	5,896,392
Rhode Island	6,541,356	68,987	297,618
South Carolina	37,668,126	400,317	2,500,071
South Dakota	4,988,905	56,524	311,251
Tennessee	62,828,100	582,241	4,126,465
Texas	172,960,157	1,120,382	11,588,283
Utah	15,702,637	150,030	762,671
Vermont	5,418,786	48,134	276,293
Virginia	49,166,991	425,346	2,731,669
Washington	46,164,427	313,279	1,945,860
West Virginia	9,185,967	89,639	685,587
Wisconsin	66,964,978	694,628	3,902,823
Wyoming	2,979,475	15,221	361,536

$$\widehat{\ln Output_i} = 3.8876 + 0.4683 \ln labor_{2i} + 0.5213 \ln capital_{3i}$$

$$t = (9.8115) \quad (4.7342) \qquad (5.3803)$$

$$R^2 = 0.9642$$

$$\bar{R}^2 = 0.9627$$

- The U.S. manufacturing sector for 2005, the output elasticities of labor and capital were 0.4683 and 0.5213, respectively

Polynomial Regression Models

$$Y = \beta_0 + \beta_1 X + \beta_1 X^2$$

The stochastic version may be written as

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

which is called a second-degree polynomial regression

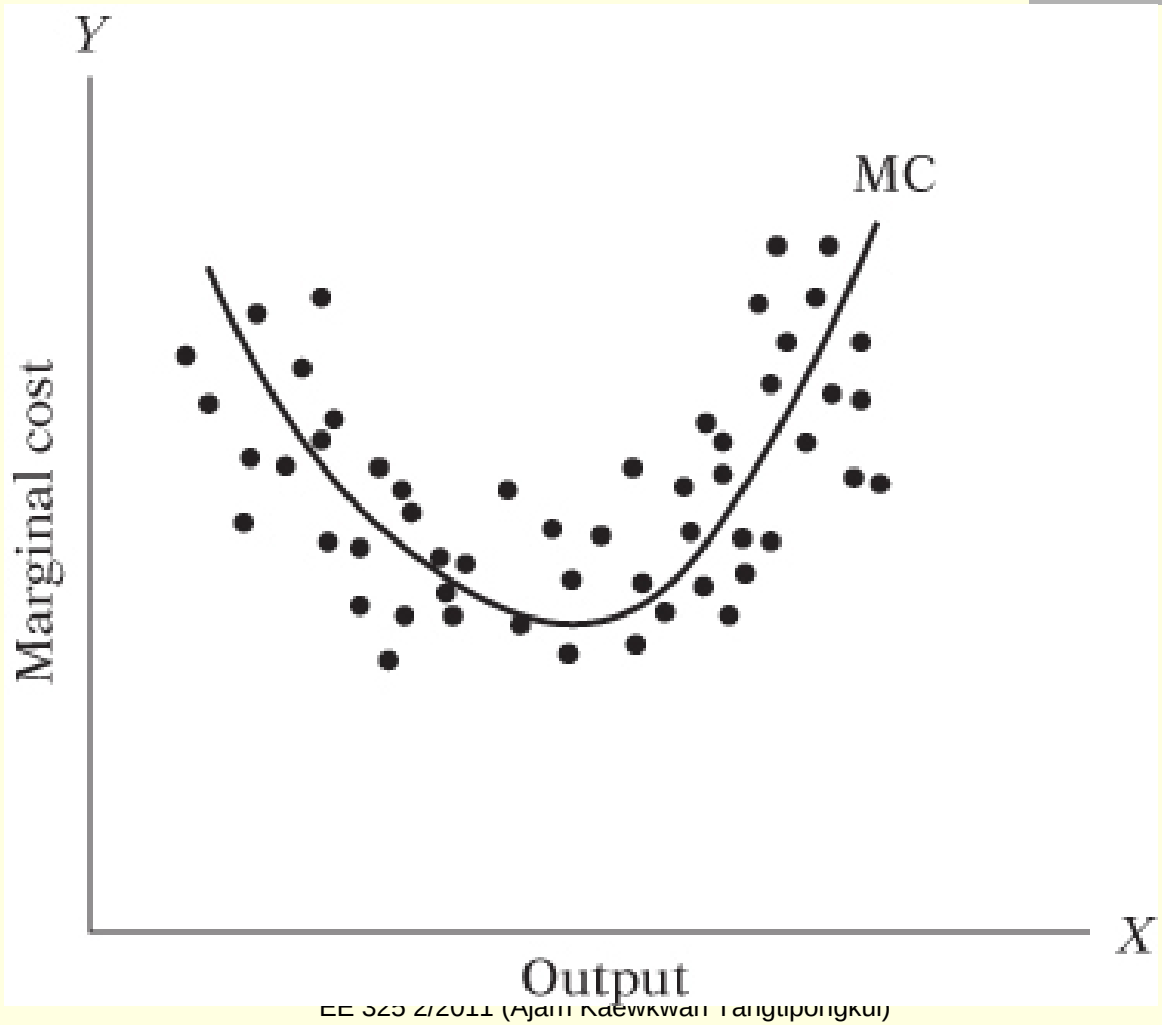
The general kth degree polynomial regression

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k + u_i$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$$

$$\frac{\partial \hat{Y}_i}{\partial X_i} = \hat{\beta}_1 + 2\hat{\beta}_2 X_i$$

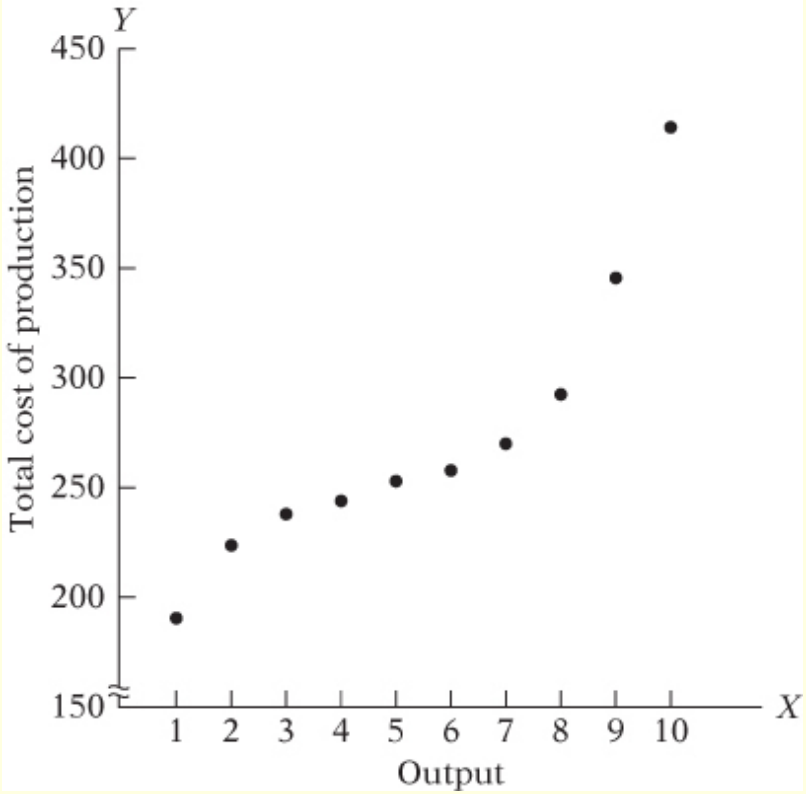
When X increases one unit, Y will increase or decrease by $\hat{\beta}_1 + 2\hat{\beta}_2 X_i$ unit


Y
MC
Marginal cost
X
Output

Example:
Estimating the total cost function

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



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + \hat{\beta}_3 X_i^3$$

Source	SS	df	MS
Model	38918.1562	3	12972.7187
Residual	64.7438228	6	10.7906371
Total	38982.9	9	4331.43333

Number of obs	10
F(3, 6)	1202.22
Prob > F	0.0000
R-squared	0.9983
Adj R-squared	0.9975
Root MSE	3.2849

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	63.47766	4.778607	13.28	0.000	51.78483	75.17049
x2	-12.96154	.9856646	-13.15	0.000	-15.37337	-10.5497
x3	.9395882	.0591056	15.90	0.000	.794962	1.084214
_cons	141.7667	6.375322	22.24	0.000	126.1668	157.3665

$$\hat{Y}_i = 141.7667 + 63.4776 X_i - 12.9615 X_i^2 + 0.9396 X_i^3$$

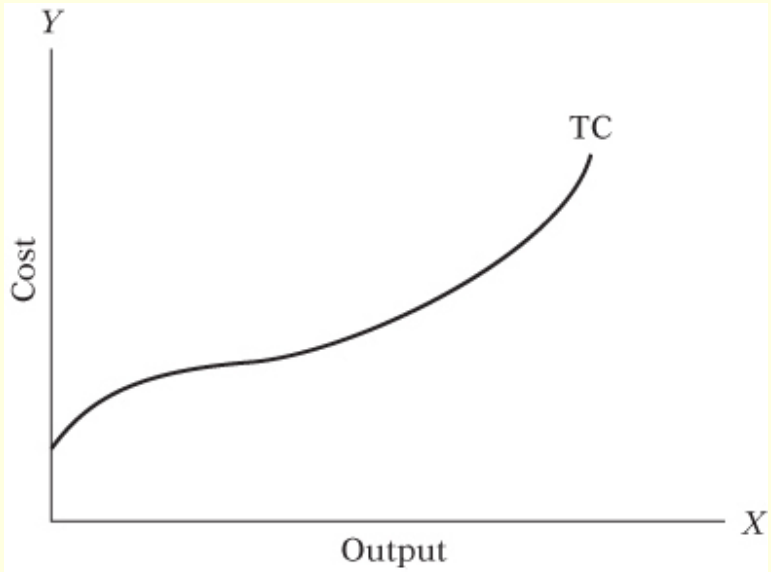
(6.3753) (4.7786) (0.9857) (0.0591)

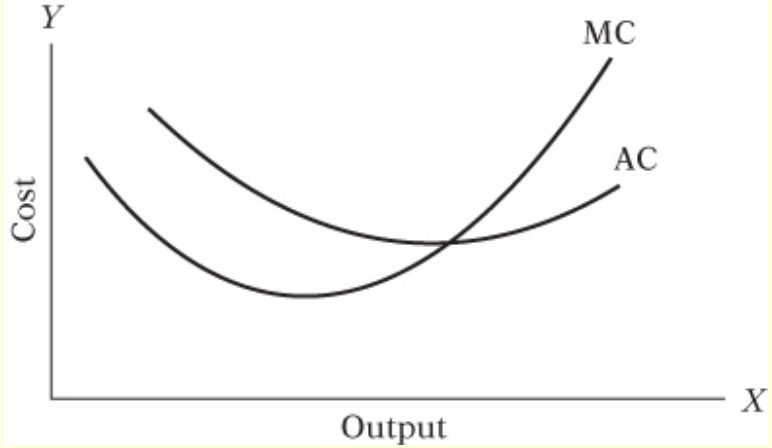
$$R^2 = 0.9983$$

X = output

Y= Total cost ($)

Elementary price theory shows that in the short run the MC and 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


Y
TC
Cost
X
Output


Y
MC
AC
Cost
X
Output

Partial correlation coefficient

- The coefficient of correlation (r) as a measure of the degree of linear association between two variables
- For the three variable regression model we can compute three correlation coefficients

  r_{12} correlation between Y and X_2

  r_{13} correlation between Y and X_3

  r_{23} correlation between X_2 and X_3

r_{12} measure the degree of linear association between Y and X_2 when a third variable X_3 may be associated with both

$r_{12.3}$ Partial correlation coefficient between Y and X_2 holding X_3 constant

Partial correlation coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

Relationship between R^2 , simple correlation coefficients and partial correlation coefficients

$$R^2 = r_{12}^2 + (1 - r_{12}^2) r_{13.2}^2$$

$$R^2 = r_{13}^2 + (1 - r_{13}^2) r_{12.3}^2$$

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2}$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2) r_{13.2}^2$$

The proportion of the variation in Y explained by X_2 and X_3 jointly is the sum of two parts:

1. The part explained by X_2 alone (r_{12}^2)
2. The part not explained by X_2 $(1 - r_{12}^2)$ time proportion that is explained by X_3 after holding the influence of X_2 constant