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: The 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 10. Multiple Regression Analysis: The Problem of Estimation

Geetha Rani Prakasam,

Professor.

Recap

- Extension of Simple Linear Regression:–
- Different functional forms
- their transformation so as to apply LS
- estimation and interpretation

Outline

- **Introduction to Multiple Linear Regression**
- The three-Variable Model: Notation & Assumptions
- **Multicollinearity**
- **Interpretation of Multiple Regression**
- The meaning of partial regression coefficients
- **OLS estimators, Var & SE in the case of three-variable regression**
- Properties of OLS Estimators

Introduction to Multiple Linear Regression

- The two-variable model studied extensively in the previous lectures is often inadequate in practice.
- In our consumption–income example, for instance, it was assumed implicitly that only income X affects consumption Y.
- But economic theory is seldom so simple for, besides income, a number of other variables are also likely to affect consumption expenditure.
- An obvious example is wealth of the consumer.
- One more example, the demand for a commodity is likely to depend not only on its own price but also on the prices of other competing or complementary goods, income of the consumer, social status, etc.
- Therefore, we need to extend our simple two-variable regression model to cover models involving more than two variables.

Introduction to Multiple Linear Regression

- Adding more variables leads us to the discussion of multiple regression models, that is, models in which the dependent variable, or regressand, Y depends on two or more explanatory variables, or regressors.
- The simplest possible multiple regression model is three-variable regression, with one dependent variable and two explanatory variables.
- In this and the next two lectures we shall study this model.
- Throughout, we are concerned with multiple linear regression models, that is, models linear in the parameters; they may or may not be linear in the variables.

Introduction to Multiple Linear Regression

- Generalizing the two-variable population regression function (PRF) (2.4.2), we may write the three-variable PRF as:
- $\mathbf{Y_i = ß_1 + ß_2 X_{2i} + ß_3 X_{3i} + u_i}$ (7.1.1)
- where Y is the dependent variable, X2 and X3 the explanatory variables (or regressors), u the stochastic disturbance term, and i the ith observation; in case the data are time series, the subscript t will denote the t_{th} observation.
- In Eq. (7.1.1) β1 is the intercept term.
- As usual, it gives the mean or average effect on Y of all the variables excluded from the model, although its mechanical interpretation is the average value of Y when X2 and X3 are set equal to zero.
- **$ß_2$, $ß_3$ are partial regression coefficients**

7-1. The three-Variable Model: Notation and Assumptions

- We continue to operate within the framework of the classical linear regression model (CLRM) first introduced in Lectures 2 & 3.
- Specifically, we assume the following:

1. Zero mean value of $U_{i:}$: $E(u_i | X_{2i}, X_{3i}) = 0. \ \forall i$ (7.1.2)
2. No serial correlation: $Cov(u_i, u_j) = 0, \ \forall i \# j$ (7.1.3)
3. **Homoscedasticity: $Var(u_i) = \sigma^2$** **(7.1.4)**
4. $Cov(u_i, X_{2i}) = Cov(u_i, X_{3i}) = 0$ (7.1.5)
5. No specification bias or model correct specified (7.1.6)

The three-Variable Model: Notation & Assumptions

6. No exact collinearity between X variables (7.1.7)

(no multicollinearity in the cases of more explanatory vars. If there is linear relationship exits, X vars. are said to be linearly dependent)

7. Model is linear in parameters

8. The Number of Observations **n** Must Be Greater than the Number of Parameters to Be Estimated.

Alternatively, the number of observations **n** must be greater than the number of explanatory variables.

The three-Variable Model: Notation & Assumptions

9. The Nature of X Variables: The X values in a given sample must not all be the same.

- Technically, var (X) must be a positive number. Furthermore, there can be no **outliers in** the values of the X variable, that is, values that are very large in relation to the rest of the observations.
- 10. Zero covariance between u_i and X_i
- $$\text{Cov}(u_i, X_i) = E(u_i, X_i) = 0$$

More on Assumptions of three variable Model

- In addition, like earlier, we assume that the multiple regression model is linear in the parameters, that the values of the regressors are fixed in repeated sampling, and that there is sufficient variability in the values of the regressors.
- Assumption (7.1.7), that there be no exact linear relationship between X_2 and X_3, technically known as the assumption of no collinearity or no multicollinearity if more than one exact linear relationship is involved, is new and needs some explanation.
- Informally, no collinearity means none of the regressors can be written as exact linear combinations of the remaining regressors in the model.

More on Assumptions of three variable Model

- Formally, no collinearity means that there exists no set of numbers, λ_2 and λ_3, not both zero such that

$$\lambda_2 X_{2i} + \lambda_3 X_{3i} = 0 \qquad (7.1.8)$$

- If such an exact linear relationship exists, then X2 and X3 are said to be collinear or linearly dependent.
- On the other hand, if (7.1.8) holds true only when $\lambda_2 = \lambda_3 = 0$, then X_2 and X_3 are said to be linearly independent.

More on Assumptions of three variable Model

- Thus, if
- $X_2i = -4X_3i$ or $X_2i + 4X_3i = 0$ (7.1.9)
- the two variables are linearly dependent, and if both are included in a regression model, we will have perfect collinearity or an exact linear relationship between the two regressors.
- Intuitively the logic behind the assumption of no multicollinearity is not too difficult to grasp.
- Suppose that in (7.1.1) Y, X_2, and X_3 represent consumption expenditure, income, and wealth of the consumer, respectively.
- In postulating that consumption expenditure is linearly related to income and wealth, economic theory presumes that wealth and income may have some independent influence on consumption.

Multicollinearity

- If not, there is no sense in including both income and wealth variables in the model.
- In the extreme, if there is an exact linear relationship between income and wealth, we have only one independent variable, not two, and there is no way to assess the separate influence of income and wealth on consumption.
- To see this clearly, let $X_{3i} = 2X_{2i}$ in the consumption–income–wealth regression.
- Then the regression (7.1.1) becomes

Multicollinearity

$$
\begin{aligned}
Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i \\
&= \beta_1 + (\beta_2 + 2\beta_3) X_{2i} + u_i \qquad \textbf{(7.1.10)} \\
&= \beta_1 + \alpha X_{2i} + u_i
\end{aligned}
$$

- where $\alpha = (\beta_2 + 2\beta_3)$.
- That is, we in fact have a two-variable and not a three variable regression.
- Moreover, if we run the regression (7.1.10) and obtain α, there is no way to estimate the separate influence of X_2 (= β_2) and X_3 (= β_3) on Y, for α gives the combined influence of X_2 and X_3 on Y

Multicollinearity

- In short the assumption of no multicollinearity requires that in the PRF we include only those variables that are not exact linear functions of one or more variables in the model.
- A couple of points may be noted here.
- First, the assumption of no multicollinearity pertains to our theoretical (i.e., PRF) model.
- In practice, when we collect data for empirical analysis there is no guarantee that there will not be correlations among the regressors.
- As a matter of fact, in most applied work it is almost impossible to find two or more (economic) variables that may not be correlated to some extent, as we will show in our illustrative examples later in lecture 12.
- What we require is that there be no exact relationships among the regressors, as in Eq. (7.1.9).

Multicollinearity

- Second, keep in mind that we are talking only about perfect linear relationships between two or more variables.
- Multicollinearity does not rule out nonlinear relationships between variables.
- Suppose $X_{3i} = X^2_{2i}$. This does not violate the assumption of no perfect collinearity, as the relationship between the variables here is nonlinear.

7-2. Interpretation of Multiple Regression

- $\mathbf{E(Y_i | X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}}$ (7.2.1)
- **(7.2.1) gives conditional mean or expected value of Y conditional upon the given or fixed value of the X_2 and X_3.**
- In words, (7.2.1) gives the conditional mean or expected value of Y conditional upon the given or fixed values of X_2 and X_3.
- Therefore, as in the two-variable case, multiple regression analysis is regression analysis conditional upon the fixed values of the regressors, and what we obtain is the average or mean value of Y or the mean response of Y for the given values of the regressors.

7-3. The meaning of partial regression coefficients

- $\mathbf{Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_3 + \ldots + \beta_s X_s + u_i}$
- $\boldsymbol{\beta}_k$ measures the change in the mean value of Y per unit change in X_k, holding the rest of the explanatory variables constant.
- It gives the "direct" effect of unit change in X_k on the $E(Y_i)$, net of X_j (j # k)
- How to control the "true" effect of a unit change in X_k on Y? – meaning of partial regression coefficients

7-3. The meaning of partial regression coefficients

- Regression coefficients $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ are known as partial regression or partial slope coefficients.
- The meaning of partial regression coefficient is as follows: $\boldsymbol{\beta_2}$ measures the change in the mean value of Y, E(Y), per unit change in X_2, holding the value of X_3 constant.
- Put differently, it gives the "direct" or the "net" effect of a unit change in X_2 on the mean value of Y, net of any effect that X_3 may have on mean Y.
- Likewise, β_3 measures the change in the mean value of Y per unit change in X_3, holding the value of X_2 constant.
- That is, it gives the "direct" or "net" effect of a unit change in X_3 on the mean value of Y, net of any effect that X_2 may have on mean Y.

Partial Regression Coefficients: Example

- How do we actually go about holding the influence of a regressor constant?
- To explain this, let us take the child mortality example.
- Example, Y = child mortality (CM), X2 = per capita GNP (PGNP), and X_3 = female literacy rate (FLR).
- Let us suppose we want to hold the influence of FLR constant. Since FLR may have some effect on CM as well as PGNP in any given concrete data, what we can do is to remove the (linear) influence of FLR from both CM and PGNP by running the regression of CM on FLR and that of PGNP on FLR separately and then looking at the residuals obtained from these regressions.
- Using the data given in Table 6.4, we obtain the following regressions

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

Observation	CM	FLFP	PGNP	TFR	Observation	CM	FLFP	PGNP	TFR
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

Partial Regression Coefficients: Example

- $CM_i = 263.8635 - 2.3905\ FLR_i + \hat{u}_{1i}$ (7.3.1)
- se = (12.2249) (0.2133) $r^2 = 0.6695$ where $\hat{u}_{1i}$ represents the residual term of this regression.
- $PGNP_i = -39.3033 + 28.1427\ FLR_i + \hat{u}_{2i}$ (7.3.2)
- se = (734.9526) (12.8211) $r^2 = 0.0721$ where $\hat{u}_{2i}$ represents the residual term of this regression.
- Now $\hat{u}_{1i} = (CMi_i - 263.8635 + 2.3905\ FLR_i)$ (7.3.3) represents that part of CM left after removing from it the (linear) influence of FLR.
- Likewise, $\hat{u}_{2i} = (PGNP_i + 39.3033 - 28.1427\ FLR_i)$ (7.3.4)

Partial Regression Coefficients: Example

- represents that part of PGNP left after removing from it the (linear) influence of FLR.
- Therefore, if we now regress $\hat{u}_{1i}$ on $\hat{u}_{2i}$, which are "purified" of the (linear) influence of FLR, wouldn't we obtain the net effect of PGNP on CM?
- That is indeed the case. The regression results are as follows:
- $\hat{u}_{1i} = -0.0056\ \hat{u}_{2i}$ (7.3.5)
  - se = (0.0019) r2 = 0.1152
- Note: This regression has no intercept term because the mean value of the OLS residuals $\hat{u}_{1i}$ and $\hat{u}_{2i}$ is zero.

Partial Regression Coefficients: Example

- The slope coefficient of −0.0056 now gives the “true” or net effect of a unit change in PGNP on CM or the true slope of CM with respect to PGNP.
- That is, it gives the partial regression coefficient of CM with respect to PGNP, β2.
- Do we have to go through this multistep procedure every time we want to find out the true partial regression coefficient?
- Fortunately, not – the same can be accomplished fairly quickly and routinely by the OLS procedure discussed in the next section.
- The multistep procedure – for pedagogic purposes to understand the meaning of “partial” regression coefficient.

OLS estimators in the case of three-variable regression

- Yi = βˆ 1 + βˆ 2X2i + βˆ 3X3i + ˆui (7.4.1)
- residual sum of squares (RSS) u^2 i is as small as possible: LS criterion

$$\min \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2 \qquad \textbf{(7.4.2)}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \qquad \textbf{(7.4.6)}$$

OLS estimators in the case of three-variable regression

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \qquad (7.4.7)$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \qquad (7.4.8)$$

Note: (1) Equations (7.4.7) and (7.4.8) are symmetrical in nature because one can be obtained from the other by interchanging the roles of X2 and X3;
(2) the denominators of these two equations are identical; and
(3) the three-variable case is a natural extension of the two-variable case.

Variances and Standard Errors of OLS Estimators

- Variance of a partial regression coefficient
- $\text{Var}(\hat{\beta}_k) = \sigma^2 / \sum x^2_k \, (1/(1-R^2_k))$ (7.5.6)
- Where $\hat{\beta}_k$ is the partial regression coefficient of regressor X_k and R^2_k is the R^2 in the regression of X_k on the rest regressors.
- An unbiased estimator of σ 2 is given by
- $\hat{\sigma}^2 = \sum u_i^2 / n-3$ (7.4.18)
- Note the similarity between this estimator of σ 2 and its two-variable counterpart $[\hat{\sigma}^2 = (\sum \hat{u}_i^2)/(n-2)]$.
- The degrees of freedom are now (n− 3) because in estimating $\sum \hat{u}_i^2$ we must first estimate β_1 , β_2, and β_3, which consume 3 df.

Variances and Standard Errors of OLS Estimators

- (The argument is quite general. Thus, in the four-variable case the df will be n− 4.)
- The estimator $\hat{\sigma}^2$ can be computed from (7.4.18) once the residuals are available,
- but it can also be obtained more readily by using the following relation :) $\sum \hat{u}_i^2 = y_i^2 - \hat{\beta}_2 y_i x_{2i} - \hat{\beta}_3 y_i x_{3i}$ (7.4.19) which is the three-variable counterpart of the relation given in (3.3.6).

Properties of OLS Estimators

- The properties of OLS estimators of the multiple regression model parallel those of the two-variable model.
- Specifically:
- 1. The three-variable regression line (surface) passes through the means $\bar{y}$, $\bar{x}_2$, and $\bar{x}_3$, which is evident from (7.4.3) [cf. Eq. (3.1.7) of the two variable model].
- This property holds generally.
- Thus in the k-variable linear regression model [a regressand and (k − 1) regressors].

Properties of OLS Estimators

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad \textbf{(7.4.20)}$$

we have

$$\hat{\beta}_1 = \bar{Y} - \beta_2 \bar{X}_2 - \beta_3 \hat{X}_3 - \cdots - \beta_k \bar{X}_k \qquad \textbf{(7.4.21)}$$

- 2. The mean value of the estimated Yi (= Yˆ i) is equal to the mean value of the actual Yi, which is easy to prove:

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3) + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \qquad \text{(Why?)} \\ &= \bar{Y} + \hat{\beta}_2 (X_{2i} - \bar{X}_2) + \hat{\beta}_3 (X_{3i} - \bar{X}_3) \\ &= \bar{Y} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \end{aligned} \qquad \textbf{(7.4.22)}$$

Properties of OLS Estimators

- where as usual small letters indicate values of the variables as deviations from their respective means.
- Summing both sides of (7.4.22) over the sample values and dividing through by the sample size n gives $\bar{y}\hat{} = \bar{y}$.
- (Note: $x_{2i} = x_{3i} = 0$. Why?) Notice that by virtue of (7.4.22) we can write
- $y\hat{}_i = \beta\hat{}_2 x_{2i} + \beta\hat{}_3 x_{3i}$ (7.4.23)
- where $y\hat{}_i = (Y\hat{}_i - \bar{y})$.
- Therefore, the SRF (7.4.1) can be expressed in the deviation form as:

Properties of OLS Estimators

- $Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_2 x_{23} + \hat{\beta}_3 x_{3i} + \hat{u}i$ (7.4.24)
- 3. $\hat{u}_i = \bar{u}\hat{} = 0$, which can be verified from (7.4.24).
- [Sum both sides of (7.4.24) over the sample values.]
- 4. The residuals $\hat{u}_i$ are uncorrelated with X_{2i} and X_{3i}, that is, $\hat{u}_i X_{2i} = \hat{u}_i X_{3i} = 0$.
- 5. The residuals $\hat{u}_i$ are uncorrelated with $\hat{Y}_i$; that is, $\hat{u}_i \hat{Y}_i = 0$.
- [Multiply (7.4.23) on both sides by $\hat{u}_i$ and sum over the sample values.]
- 6. From (7.4.12) and (7.4.15) it is evident that as r_{23}, the correlation coefficient between X_2 and X_3, increases toward 1, the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$ increase for given values of σ^2 and Σx_{2i}^2 or Σx_{3i}^2.

Properties of OLS Estimators

- In the limit, when $r_{23} = 1$ (i.e., perfect collinearity), these variances become infinite.
- Intuitively the we can see that as r_{23} increases it is going to be increasingly difficult to know what the true values of β_2 and β_3 are. [refer to Eq. (7.1.10).]
- 7. It is also clear from (7.4.12) and (7.4.15) that for given values of r_{23} and x_{2i}^2 or x_{3i}^2, the variances of the OLS estimators are directly proportional to σ^2; that is, they increase as σ^2 increases.

Properties of OLS Estimators

- Similarly, for given values of σ^2 and r_{23}, the variance of $\hat{\beta}_2$ is inversely proportional to 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.
- A similar statement can be made about the variance of $\hat{\beta}_3$.
- 8. Given the assumptions of the classical linear regression model, which are spelled out in Section 7.1, one can prove that the OLS estimators of the partial regression coefficients not only are linear and unbiased but also have minimum variance in the class of all linear unbiased estimators.
- In short, they are BLUE: Put differently, they satisfy the Gauss-Markov theorem.

7-5. The multiple coefficient of determination R^2

1. In the two-variable case we saw that r^2 as defined in (3.5.5) measures the goodness of fit of the regression equation; that is, it gives the proportion or percentage of the total variation in the dependent variable Y explained by the (single) explanatory variable X.
   This notation for r^2 can be easily extended to regression models containing more than two variables. Thus, in the three variable model we would like to know the proportion of the variation in Y explained by the variables X2 and X3 jointly.
   The quantity that gives this information is known as the multiple coefficient of determination and is denoted by R^2; conceptually it is akin to r^2.
2. **Definition of R^2 in the context of multiple regression like r^2 in the case of two-variable regression 2.**

7-5. The multiple coefficient of determination R^2

- To derive R^2, we may follow the derivation of r^2 given in Section 3.5.
- Recall that

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$
$$= \hat{Y}_i + \hat{u}_i \qquad \textbf{(7.5.1)}$$

- Squaring (7.5.2) on both sides and summing over the sample values, we obtain

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i$$
$$= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \qquad \text{(Why?)} \qquad \textbf{(7.5.3)}$$

7-5. The multiple coefficient of determination R^2

- Verbally, Eq. (7.5.3) states that the total sum of squares (TSS) equals the explained sum of squares (ESS) + the residual sum of squares (RSS). Now substituting for uˆ2 i from (7.4.19), we obtain

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

- which, on rearranging, gives

$$\text{ESS} = \sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} \qquad \textbf{(7.5.4)}$$

7-5. The multiple coefficient of determination R^2

- Now, by definition R^2 = ESS / TSS

$$= \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \qquad (7.5.5)^9$$

- [cf. (7.5.5) with (3.5.6)].
- Since the quantities entering (7.5.5) are generally computed routinely, R^2 can be computed easily.
- Note that R^2, like r^2, lies between 0 and 1. If it is 1, the fitted regression line explains 100 percent of the variation in Y.
- On the other hand, if it is 0, the model does not explain any of the variation in Y. Typically, however, R^2 lies between these extreme values.
- The fit of the model is said to be "better'' the closer is R^2 to 1.

7-5. The multiple coefficient of determination R^2

- Recall that in the two-variable case we defined the quantity r as the coefficient of correlation and indicated that it measures the degree of (linear) association between two variables.
- The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly.
- Although r can be positive or negative, R is always taken to be positive. In practice, however, R is of little importance.
- The more meaningful quantity is R^2.
- Before proceeding further, let us note the following relationship between R^2 and the variance of a partial regression coefficient in the k-variable multiple regression model given in (7.4.20):

7-5. The multiple coefficient of determination R^2

$$\text{var}\,(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1 - R_j^2}\right) \qquad \textbf{(7.5.6)}$$

- where $\hat{\beta}_j$ is the partial regression coefficient of regressor X_j and R^2 j is the R^2 in the regression of X_j on the remaining (k − 2) regressors. [Note:
- There are (k − 1) regressors in the k-variable regression model.]
- Although the utility of Eq. (7.5.6) will become apparent later.
- Now observe that this equation is simply an extension of the formula given in (7.4.12) or (7.4.15) for the three-variable regression model, one regressand and two regressors.

7-5. The multiple coefficient of determination R^2 & the multiple coefficient of correlation R

- **$R = \pm\sqrt{R^2}$ is the coefficient of multiple regression, it measures the degree of association between Y and all the explanatory variables jointly**

EXAMPLE 7.1: CHILD MORTALITY IN RELATION TO PER CAPITA GNP & FEMALE LITERACY RATE

- Our model is: $CM_i = \beta_1 + \beta_2 PGNP_i + \beta_3 FLR_i + u_i$ (7.6.1)
- Keep in mind that CM is the number of deaths of children under five per 1000 live births, PGNP is per capita GNP in 1980, and FLR is measured in percent. Our sample consists of 64 countries.
- Data in the next two slides
- 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

Observation	CM	FLFP	PGNP	TFR	Observation	CM	FLFP	PGNP	TFR
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

EX. 7.1: Child Mortality in Relation to Per Capita GNP & Female Literacy Rate

- $CM_i = 263.6416 - 0.0056\ PGNP_i - 2.2316\ FLR_i$
- se = (11.5932) (0.0019) (0.2099) (7.6.2)
- $R^2 = 0.7077$ Adjtd.$R^{-2} = 0.6981$*
- interpretation: −0.0056 is the partial regression coefficient of PGNP and tells us that with the influence of FLR held constant, as PGNP increases, say, by a dollar, on average, child mortality goes down by 0.0056 units.
- To make it more economically interpretable, if the per capita GNP goes up by a thousand dollars, on average, the number of deaths of children under age 5 goes down by about 5.6 per thousand live births.

EX. 7.1: Child Mortality in Relation to Per Capita GNP & Female Literacy Rate

- The coefficient −2.2316 tells us that holding the influence of PGNP constant, on average, the number of deaths of children under 5 goes down by about 2.23 per thousand live births as the female literacy rate increases by one percentage point.
- The intercept value of about 263, mechanically interpreted, means that if the values of PGNP and FLR rate were fixed at zero, the mean child mortality would be about 263 deaths per thousand live births.

EX. 7.1: Child Mortality in Relation to Per Capita GNP & Female Literacy Rate

- Such an interpretation should be taken with a grain of salt.
- All one could infer is that if the two regressors were fixed at zero, child mortality will be quite high, which makes practical sense.
- The R^2 value of about 0.71 means that about 71 percent of the variation in child mortality is explained by PGNP and FLR, a fairly high value considering that the maximum value of R^2 can at most be 1.
- All told, the regression results do make sense.

Summary and Conclusions

- 1. This lecture introduced the simplest possible multiple linear regression model, namely, the three-variable regression model. It is understood that the term linear refers to linearity in the parameters and not necessarily in the variables.
- 2. Although a three-variable regression model is in many ways an extension of the two-variable model, there are some new concepts involved, such as partial regression coefficients, adjusted and unadjusted (for degrees of freedom) R^2, and multicollinearity.

Summary and Conclusions

- 3. Although R^2 and adjusted R^2 are overall measures of how the chosen model fits a given set of data, their importance should not be overplayed.
- What is critical is the underlying theoretical expectations about the model in terms of a priori signs of the coefficients of the variables entering the model and, as it is shown in the following lecture, their statistical significance.
- The results presented in this lecture can be easily generalized to a multiple linear regression model involving any number of regressors.

References

Chapter 7:MULTIPLE REGRESSION ANALYSIS: The Problem of Estimation in Basic Econometrics by Domadar Gujarati.

What next?

- Comparing two R^2

- **Regression on Standardised Variables**
- **Introduction to Specification Bias**

- MULTIPLE REGRESSION ANALYSIS**:- The Problem of Inference**