Regression Analysis

Professorial fellow. G. Panda

IIT, Bhubaneswar, India

Regression Analysis

Regression analysis examines associative relationships between a metric dependent variable and one or more independent variables in the following ways:

- Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
- Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
- Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
- Predict the values of the dependent variable.
- Control for other independent variables when evaluating the contributions of a specific variable or set of variables.
- Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.

Statistics Associated with Bivariate Regression Analysis

- **Bivariate regression model**. The basic regression equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where Y = dependent or criterion variable, X = independent or predictor variable, β_0 = intercept of the line, β_1 = slope of the line, and e_i is the error term associated with the *i* th observation.

- **Coefficient of determination**. The strength of association is measured by the coefficient of determination, r^2. It varies between 0 and 1 and signifies the proportion of the total variation in Y that is accounted for by the variation in X.

- **Estimated or predicted value**. The estimated or predicted value of Y_i is $\hat{Y}_i = a + b\,x$, where $\hat{Y}_i$ is the predicted value of Y_i, and a and b are estimators of β_0 and β_1, respectively.

Statistics Associated with Bivariate Regression Analysis

- **Regression coefficient**. The estimated parameter *b* is usually referred to as the non-standardized regression coefficient.
- **Scattergram**. A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations.
- **Standard error of estimate**. This statistic, SEE, is the standard deviation of the actual *Y* values from the predicted $\hat{Y}$ values.
- **Standard error.** The standard deviation of *b*, SE_b, is called the standard error.

Statistics Associated with Bivariate Regression Analysis

- **Standardized regression coefficient**. Also termed the beta coefficient or beta weight, this is the slope obtained by the regression of *Y* on *X* when the data are standardized.

- **Sum of squared errors**. The distances of all the points from the regression line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, Σe_j^2

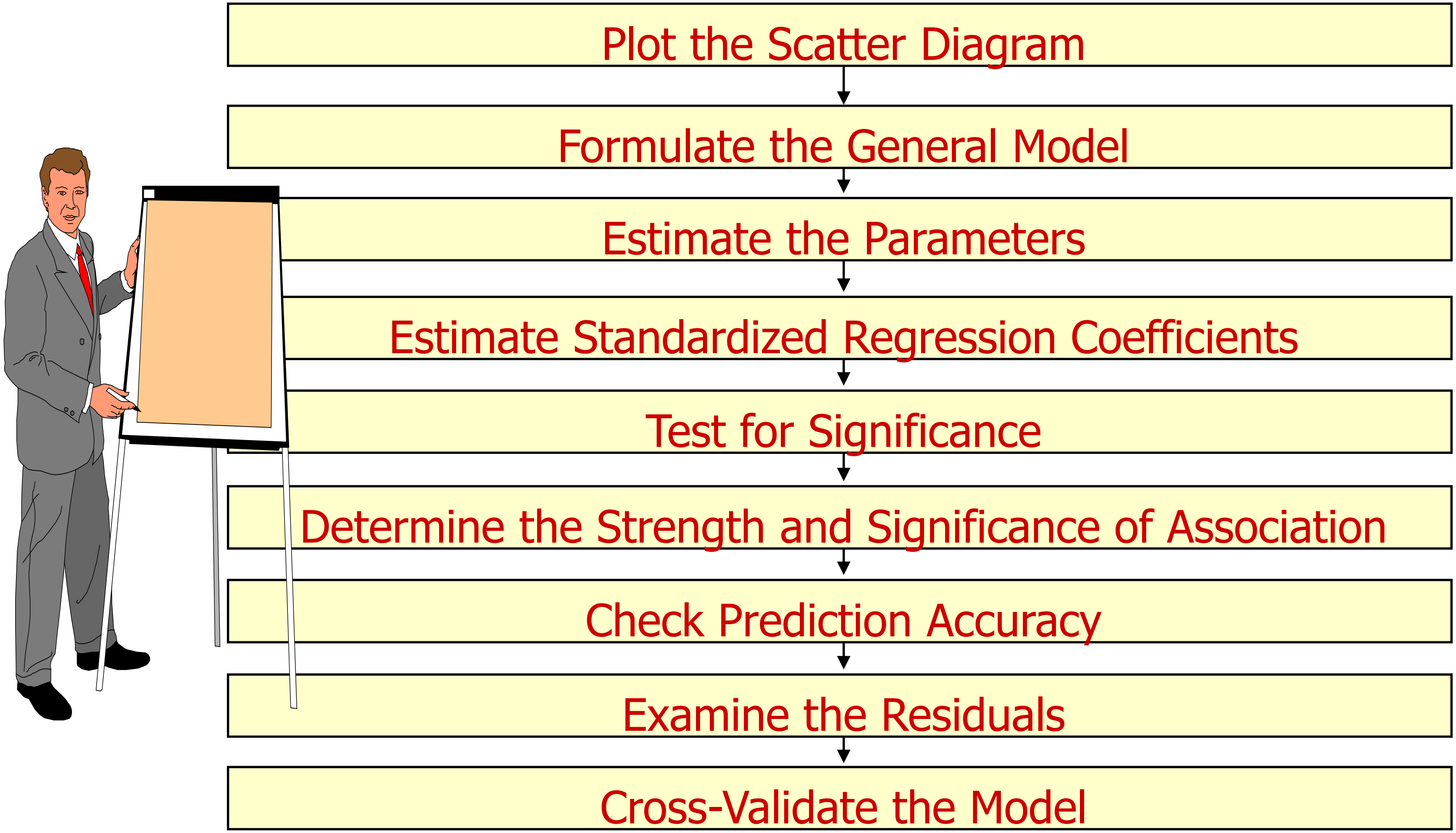
- ***t* statistic**. A *t* statistic with *n* - 2 degrees of freedom can be used to test the null hypothesis that no linear relationship exists between *X* and *Y*, or H_0: β = 0, where $t = b / SE_b$

Conducting Bivariate Regression Analysis
Plot the Scatter Diagram

- A **scatter diagram**, or **scattergram**, is a plot of the values of two variables for all the cases or observations.

- The most commonly used technique for fitting a straight line to a scattergram is the **least-squares procedure**. In fitting the line, the least-squares procedure minimizes the sum of squared errors, Σe_j^2.

Conducting Bivariate Regression Analysis



Plot the Scatter Diagram

Formulate the General Model

Estimate the Parameters

Estimate Standardized Regression Coefficients

Test for Significance

Determine the Strength and Significance of Association

Check Prediction Accuracy

Examine the Residuals

Cross-Validate the Model

Conducting Bivariate Regression Analysis
Formulate the Bivariate Regression Model

In the bivariate regression model, the general form of a straight line is: $Y = \beta_0 + \beta_1 X$

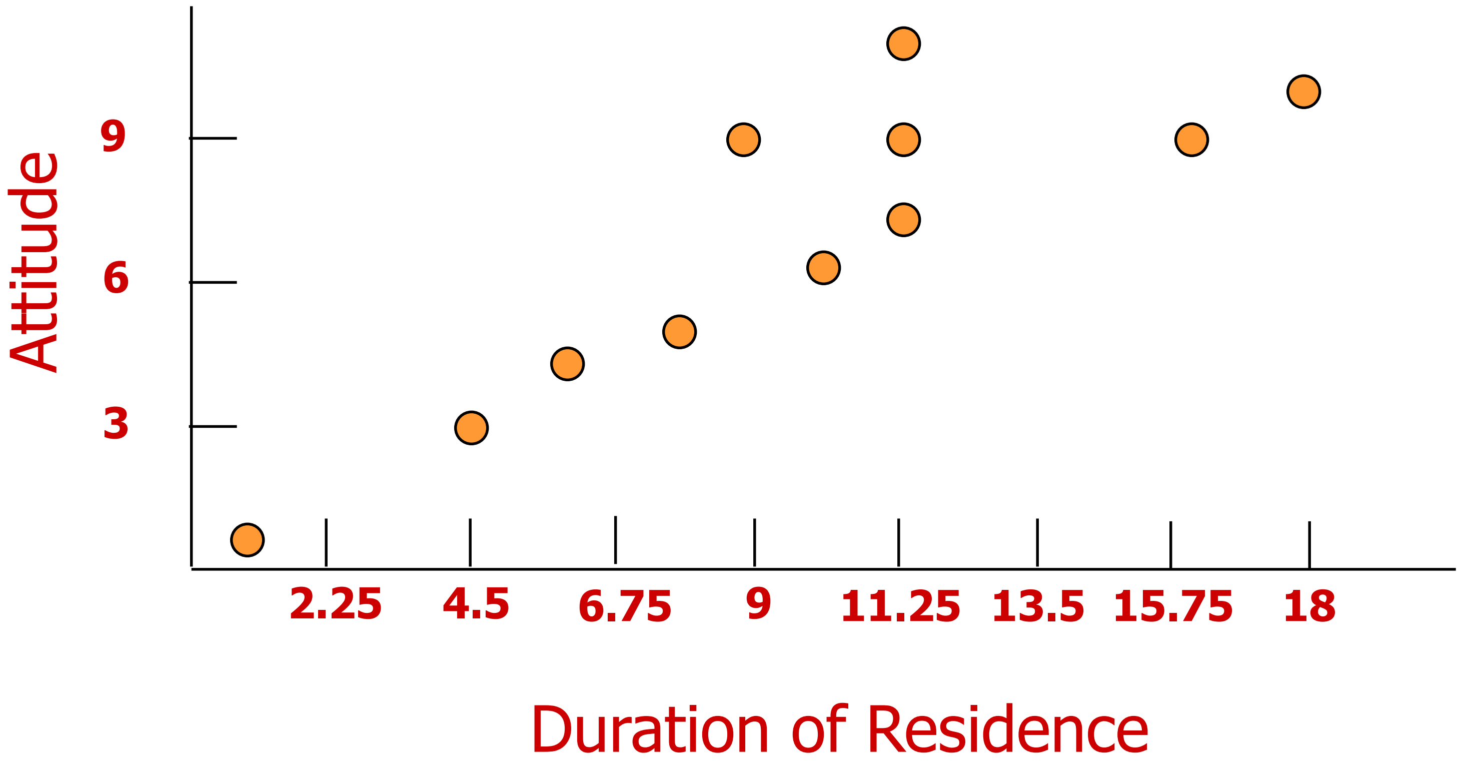
where
Y = dependent or criterion variable
X = independent or predictor variable
β_0= intercept of the line
β_1= slope of the line

The regression procedure adds an error term to account for the probabilistic or stochastic nature of the relationship:

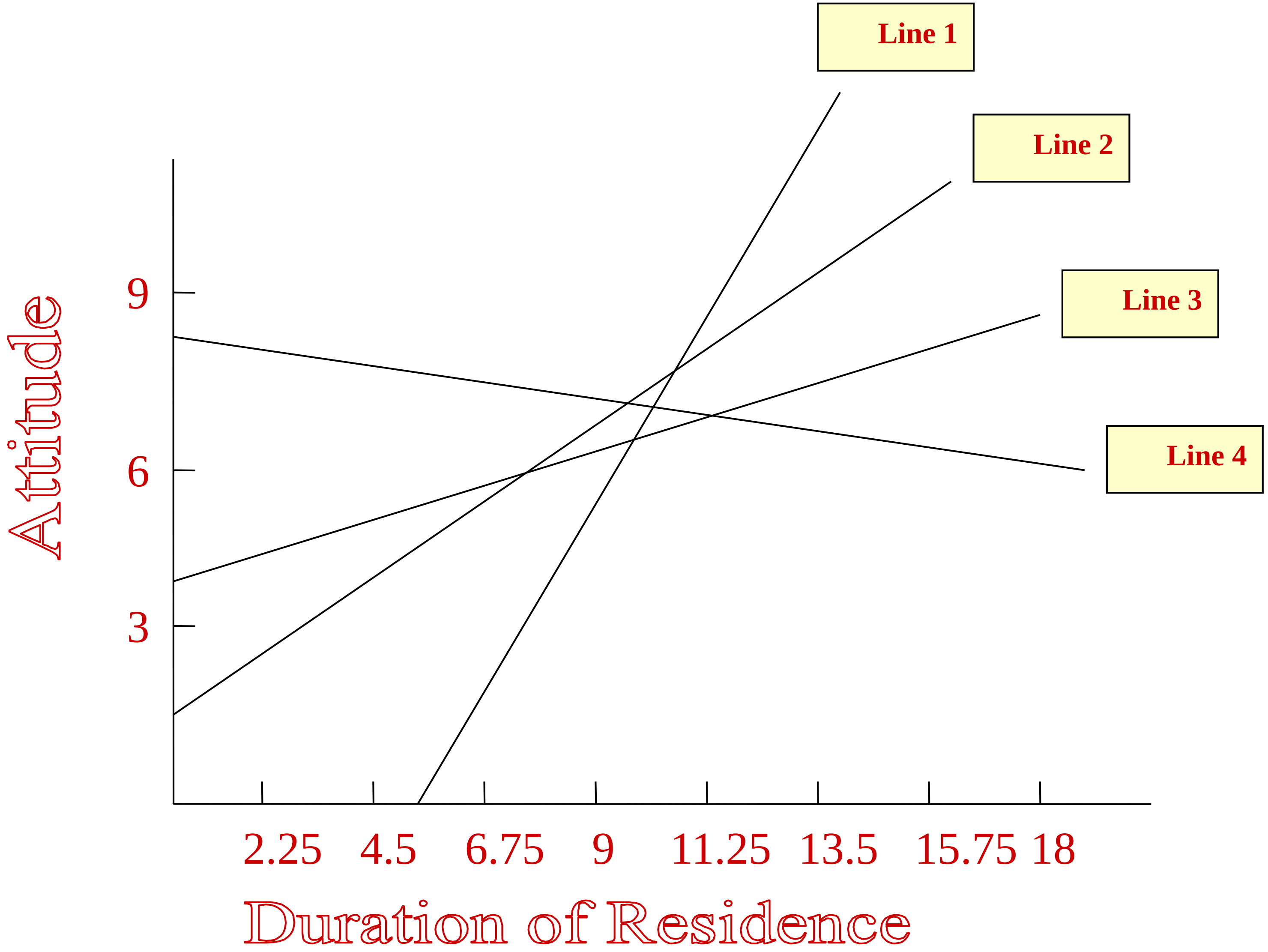
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

where e_i is the error term associated with the i th observation.

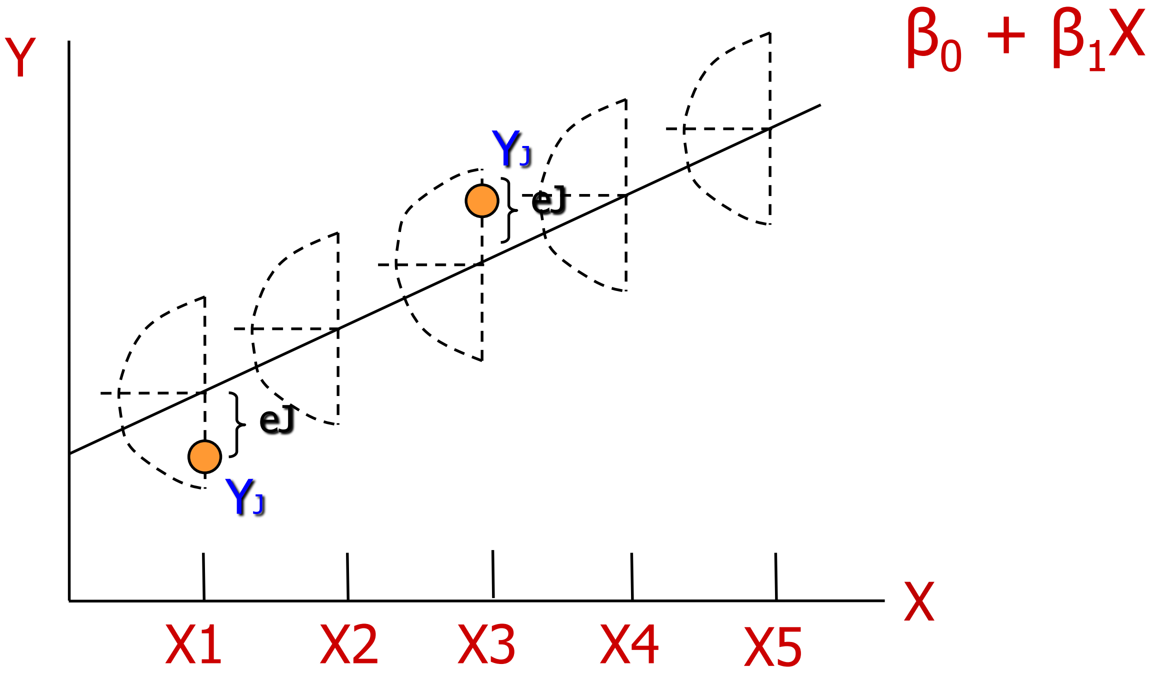
Plot of Attitude with Duration



Which Straight Line Is Best?



Bivariate Regression



Conducting Bivariate Regression Analysis Estimate the Parameters

In most cases, β_0 and β_1 are unknown and are estimated from the sample observations using the equation

$$\hat{Y}_i = a + b\,x_i$$

where $\hat{Y}_i$ is the estimated or predicted value of Y_i, and *a* and *b* are estimators of β_0 and β_1, respectively.

$$b = \frac{COV_{xy}}{S_x^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Conducting Bivariate Regression Analysis
Estimate the Parameters

The intercept, *a*, may then be calculated using:

$$a = \overline{Y} - b\overline{X}$$

For the data in Table 17.1, the estimation of parameters may be illustrated as follows:

$$\sum_{i=1}^{12} X_i Y_i$$

= (10) (6) + (12) (9) + (12) (8) + (4) (3) + (12) (10) + (6) (4)
+ (8) (5) + (2) (2) + (18) (11) + (9) (9) + (17) (10) + (2) (2)
= 917

$\sum_{i=1}^{12} X_i^2$ = 102 + 122 + 122 + 42 + 122 + 62
+ 82 + 22 + 182 + 92 + 172 + 22
= 1350

Conducting Bivariate Regression Analysis Estimate the Parameters

It may be recalled from earlier calculations of the simple correlation that:

$\overline{X}$ = 9.333

$\overline{Y}$ = 6.583

Given *n* = 12, *b* can be calculated as:

$$b = \frac{917 - (12)\,(9.333)\,(6.583)}{1350 - (12)\,(9.333)^2}$$

= 0.5897

a = $\overline{Y}$ - *b* $\overline{X}$

= 6.583 - (0.5897) (9.333)

= 1.0793

Conducting Bivariate Regression Analysis Estimate the Standardized Regression Coefficient

- **Standardization** is the process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1 (Chapter 14).
- When the data are standardized, the intercept assumes a value of 0.
- The term **beta coefficient** or **beta weight** is used to denote the standardized regression coefficient.

$$B_{yx} = B_{xy} = r_{xy}$$

- There is a simple relationship between the standardized and non-standardized regression coefficients:

$$B_{yx} = b_{yx}\,(S_x / S_y)$$

Conducting Bivariate Regression Analysis Test for Significance

The statistical significance of the linear relationship between *X* and *Y* may be tested by examining the hypotheses:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

A *t* statistic with *n* - 2 degrees of freedom can be used, where $t = \dfrac{b}{SE_b}$

SE_b denotes the standard deviation of *b* and is called the **standard error**.

Conducting Bivariate Regression Analysis Test for Significance

Using a computer program, the regression of attitude on duration of residence, using the data shown in Table 17.1, yielded the results shown in Table 17.2. The intercept, *a*, equals 1.0793, and the slope, *b*, equals 0.5897. Therefore, the estimated equation is:

Attitude ($\hat{Y}$) = 1.0793 + 0.5897 (Duration of residence)

The standard error, or standard deviation of *b* is estimated as 0.07008, and the value of the *t* statistic as *t* = 0.5897/0.0700 = 8.414, with *n* - 2 = 10 degrees of freedom.

From Table 4 in the Statistical Appendix, we see that the critical value of *t* with 10 degrees of freedom and α = 0.05 is 2.228 for a two-tailed test. Since the calculated value of *t* is larger than the critical value, the null hypothesis is rejected.

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

The total variation, SS_y, may be decomposed into the variation accounted for by the regression line, SS_{reg}, and the error or residual variation, SS_{error} or SS_{res}, as follows:

$$SS_y = SS_{reg} + SS_{res}$$

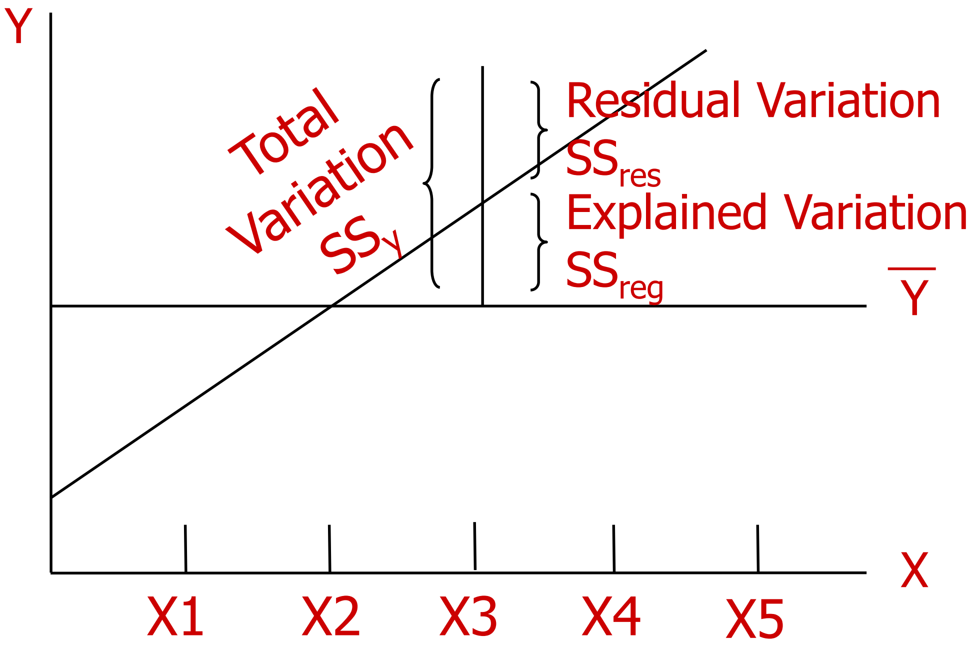
where

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Decomposition of the Total Variation in Bivariate Regression



Conducting Bivariate Regression Analysis
Determine the Strength and Significance of Association

The **strength of association** may then be calculated as follows:

$$r^2 = \frac{SS_{reg}}{SS_y}$$

$$= \frac{SS_y - SS_{res}}{SS_y}$$

To illustrate the calculations of r^2, let us consider again the effect of attitude toward the city on the duration of residence. It may be recalled from earlier calculations of the simple correlation coefficient that:

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

= 120.9168

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

The predicted values ($\hat{Y}$) can be calculated using the regression equation:

Attitude ($\hat{Y}$) = 1.0793 + 0.5897 (Duration of residence)

For the first observation in Table 17.1, this value is:

($\hat{Y}$) = 1.0793 + 0.5897 x 10 = 6.9763.

For each successive observation, the predicted values are, in order,

8.1557, 8.1557, 3.4381, 8.1557, 4.6175, 5.7969, 2.2587, 11.6939, 6.3866, 11.1042, and 2.2587.

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

Therefore,

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$= (6.9763-6.5833)^2 + (8.1557-6.5833)^2$

$+ (8.1557-6.5833)^2 + (3.4381-6.5833)^2$

$+ (8.1557-6.5833)^2 + (4.6175-6.5833)^2$

$+ (5.7969-6.5833)^2 + (2.2587-6.5833)^2$

$+ (11.6939 -6.5833)^2 + (6.3866-6.5833)^2$
$+ (11.1042 -6.5833)^2 + (2.2587-6.5833)^2$

$=0.1544 + 2.4724 + 2.4724 + 9.8922 + 2.4724$

$+ 3.8643 + 0.6184 + 18.7021 + 26.1182$

$+ 0.0387 + 20.4385 + 18.7021$

$= 105.9524$

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (6-6.9763)^2 + (9-8.1557)^2 + (8-8.1557)^2$$
$$+ (3-3.4381)^2 + (10-8.1557)^2 + (4-4.6175)^2$$
$$+ (5-5.7969)^2 + (2-2.2587)^2 + (11-11.6939)^2$$
$$+ (9-6.3866)^2 + (10-11.1042)^2 + (2-2.2587)^2$$

$$= 14.9644$$

It can be seen that $SS_y = SS_{reg} + SS_{res}$. Furthermore,

$$r^2 = SS_{reg} / SS_y$$
$$= 105.9524/120.9168$$
$$= 0.8762$$

Conducting Bivariate Regression Analysis
Determine the Strength and Significance of Association

Another, equivalent test for examining the significance of the linear relationship between *X* and *Y* (significance of *b*) is the test for the significance of the coefficient of determination. The hypotheses in this case are:

$$H_0\text{: } R^2_{pop} = 0$$

$$H_1\text{: } R^2_{pop} > 0$$

Conducting Bivariate Regression Analysis Determine the Strength and Significance of Association

The appropriate test statistic is the *F* statistic:

$$F = \frac{SS_{reg}}{SS_{res}/(n\text{-}2)}$$

which has an *F* distribution with 1 and *n* - 2 degrees of freedom. The *F* test is a generalized form of the *t* test (see Chapter 15). If a random variable is *t* distributed with *n* degrees of freedom, then t^2 is *F* distributed with 1 and *n* degrees of freedom. Hence, the *F* test for testing the significance of the coefficient of determination is equivalent to testing the following hypotheses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

or

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Conducting Bivariate Regression Analysis
Determine the Strength and Significance of Association

From Table 17.2, it can be seen that:

r^2 = 105.9522/(105.9522 + 14.9644)

= 0.8762

Which is the same as the value calculated earlier. The value of the *F* statistic is:

F = 105.9522/(14.9644/10)
= 70.8027

with 1 and 10 degrees of freedom. The calculated *F* statistic exceeds the critical value of 4.96 determined from Table 5 in the Statistical Appendix. Therefore, the relationship is significant at α = 0.05, corroborating the results of the *t* test.

Bivariate Regression

Multiple *R*	0.93608
R^2	0.87624
Adjusted R^2	0.86387
Standard Error	1.22329

ANALYSIS OF VARIANCE

	df	Sum of Squares	Mean Square
Regression	1	105.95222	105.95222
Residual	10	14.96444	1.49644
F = 70.80266		Significance of F = 0.0000	

VARIABLES IN THE EQUATION

Variable	b	SE_b	Beta (ß)	T	Significance of T
Duration	0.58972	0.07008	0.93608	8.414	0.0000
(Constant)	1.07932	0.74335		1.452	0.1772

Conducting Bivariate Regression Analysis Check Prediction Accuracy

To estimate the accuracy of predicted values, $\hat{Y}$, it is useful to calculate the standard error of estimate, SEE.

$$SEE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

or

$$SEE = \sqrt{\frac{SS_{res}}{n-2}}$$

or more generally, if there are *k* independent variables,

$$SEE = \sqrt{\frac{SS_{res}}{n-k-1}}$$

For the data given in Table 17.2, the SEE is estimated as follows:

$$\mathrm{SEE} = \sqrt{14.9644/(12\text{-}2)}$$

= 1.22329

Assumptions

- The error term is **normally** distributed. For each fixed value of *X*, the distribution of *Y* is normal.
- The means of all these normal distributions of *Y*, given *X*, lie on a **straight line** with slope *b*.
- The mean of the error term is **0**.
- The variance of the error term is **constant**. This variance does not depend on the values assumed by *X*.
- The error terms are **uncorrelated**. In other words, the observations have been drawn independently.

Multiple Regression

The general form of the **multiple regression model** is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \mathbf{e}$$

which is estimated by the following equation:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

As before, the coefficient *a* represents the intercept, but the *b*'s are now the partial regression coefficients.

Statistics Associated with Multiple Regression

- **Adjusted *R*^{2}**. *R*^{2}, coefficient of multiple determination, is adjusted for the number of independent variables and the sample size to account for the diminishing returns. After the first few variables, the additional independent variables do not make much contribution.

- **Coefficient of multiple determination**. The strength of association in multiple regression is measured by the square of the multiple correlation coefficient, *R*^{2}, which is also called the coefficient of multiple determination.

- ***F* test**. The *F* test is used to test the null hypothesis that the coefficient of multiple determination in the population, R^2_{pop}, is zero. This is equivalent to testing the null hypothesis. The test statistic has an *F* distribution with *k* and (n - k - 1) degrees of freedom.

Statistics Associated with Multiple Regression

- **Partial *F* test**. The significance of a partial regression coefficient,β_i , of X_i may be tested using an incremental *F* statistic. The incremental *F* statistic is based on the increment in the explained sum of squares resulting from the addition of the independent variable X_i to the regression equation after all the other independent variables have been included.

- **Partial regression coefficient**. The partial regression coefficient, b_1, denotes the change in the predicted value, $\hat{Y}$, per unit change in X_1 when the other independent variables, X_2 to X_k, are held constant.

Conducting Multiple Regression Analysis Partial Regression Coefficients

To understand the meaning of a partial regression coefficient, let us consider a case in which there are two independent variables, so that:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

- First, note that the relative magnitude of the partial regression coefficient of an independent variable is, in general, different from that of its bivariate regression coefficient.
- The interpretation of the partial regression coefficient, b_1, is that it represents the expected change in Y when X_1 is changed by one unit but X_2 is held constant or otherwise controlled. Likewise, b_2 represents the expected change in Y for a unit change in X_2, when X_1 is held constant. Thus, calling b_1 and b_2 partial regression coefficients is appropriate.

Conducting Multiple Regression Analysis
Partial Regression Coefficients

- It can also be seen that the combined effects of X_1 and X_2 on Y are additive. In other words, if X_1 and X_2 are each changed by one unit, the expected change in Y would be (b_1+b_2).

- Suppose one was to remove the effect of X_2 from X_1. This could be done by running a regression of X_1 on X_2. In other words, one would estimate the equation $\hat{X}_1 = a + b\,X_2$ and calculate the residual $X_r = (X_1 - \hat{X}_1)$. The **partial regression coefficient,** b_1, is then equal to the bivariate regression coefficient, b_r , obtained from the equation $\hat{Y} = a + b_r\,X_r$.

Conducting Multiple Regression Analysis Partial Regression Coefficients

- Extension to the case of *k* variables is straightforward. The partial regression coefficient, b_1, represents the expected change in *Y* when X_1 is changed by one unit and X_2 through X_k are held constant. It can also be interpreted as the bivariate regression coefficient, *b*, for the regression of *Y* on the residuals of X_1, when the effect of X_2 through X_k has been removed from X_1.
- The relationship of the standardized to the non-standardized coefficients remains the same as before:

  $B_1 = b_1 \, (S_{x1}/Sy)$

  $B_k = b_k \, (S_{xk}/S_y)$

The estimated regression equation is:

$(\hat{Y}) = 0.33732 + 0.48108 \, X_1 + 0.28865 \, X_2$

or

Attitude = 0.33732 + 0.48108 (Duration) + 0.28865 (Importance)

Multiple Regression

Multiple R	0.97210
R^2	0.94498
Adjusted R^2	0.93276
Standard Error	0.85974

ANALYSIS OF VARIANCE

	df	Sum of Squares	Mean Square
Regression	2	114.26425	57.13213
Residual	9	6.65241	0.73916

F = 77.29364 Significance of F = 0.0000

VARIABLES IN THE EQUATION

Variable	b	SE_b	Beta (ß)	T	Significance of T
IMPORTANCE	0.28865	0.08608	0.31382	3.353	0.0085
DURATION	0.48108	0.05895	0.76363	8.160	0.0000
(Constant)	0.33732	0.56736		0.595	0.5668

Conducting Multiple Regression Analysis Strength of Association

$$SS_y = SS_{reg} + SS_{res}$$

where

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Conducting Multiple Regression Analysis
Strength of Association

The strength of association is measured by the square of the multiple correlation coefficient, *R*², which is also called the **coefficient of multiple determination**.

$$R^2 = \frac{SS_{reg}}{SS_y}$$

*R*² is adjusted for the number of independent variables and the sample size by using the following formula:

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

Conducting Multiple Regression Analysis Significance Testing

$H_0 : R^2_{pop} = 0$

This is equivalent to the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = . . . = \beta_k = 0$$

The overall test can be conducted by using an *F* statistic:

$$F = \frac{SS_{reg}/k}{SS_{res}/(n - k - 1)}$$

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

which has an *F* distribution with *k* and (*n* - *k* -1) degrees of freedom.

Conducting Multiple Regression Analysis Significance Testing

Testing for the significance of the β_i's can be done in a manner similar to that in the bivariate case by using *t* tests. The significance of the partial coefficient for importance attached to weather may be tested by the following equation:

$$t = \frac{b}{SE_b}$$

which has a *t* distribution with *n* - *k* -1 degrees of freedom.

Conducting Multiple Regression Analysis Examination of Residuals

- A **residual** is the difference between the observed value of Y_i and the value predicted by the regression equation $\hat{Y}_i$.

- Scattergrams of the residuals, in which the residuals are plotted against the predicted values, $\hat{Y}_i$, time, or predictor variables, provide useful insights in examining the appropriateness of the underlying assumptions and regression model fit.

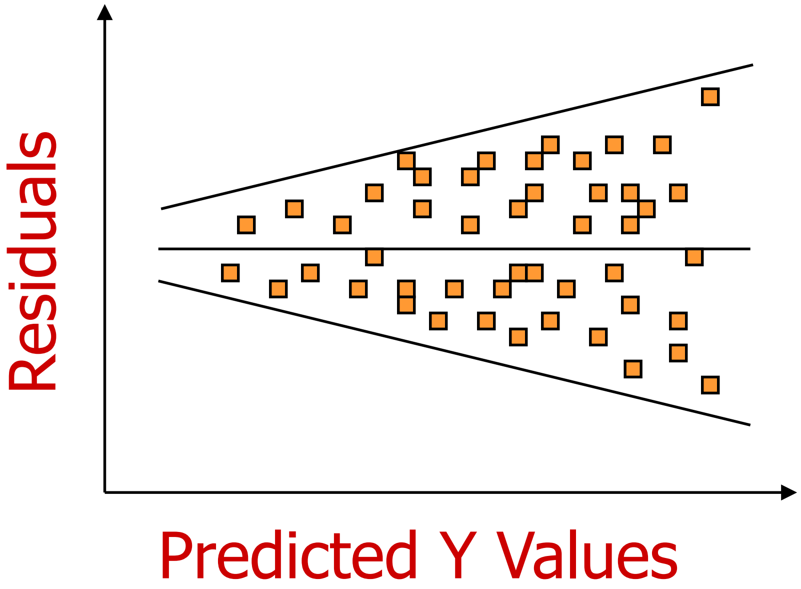
- The assumption of a normally distributed error term can be examined by constructing a histogram of the residuals.

- The assumption of constant variance of the error term can be examined by plotting the residuals against the predicted values of the dependent variable, $\hat{Y}_i$.

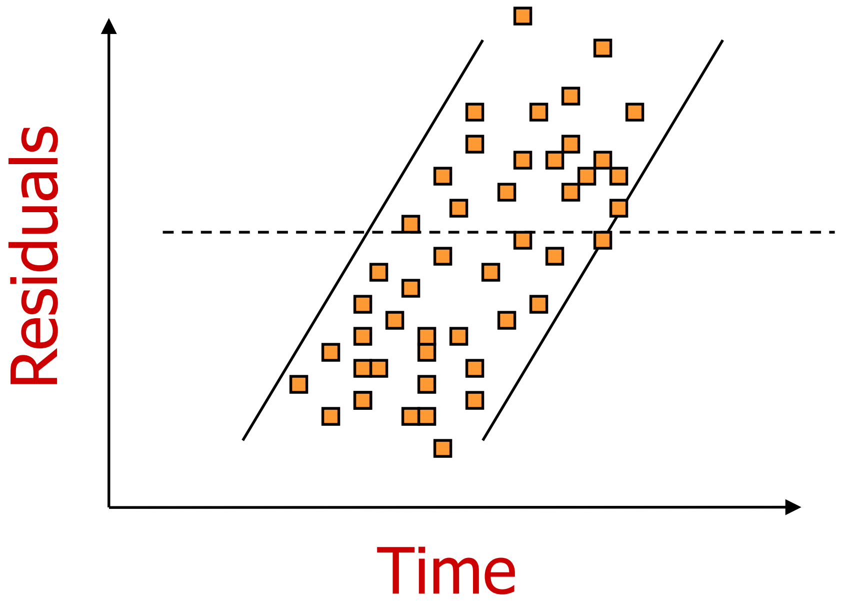
Conducting Multiple Regression Analysis
Examination of Residuals

- A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated.
- Plotting the residuals against the independent variables provides evidence of the appropriateness or inappropriateness of using a linear model. Again, the plot should result in a random pattern.
- To examine whether any additional variables should be included in the regression equation, one could run a regression of the residuals on the proposed variables.
- If an examination of the residuals indicates that the assumptions underlying linear regression are not met, the researcher can transform the variables in an attempt to satisfy the assumptions.

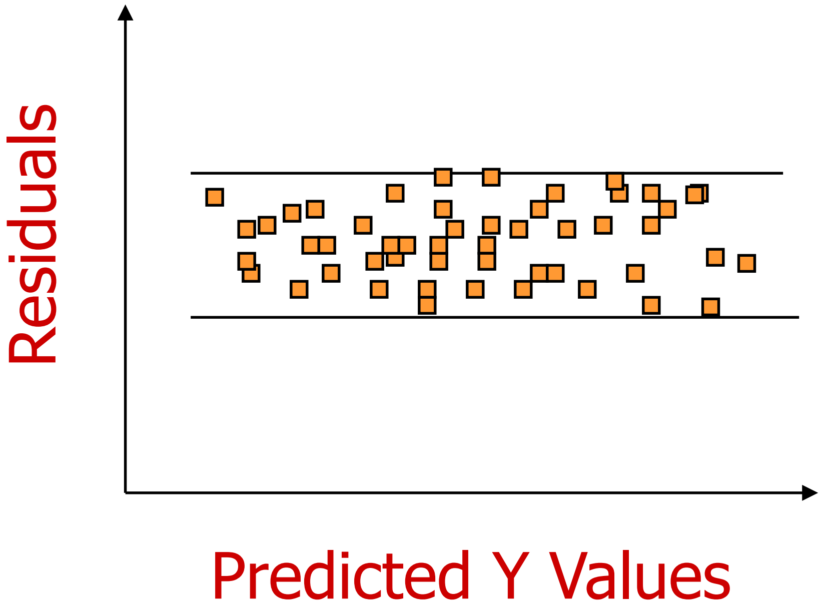
Residual Plot Indicating that Variance Is Not Constant



Residual Plot Indicating a Linear Relationship Between Residuals and Time



Plot of Residuals Indicating that a Fitted Model Is Appropriate



Stepwise Regression

The purpose of **stepwise regression** is to select, from a large number of predictor variables, a small subset of variables that account for most of the variation in the dependent or criterion variable. In this procedure, the predictor variables enter or are removed from the regression equation one at a time. There are several approaches to stepwise regression.

- **Forward inclusion**. Initially, there are no predictor variables in the regression equation. Predictor variables are entered one at a time, only if they meet certain criteria specified in terms of *F* ratio. The order in which the variables are included is based on the contribution to the explained variance.
- **Backward elimination**. Initially, all the predictor variables are included in the regression equation. Predictors are then removed one at a time based on the *F* ratio for removal.
- **Stepwise solution**. Forward inclusion is combined with the removal of predictors that no longer meet the specified criterion at each step.