8 Linear Regression

- In many business research situations the path to decision making lies in understanding the relationship between two or more variables.

- For example, can the price of airline stock be predicted using a variable like the cost of oil? or how strong a relationship is there between a company's sales figures and their advertising budget?

- ***Regression Analysis*** *is the process of constructing a mathematical model or function that can be used to predict one variable using another variable (simple linear regression) or variables (multiple linear regression).*

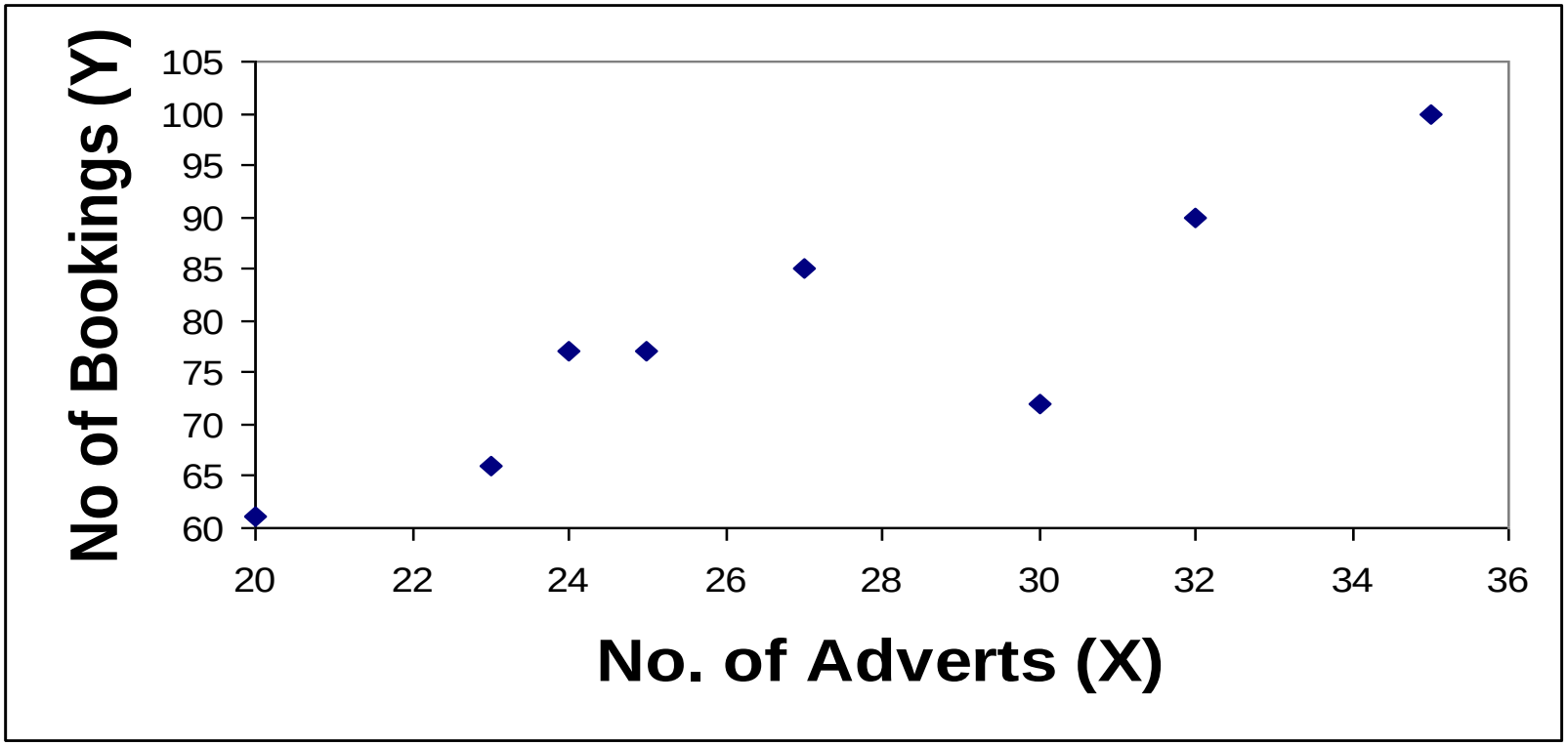
- ***Example (Fedbus)***

Fedbus have been using radio advertisements to advertise their weekend excursions for the past 8 months.

The managing director has asked the sales team to assess the effect of these ads on the number of bookings

Month	No. of Adverts (X)	No. of bookings (Y)
1	20	61
2	24	77
3	30	72
4	23	66
5	27	85
6	25	77
7	32	90
8	35	100

Figure A



The scatter plot can give some indication on the nature and strength of the relationship between X and Y, but clearly we need a more precise technique to quantify this relationship.

2.1.2 Determining the Regression Equation

Try to imagine a line passing through the points in scatter plot in Figure A. The Regression technique attempts to fit a straight line to represent these points.

The equation of the Regression line takes the form

$$Y = b_0 + b_1X$$

where

b_0 = Intercept and b_1 = Slope of the line

b_0 and b_1 are calculated using the *Method of Least Squares*.

Fedbus example

$$Y = 18.67 + 2.22X$$

No. of Bookings = 18.67 + 2.22 x (No. of Adverts)

2.1.3 What does the Regression Equation Mean?

- A Regression Equation can be seen as a model of the relationship between Y and X or can be extended and used for prediction purposes.

- In the Fedbus example the following relationship between bookings(Y) and adverts(X) was estimated

$$Y = 18.67 + 2.22X$$

- The slope of the line b_1 tells us that if X increases by one unit then Y will increase by 2.22 units.

- We can also use the equation for prediction purposes

- It is important to note that regression cannot be used reliably for extrapolation

2.2.1 Multiple Linear Regression

- Simple Linear Regression dealt with the relationship between One Dependent Variable Y and One Independent Variable X.

- Rarely will the variations of a business variable depend on only one factor.

- **Multiple Regression Model**

  Given Y (Dependent Variable) and $X_1, X_2, \ldots X_n$ (a set of n Independent Variables) then

  $$Y = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots + b_nX_n$$

  b_i = Regression Coefficient for Xi.

Example (Automobile Battery Sales)

- Before entering a marketplace it is common for a company to examine the sales potential in that area. One approach is to develop a Regression Model of Sales using data from other regions.
- In this example An Automobile Battery Company have figures on Sales from thirty Sales Areas.

Dependent Variable (Y): Battery Sales

Independent Variables:

Number of Registered Vehicles (X_1)

Average Weekly Earnings (X_2)

Advertising Expenditure (X_3)

Variables:

Y = Number of Batteries Sold X_1 = Registered Vehicles

X_2 = Advertising Expenditure X_3 = Weekly Earnings

Model:

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

$Y = -46.6 + .17X_1 + .87X_2 - .03X_3$

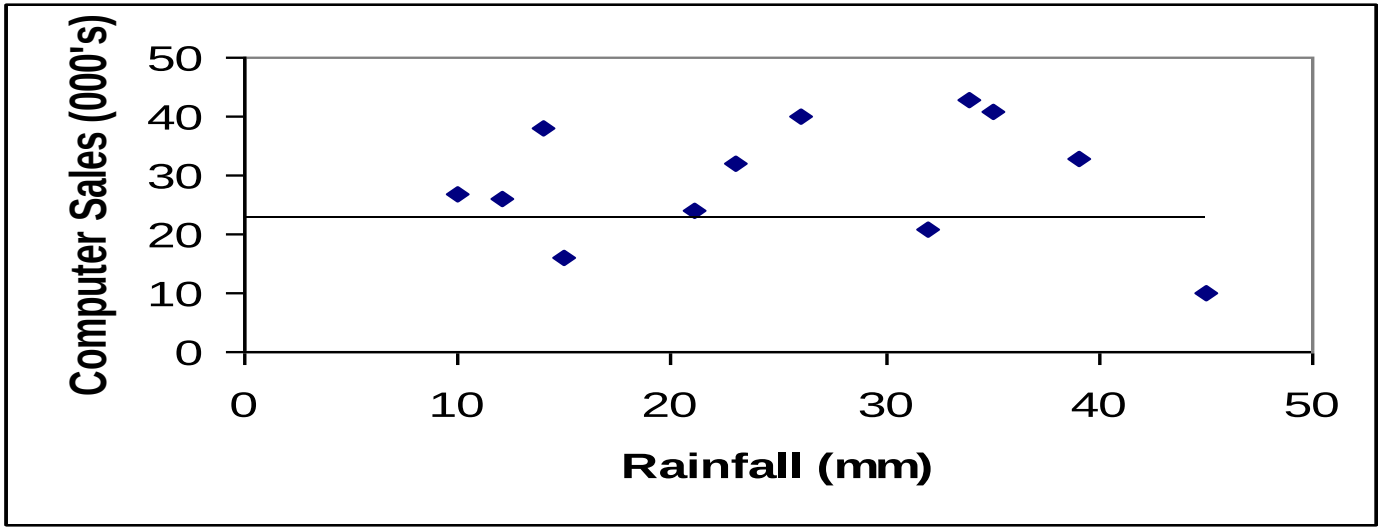
or

Battery Sales = - 46.6
+ .17 x (Vehicle Registrations)
+ .87 x (Advertising Expend.)
- .03 x (Weekly Earnings)

How good is the Regression Model?

Constructing the Regression Model is only the first step in a Regression Analysis. Unfortunately the Regression technique will fit a line to any set of points, even if no strong relationship exists

Here is the relationship between a company's computer sales and rainfall in their region.



We can see that there is no relationship between the two variables, and because of this the Regression line is a very bad fit.

How good is Regression Model?

- So, all Regression models are not necessarily good predictors of the dependent variable Y. This means we need a set of statistical tools to help assess the validity of any model.

(i) Coefficient of determination

(ii) T-test

Battery Example

	B	Sig t (P-Value)
(Registered Vehicles)	.17	.157
(Advertising Expend.)	.87	.000
(Weekly Earnings	-.03	.000

- Weekly Earnings and Advertising Expenditure are the significant independent variables (both p-values are below .05).
- However the test shows that the number of Registered Vehicles is not a significant variable (p-value equal to .15, which is greater than .05)
- Rerun model without Number of Registered Vehicles

2.2.2 Measurements in Regression

- The dependent variable must **always** be a **ratio/interval** variable.

- Independent variables should also be ratio. Many business analysts however tend to relax these assumptions slightly to include ordinal measurements.

- Nominal data can also be used as independent variables in a Regression Analysis using a technique called dummy variables

2.2.3 Analysis of Residuals

- The difference between the an observed value and the value predicted by a regression equation is called a residual.

- Regression Analysis is based on a number of assumptions

  - If any of these assumptions are violated model may be incorrectly estimated.

  - Important therefore to test that the assumptions have not been violated.

- Most important tool for testing assumptions is called residual analysis.

  - Plotting the residuals and looking for patterns in these plots.

  - If there are any patterns in the residuals - one or more of the assumptions may have been violated