

# 4 Statistical Modelling

- In many practical situations we want to identify various types of relationships between variables.
- Sometimes we want to estimate how one variable is related to or affected by some other variables.

## 4.1 Choosing the Right Statistical Test

		Dependent Variable (Y)			
		Categorical		Continuous	
Categorical	1 Variable, 2 Categories, Between-subjects	1 Variable, 2 Categories	1 Variable, >2 Categories	1 Variable	>1 Variable
	1 Variable, 2 Categories, Within-subjects	Chi-square Test		Independent t Test	
	1 Variable, >2 Categories, Between-subjects			Paired t Test	
	1 Variable, >2 Categories, Within-subjects			One-Way ANOVA	One-Way MANOVA
	>1 Variable, All Categorical, Between Subjects	Binomial Logistic Regression with Categorical Predictors	Multinomial Logistic Regression	Repeated Measures ANOVA	Repeated Measures MANOVA
	>1 Variable, All Categorical, Mixed Within- & Between-Subjects			Factorial ANOVA	Factorial MANOVA
	>1 Variable, Mixed Categorical & Continuous			Mixed-Design ANOVA	Mixed-Design MANOVA
Continuous	1 Variable	Binomial Logistic Regression		One-Way ANCOVA	One-Way MANCOVA
	>1 Variable			Simple Linear Regression Multiple Linear Regression	Multivariate Linear Regression

Figure 4.1: Selection of statistical tests based on variable type

- The **dependent variable** (response variable) is the variable that is being measured or tested in an experiment.
- The **independent variable** (explanatory) is the variable the experimenter manipulates or changes, and is assumed to have a direct effect on the **dependent variable**.
- **Between-subjects** (or between-groups) study design: different people test each condition, so that each person is only exposed to one condition.
- **Within-subjects** (or repeated-measures) study design: the same person tests all the conditions.

## 4.2 Regression analysis

- **Regression analysis** is a statistical technique for investigating and **modeling the relationship between variables**.
- There are numerous applications of regression in many fields.

### *Examples*

- A production company may need to determine how its sales related to advertising
- How the growth of the bacteria is related to moisture level of the environment.
- The relationship between blood pressure and the age of a person.
- the relationship between transaction time and transaction amount in fraud detection.
- Usually, the first step in regression analysis is to construct a scatter plot (or scatter matrix).
- Graphing the data in a scatter plot yields preliminary information about the *shape* and *spread* of the data.

### 4.2.1 Simple Linear Regression

- The most elementary regression model is called **simple linear regression**.
- It is also known as *bivariate linear regression*, which means that it involves only two variables.
- One variable is predicted by another variable.
- The variable to be predicted is called the *dependent variable* and is denoted by  $y$ .
- The *predictor* is called the *independent variable* or *explanatory variable* and is denoted by  $x$
- In simple *linear* regression analysis, only a straight-line relationship between two variables is examined.
- Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models.

#### 4.2.1.1 Determining the equation of the regression line

- The first step in determining the equation of the regression line that passes through the sample data is to establish the equation's form.
- In mathematics, the equation of a line can be written as

$$y = mx + c$$

where:  $m$  = slope of the line

$c$  =  $y$  intercept of the line.

- In statistics, the slope-intercept form of the equation of the regression line through the population points is:

$$\hat{y} = \beta_0 + \beta_1 x$$

where:  $\hat{y}$  = the predicted value of  $y$

$\beta_0$  = the population  $y$  intercept

$\beta_1$  = the population slope.

- For any specific dependent variable value,  $y_i$ :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where:  $x_i$  = the value of the independent variable for the  $i$ th value

$y_i$  = the value of the dependent variable for the  $i$ th value

$\beta_0$  = the population  $y$  intercept

$\beta_1$  = the population slope

$\epsilon_i$  = the error of prediction for the  $i$ th value.

- Unless the points being fitted by the regression equation are in perfect alignment, the regression line will miss at least some of the points.
- In the above equation,  $\epsilon_i$  represents the error of the regression line in fitting these points. If a point is on the regression line,  $\epsilon_i = 0$

#### 4.2.1.2 Deterministic models vs probabilistic models

- These mathematical models can be either deterministic models or probabilistic models.

##### Deterministic models

- Deterministic models are *mathematical models that produces an exact output for a given input*

- For example, consider the equation of a regression line is:

$$\hat{y} = 1.68 + 2.40x$$

For a value of  $x = 5$ , the exact predicted value of  $y$  is

$$\hat{y} = 1.68 + 2.40 \times 5 = 13.68$$

- However, the most of the time the values of  $y$  will not equal exactly the values yields by the equations.
- Random error will occur in the prediction of the  $y$  values for values of  $x$ , because it is likely that the variable  $x$  does not explain all the variability of the variable  $y$ .

### *Example*

- Suppose we want to predict the sales volume ( $y$ ) for a mobile phone company through regression analysis by using the annual amount of advertising (in Rupees) ( $x$ ) as the predictor.
- Although sales are often related to advertising, there can be other factors related to sales that are not accounted for by the amount of advertising.
- Therefore, a regression model to predict sales volume by the amount of advertising probably involves some error.
- For this reason, in regression, we present the general model as a probabilistic model.

### **Probabilistic models**

- A probabilistic model is *one that includes an error term that allows for the  $y$  values to vary for any given value of  $x$* .
- The deterministic regression model is

$$y = \beta_0 + \beta_1 x$$

- The probabilistic regression model is

$$y = \beta_0 + \beta_1 x + \epsilon.$$

- $\beta_0 + \beta_1 x$  is the deterministic portion of the probabilistic model,  $\beta_0 + \beta_1 x + \epsilon$ .
- In deterministic model, all points are assumed to be on the line and in all cases  $\epsilon$  is zero.

#### 4.2.1.3 Least squares estimation of the parameters

- The parameters  $\beta_0$  and  $\beta_1$  are unknown and need to be estimated using *sample data*.
- The equation of the regression line contains the sample  $y$  intercept,  $\hat{\beta}_0$ , and the sample slope,  $\hat{\beta}_1$ .

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where:  $\hat{y}$  = the predicted value of  $y$

$\hat{\beta}_0$  = the sample  $y$  intercept

$\hat{\beta}_1$  = the sample slope

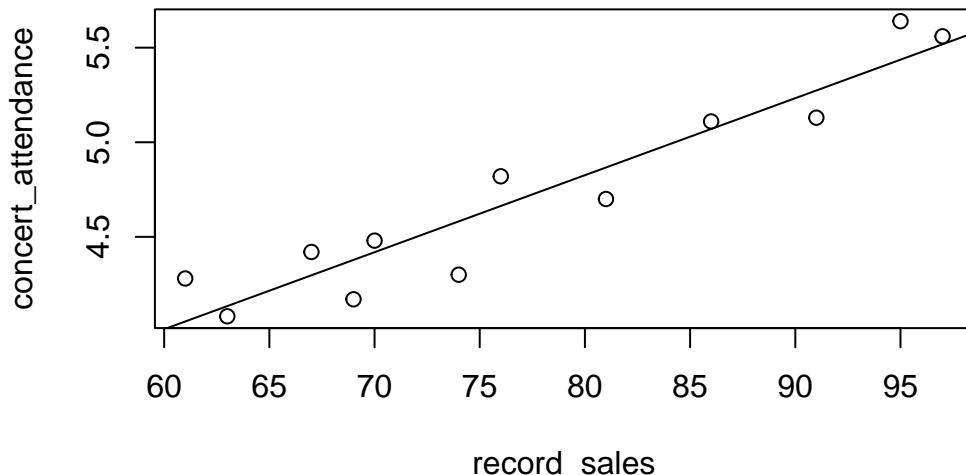
- To determine the equation of the regression line for a sample of data, the researcher must determine the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .
- This process is sometimes referred to as **least squares analysis**.
- Least squares analysis is a process whereby a regression model is developed by *producing the minimum sum of the squared error values*.
- The least squares regression line is the **regression line that results in the smallest sum of errors squared**.

*Example*

The data in the table are the twelve observations corresponding to concert attendance (in thousand) and total worldwide CD sales by the performing artist (or band) in the previous year (also in thousand).

- Draw a scatter diagram for the data and determine by inspection if there exists an approximate linear relationship between  $X$  and  $Y$ .

Number of CD sales ('000) $x$	Concert attendance ('000) $y$	$x^2$	$xy$	Predicted value $\hat{y}$	Residual $y - \hat{y}$
61	4.280				
63	4.080				
67	4.420				
69	4.170				
70	4.480				
74	4.300				
76	4.820				
81	4.700				
86	5.110				
91	5.130				
95	5.640				
97	5.560				



```
integer(0)
```

- Use the data to develop a regression model to predict concert attendance by CD sales.

*Least squares analysis cntd...*

*Least squares analysis cntd...*

#### 4.2.2 Residual analysis

- Each difference between the actual  $y$  values and the predicted  $\hat{y}$  values is the error of the regression line at a given point,  $y - \hat{y}$ , and is referred to as the **residual**.
- Except for rounding error, the sum of the residuals is approximately zero
- The analysis of residuals plays an important role in validating the regression model.
- Residual analysis plots are a very useful tool for assessing aspects of accuracy of a linear regression model on a particular dataset and testing that the attributes of a dataset meet the requirements for linear regression.
- The following are the assumptions of simple linear regression analysis
  1. The model is linear.
  2. The error terms have constant variances. (*The assumption of constant variance is called homoskedasticity. If the error variances are not constant, it is called heteroskedasticity.*)
  3. The error terms are independent.
  4. The error terms are normally distributed.
- Residual plots can be used to show if there are problems with the dataset and the model produced that need to be considered in looking at the validity of the model.

##### Residual plot

- A particular method for studying the behaviour of residuals is the residual plot.
- The residual plot is a type of graph in which the residuals for a particular regression model are plotted along with their associated value of  $x$  and an ordered pair  $(x, y - \hat{y})$ .
- Information about how well the regression assumptions are met by the particular regression model can be obtained by examining the plots.
- Residual plots are more meaningful with larger sample sizes.
- For small sample sizes, residual plot analyses can be problematic and subject to over-interpretation.

##### Normal Q–Q (quantile-quantile) Plot

- Residuals should be normally distributed.
- A normal quantile plot (also known as a quantile-quantile plot or QQ plot) is a graphical way of checking whether your data are normally distributed.
- If residuals follow close to a straight line on this plot, it is a good indication they are normally distributed.

#### 4.2.3 Coefficient of determination

- A widely used measure of fit for regression models is the **coefficient of determination** ( $R^2$ )

- It is a statistical measure of **how well the regression predictions approximate the real data points.**
- In statistical point of view, the coefficient of determination is the proportion of variability of the dependent variable ( $y$ ) accounted for, or explained by, the independent variable ( $x$ )
- The coefficient of determination ranges from 0 to 1.
- An  $R^2$  of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of  $y$  by  $x$ .
- An  $R^2$  of 1 means perfect prediction of  $y$  by  $x$  and that 100% of the variability of  $y$  is accounted for by  $x$  (i.e. an  $R^2$  of 1 indicates that the regression predictions perfectly fit the data.)

#### 4.2.4 Relationship between $r$ and $R^2$

- Let  $r$  be the coefficient of correlation.
- The coefficient of determination ( $R^2$ ) is the square of the coefficient of correlation

$$r = \sqrt{R^2}$$

- The researcher must examine the sign of the slope of the regression line to determine whether a positive or negative relationship exists between the variables and then assign the same sign to the correlation value.