Chapter 1

Introduction

What do we mean by Statistical Modelling and a Statistical Model?

Think back to Introduction to Statistics or Statistical Methods modules. There were statements like: "Y_1, Y_2,. . .,Y_n are independent and identically distributed normal random variables with mean μ and variance σ^2". Another way of writing this is

$$Y_i = \mu + \varepsilon_i \qquad i = 1, 2, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and are independent. We wanted to estimate μ, which can be done by using $\bar{Y}$, or to test a hypothesis such as $H_0 : \mu = \mu_0$.

This statistical model has two components, a part which tells us about the average behavior of Y, which is constant, and a random part.

In Statistical Modeling I course we are interested in models where the mean depends on values of other variables. In the simplest case, we have a response variable Y and one explanatory variable X. Then μ depends on the value of X, say x_i, and we may write $\mu_i = \beta_0 + \beta_1 x_i$, where β_0 and β_1 are some unknown constant parameters.

In practice, we start with a real life problem for which we have some data. We think of a statistical model as a mathematical representation of the variables we have measured. This model usually involves some parameters. We may then try to estimate the values of these parameters or to test hypotheses about them. We may wish to use the model to predict what would happen in the future in a similar situation. In order to test hypotheses or to make predictions we usually have to make some assumptions. Part of the modelling process is to test these assumptions. Having found an adequate model we must compare its predictions

with reality to check that it gives reasonable answers.

We can illustrate these ideas using a simple example. Suppose that we are interested in some items, widgets say, which are manufactured in batches. The size of the batch and the time to make the batch in man hours are recorded, see Table 1.1.

x (batch size)	y (man-hours)
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

Table 1.1: Data on batch size and time to make each batch

We begin by plotting the data to see what sort of relationship might hold.

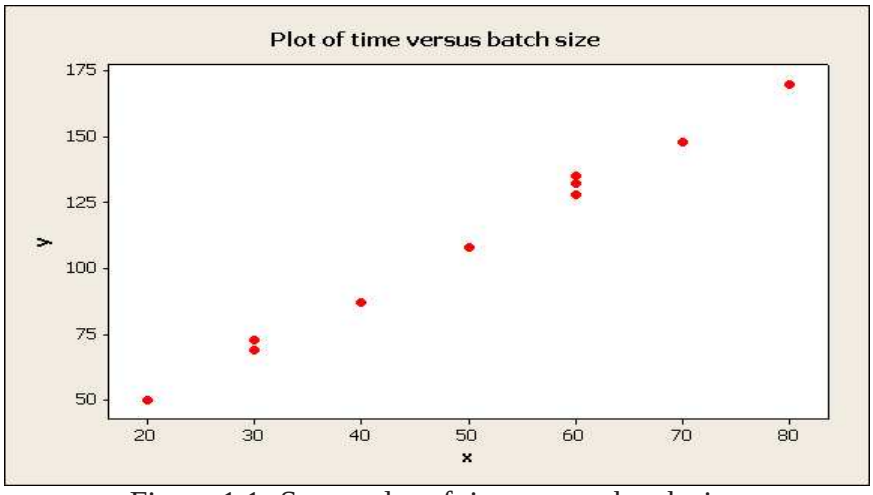

Figure 1.1: Scatterplot of time versus batch size.

MINITAB

Graph → Scatterplot Use Labels... to add title.

From this plot, Figure 1.1, it seems that a straight line relationship is a good representation of the data although it is not an exact relationship. Using MINITAB we can fit this model and obtain the fitted line plot, Figure 1.2.

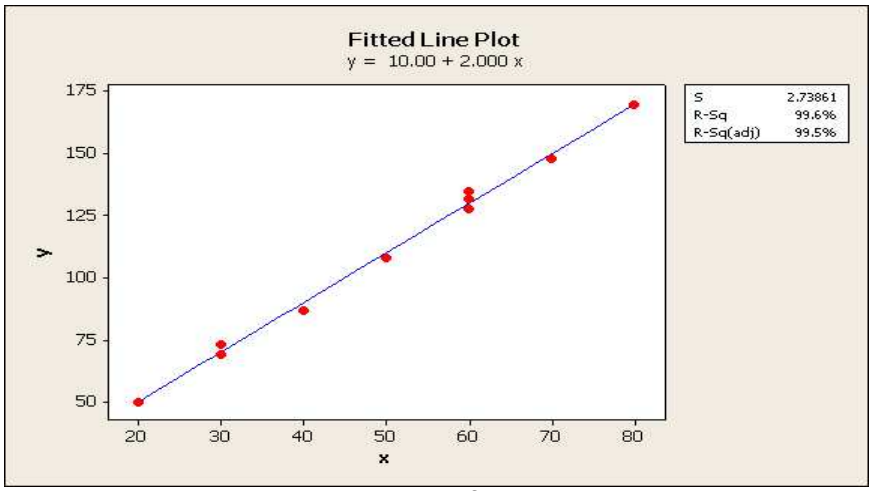

Figure 1.2: Fitted line plot of time versus batch size.

MINITAB

Stat → Regression → Fitted Line Plot...

The fitted line is $\widehat{y} = 10 + 2x$. One interpretation of this is that on average it takes 10 hours to set up the machinery to make widgets and then it takes 2 hours to make each widget.

But before we come to this conclusion we should check that our data satisfy the assumptions of the statistical model. One way to do this is to look at residual plots, as in Figure1.3. We shall discuss these later in the course and in the practicals but here we see that there is no apparent reason to doubt our model. In fact, for small data sets histograms do not represent the distribution well. It is better to examine the Normal Probability Plot.

MINITAB

Stat → Regression → Fitted Line Plot... Graphs... ⊙ Four in one

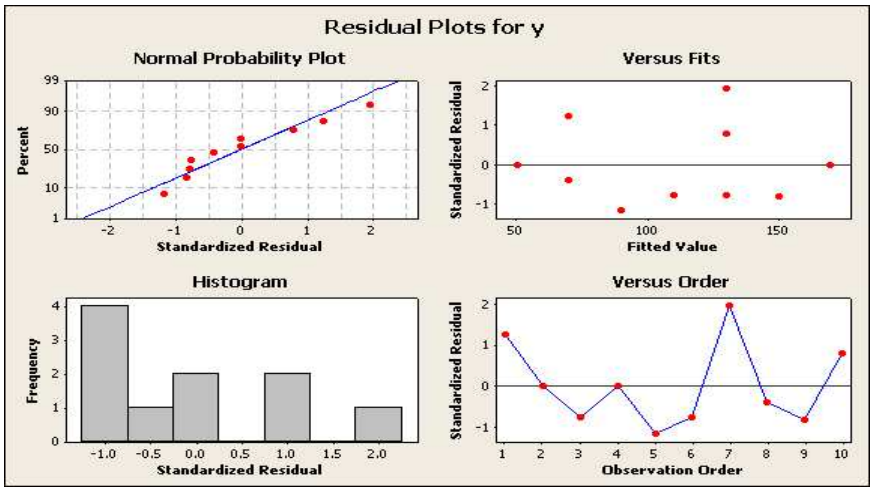

Figure 1.3: Residual plots.

Statistical modelling is iterative. We think of a model we believe will fit the data. We fit it and then check the model. If it is OK we use the model to explain what is happening or to predict what may happen. Note that we should be very wary of making predictions far outside of the x values which are used to fit the model.

In general different techniques are needed depending on whether the explanatory variables are qualitative or quantitative and the random response variable is discrete or continues.

In Statistical Modelling I we will mostly study continuous Y with quantitative $X_1, X_2, \ldots, X_p$. In Statistical Modelling II you would study continuous Y with qualitative $X_1, X_2, \ldots, X_p$. SMI and SMII use ***Linear Models***.

In more advanced courses you can study both models with a mixture of quantitative and qualitative explanatory variables and also discrete Y where we no longer assume errors are normally distributed.

In Time Series we relax the assumption that errors are independent or uncorrelated.

Chapter 2

Simple Linear Regression

2.1 The Model

We start with the simplest situation where we have one response variable Y and one explanatory variable X.

In many practical situations we deal with an explanatory variable X that can be controlled and a response variable Y which can be observed. We want to estimate or to predict the mean value of Y for given values of X working from a sample on n pairs of observations

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Example 2.1. Sparrow's wings.
An ornithologist is interested in the relationship of the wing length and age of sparrows. Data were collected of 13 sparrows of known age, as follows. The last three columns show partial calculations needed for fitting a regression line.

Readings of wing's length may vary for different birds of the same age. Time, X, is known exactly and it is not random, but we may assume that Y is random, so that repeated observations of Y for the same values of X may vary.

A useful initial stage of modelling is to plot the data. Figure 2.1 shows the plot of the sparrow wing's length against sparrow's age.

The plot suggests that the wing length and age might be linearly related, although we would not expect the wing's length increasing linearly over a long period of

time. In this example the linear relationship can be considered for some short growth time only.

x_i [days]	y_i [cm]	$x_i y_i$	x_i^2	y_i^2
$x_1 = 3$	$y_1 = 1.4$	4.2	9	1.96
$x_2 = 3$	$y_2 = 1.5$	4.5	9	2.25
$x_3 = 5$	$y_3 = 2.2$	11.0	25	4.84
$x_4 = 6$	$y_4 = 2.4$	14.4	36	5.76
$x_5 = 8$	$y_5 = 2.8$	22.4	64	7.84
$x_6 = 8$	$y_6 = 3.2$	25.6	64	10.24
$x_7 = 10$	$y_7 = 3.2$	32.0	100	10.24
$x_8 = 11$	$y_8 = 3.9$	42.9	121	15.21
$x_9 = 12$	$y_9 = 4.1$	49.2	144	16.81
$x_{10} = 13$	$y_{10} = 4.7$	65.8	169	22.09
$x_{11} = 14$	$y_{11} = 4.5$	67.5	196	20.25
$x_{12} = 15$	$y_{12} = 5.2$	83.2	225	27.04
$x_{13} = 16$	$y_{13} = 5.0$	80.0	256	25.00
$\sum x_i = 124$	$\sum y_i = 44.1$	$\sum x_i y_i = 488.3$	$\sum x_i^2 = 1418$	$\sum y_i^2 = 169.53$

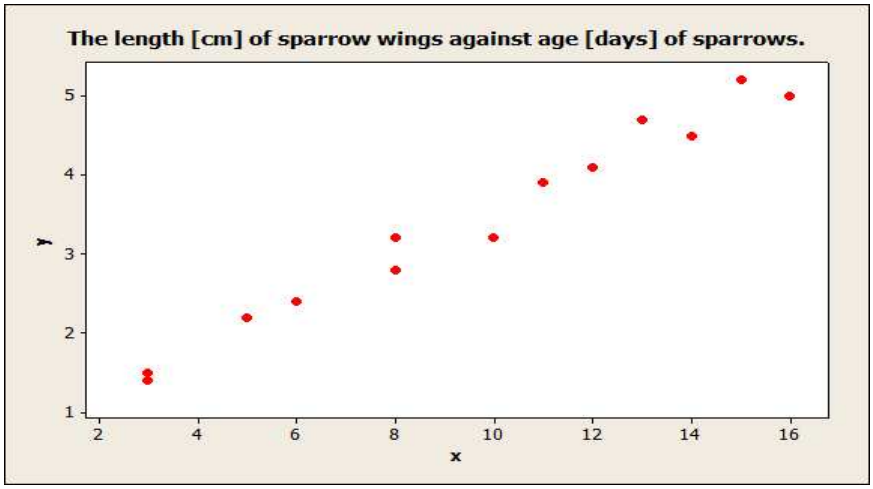

Figure 2.1: Plot of the length of sparrow wings against age of sparrows.

Other types of function could also describe the relationship well, for example a quadratic polynomial with a very small second order coefficient. However, it is better to use the simplest model which describes the relationship well. This is called ***the principle of parsimony***.

What does it mean “to describe the relationship well”?

It means to represent well the expected shape and also the variability of the response Y at each value of the explanatory variable X. We will be working on this problem throughout the course.

We can write

$$Y_i = \mathrm{E}(Y|X = x_i) + \varepsilon_i, \text{ where } \varepsilon_i \text{ is a random variable, } i = 1, 2, \ldots, n.$$

Hence, if the expected relationship is linear, we have

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } i = 1, 2, \ldots, n.$$

We call ε_i a ***random error***. Standard assumptions about the error are

1. $\mathrm{E}(\varepsilon_i) = 0$ for all $i = 1, 2, \ldots, n$,
2. $\mathrm{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$,
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i, j = 1, 2, \ldots, n, i \neq j$.

The errors are often called ***departures*** from the mean. The error ε_i is a random variable, hence Y_i is a random variable too and the assumptions can be rewritten as

1. $\mathrm{E}(Y|X = x_i) = \mu_i = \beta_0 + \beta_1 x_i$ for all $i = 1, \ldots, n$,
2. $\mathrm{var}(Y|X = x_i) = \sigma^2$ for all $i = 1, \ldots, n$,
3. $\mathrm{cov}(Y|X = x_i, Y|X = x_j) = 0$ for all $i, j = 1, \ldots, n, i \neq j$.

It means that the dependence of Y on X is linear and the variance of the response Y at each value of X is constant (does not depend on x_i) and $Y|X = x_i$ and $Y|X = x_j$ are uncorrelated.

Also, it is often assumed that the conditional distribution of Y is normal. Then, due to the assumption (3) on the covariances, the variables Y_i are independent. This is written as

$$Y|X = x_i \underset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2).$$

The graph in Figure 2.2 summarizes all the model assumptions.

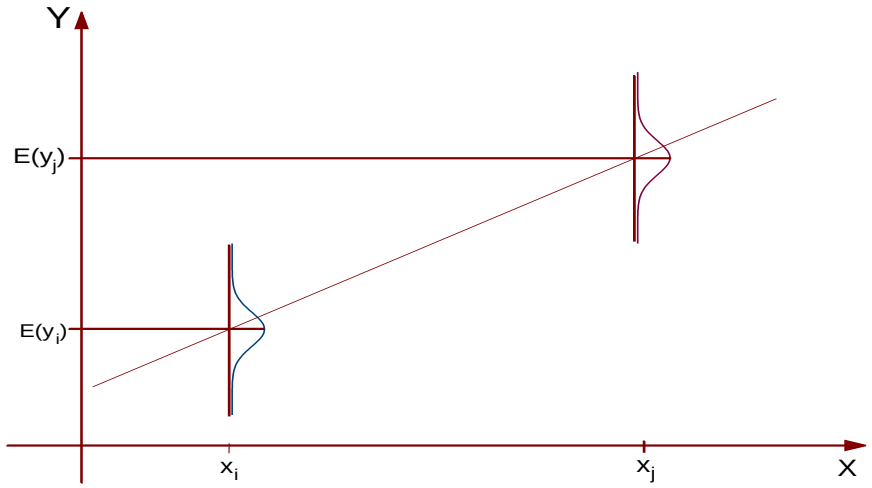

Figure 2.2: Model Assumptions about the randomness of observations.

For simplicity of notation we define

$$Y_i := Y|X = x_i. \tag{2.1}$$

Then the simple linear model can be written as

$$\begin{aligned} \mathrm{E}(Y_i) &= \beta_0 + \beta_1 x_i, \\ \mathrm{var}(Y_i) &= \sigma^2. \end{aligned}$$

If we assume normality, we have so called ***Normal Simple Linear Regression Model*** denoted in one of the equivalent ways:

- $Y_i \underset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, where $\mu_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$,
- $Y_i \underset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, 2, \ldots, n$,
- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, 2, \ldots, n$.

In all cases β_0 and β_1 are unknown constant parameters.

Example 2.2. Sparrow wings continued.
The fitted linear regression line plot, the residual diagnostics and the numerical output from MINITAB are as follows.

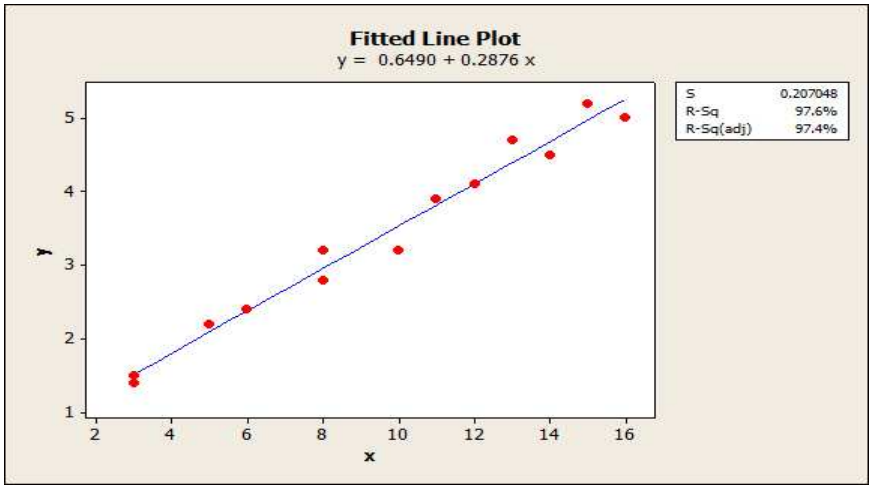

Figure 2.3: Fitted regression line for the length of sparrow wings against age of sparrows.

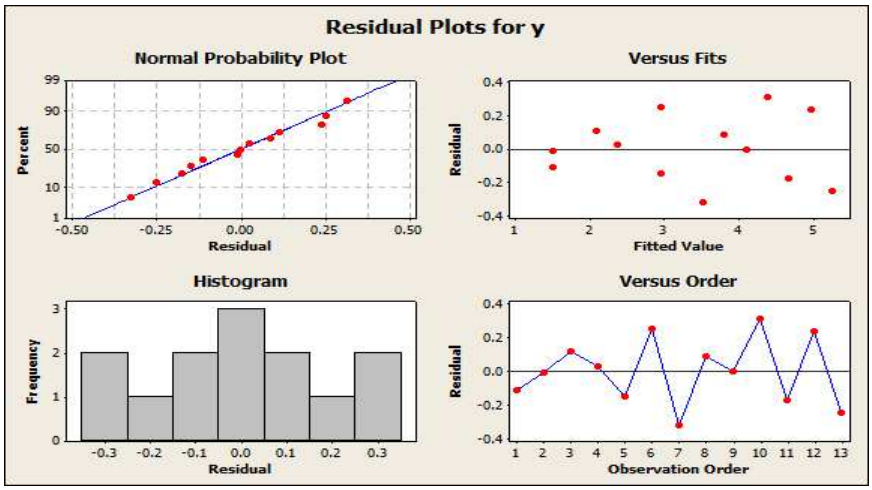

Figure 2.4: Residual diagnostics plots for the fitted regression line for the length of sparrow wings against age of sparrows.

```
Regression Analysis: y versus x

The regression equation is
y = 0.649 + 0.288 x
```

```
Predictor      Coef  SE Coef      T      P
Constant     0.6490   0.1410   4.60  0.001
x           0.28761  0.01350  21.30  0.000

S = 0.207048   R-Sq = 97.6%   R-Sq(adj) = 97.4%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  19.458  19.458  453.89  0.000
Residual Error   11   0.472   0.043
Total            12  19.929
```

MINITAB

Stat → Regression → Regression...

to obtain more numerical output than from

Stat → Regression → Fitted line plot...

The residual plots (Figure 2.4) do not contradict the assumptions of normality and a constant variance of the errors. The numerical output shows that the coefficients β_0 and β_1 are statistically significant when both are in the model. The respective values of the test function T and the corresponding p-values for testing $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ and for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ are $T = 4.60$ with $p = 0.001$ and $T = 21.30$ with $p < 0.001$.