Ch4. Correlation and simple linear regression

In this chapter we study relationships between random variables **measured together**.

Many experiments focus on establishing links between variables, for example:

- dosage of drug versus recovery time
- quantity of fertiliser versus growth of plant
- measurements of height and weight.

We discuss **two** approaches to the analysis of such data:

- **Correlation**, which measures the strength of a relationship but does not establish dependence of one variable on another
- **Regression**, which *models* the relationship by establishing a dependence.

Our data take the form of pairs of observations

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

collected together – i.e. (X, Y) is a **bivariate** random variable.

- Observations on pairs are assumed to be independent
- These data could have arisen from a random sample of n individuals from a population, or
- from an experiment in which one variable is held fixed at certain levels and measurements of the **response** variable are taken at each of these levels
- The first step is *always* to draw a scatter diagram

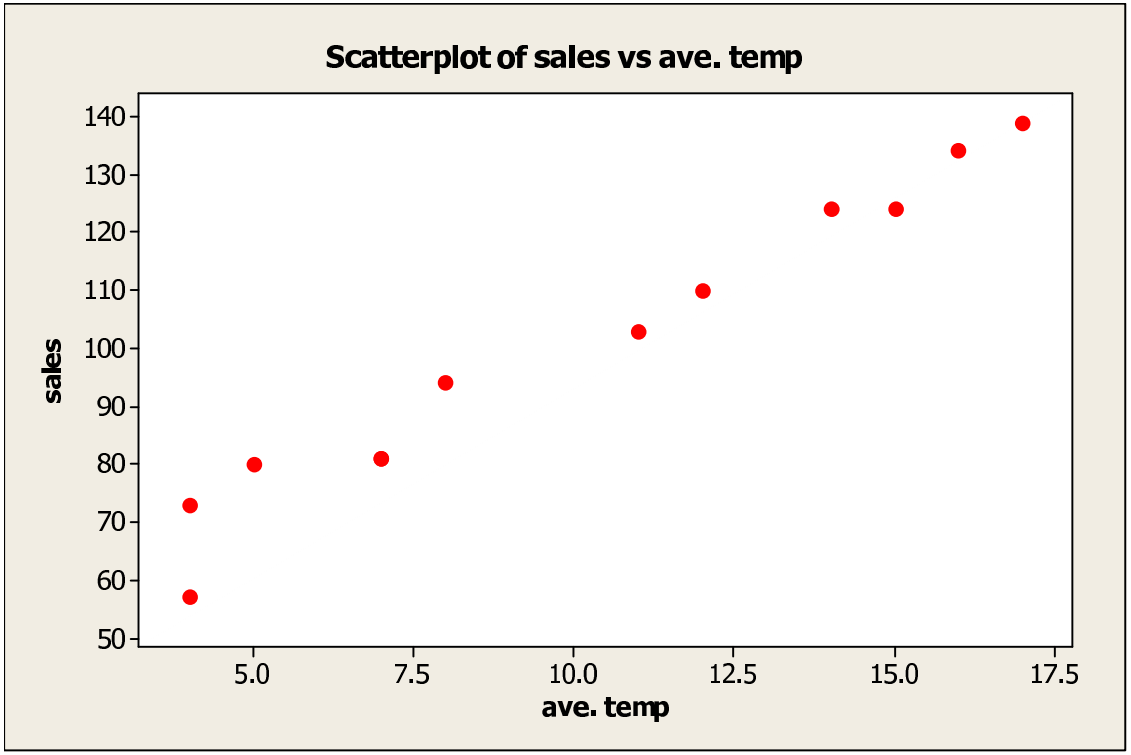
Example: ice cream sales

Ice cream sales at Luigi Minchella's ice cream parlour.

Month	**Average Temp** (^{o}C)	**Sales** (k pounds)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

- Is there any relationship between average temperature and ice cream sales?
- How would you **describe** this relationship?

We can answer such questions more easily by looking at a **scatter plot** of the data (in `Minitab` use `Graph` – `Scatterplot` – `Simple`).



Looking at the scatter plot, we see that

- as average temperature increases, sales also increase – i.e. there is a **positive** relationship
- It looks like we could draw a straight line through the data – i.e. there is a **linear** relationship
- There won't be too much scatter around this line, and so this linear relationship is **strong**
- So average temperatures and ice cream sales have a **strong**, **positive**, **linear** relationship

4.2 Correlation

The **population correlation coefficient**, ρ, is defined as

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \times \text{var}(Y)}}$$

- $-1 \leq \rho \leq 1$
- $\rho = \pm 1$ corresponds to a **perfect linear relationship**
  - If ρ is near $+1$, there is a strong *positive* linear relationship
  - If ρ is near -1 there is a strong *negative* relationship
- $\rho = 0$ indicates complete absence of such a relationship

Positive correlation

Negative Correlation

We can estimate ρ with the **Pearson product moment correlation coefficient**, r, if we have obtained n pairs of observations $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$.

The formula for r is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2, \\ S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2. \end{aligned}$$

Example: ice cream sales

To calculate r we can draw up a table (or use a calculator!)

	x	y	x^2	y^2	xy
	4	73	16	5329	292
	4	57	16	3249	228
	7	81	49	6561	567
	⋮	⋮	⋮	⋮	⋮
	5	80	25	6400	400
$\sum$	120	1200	1450	127674	13362

Thus $n = 12$,

$$\bar{x} = 120/12 = 10 \qquad \text{and } \bar{y} = 1200/12 = 100.$$

and

$$S_{XY} = \left(\sum xy\right) - n\bar{x}\bar{y} = 13362 - 12 * 10 * 100 = 1362.$$

Similarly,

$$\begin{aligned} S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2 = 1450 - 12 * (10)^2 = 250, \quad \text{and} \\ S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2 = 127674 - 12 * (100)^2 = 7674. \end{aligned}$$

Thus,

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\ &= \frac{1362}{\sqrt{250 \times 7674}} \\ &= 0.983 \text{ (to 3 decimal places).} \end{aligned}$$

This implies a **strong, positive (linear)** relationship between average temperature and ice cream sales, which agrees with what we see in the scatterplot.

4.3 Simple linear regression

A correlation analysis may establish a linear relationship but does not allow us to *use* it to, say, predict the value of one variable given the value of another.

Regression analysis allows us to do this and more.

In this model, we regard one variable, Y, as **dependent** and the other, X, as **explanatory**.

The aim is to formulate a model for predicting Y from X. We have

$$Y = \alpha + \beta X + \epsilon,$$

where α and β are unknown parameters (**intercept** and **slope**), and ϵ represents the **scatter** about the line.

If we observed n pairs of observations $(y_i, x_i), i = 1, \ldots, n$, we have

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

We assume that the **random error** $\epsilon_i \sim N\left(0, \sigma^2\right)$, independently.

To estimate α and β we use **least squares**.

This means choosing their values such that

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i}^{n} (y_i - \alpha - \beta x_i)^2, \qquad i = 1, 2, \ldots, n$$

is minimised.

Doing so gives estimates for α and β as

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{and}$$

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}}.$$

- S_{XY} and S_{XX} are as before.
- $\hat{\alpha}$ and $\hat{\beta}$ are called the **least squares estimates** of α and β.

Example: ice cream sales

We now use simple linear regression to fit a regression line through the ice cream sales data. The equation of the regression line is

$$Y = \alpha + \beta X + \epsilon,$$

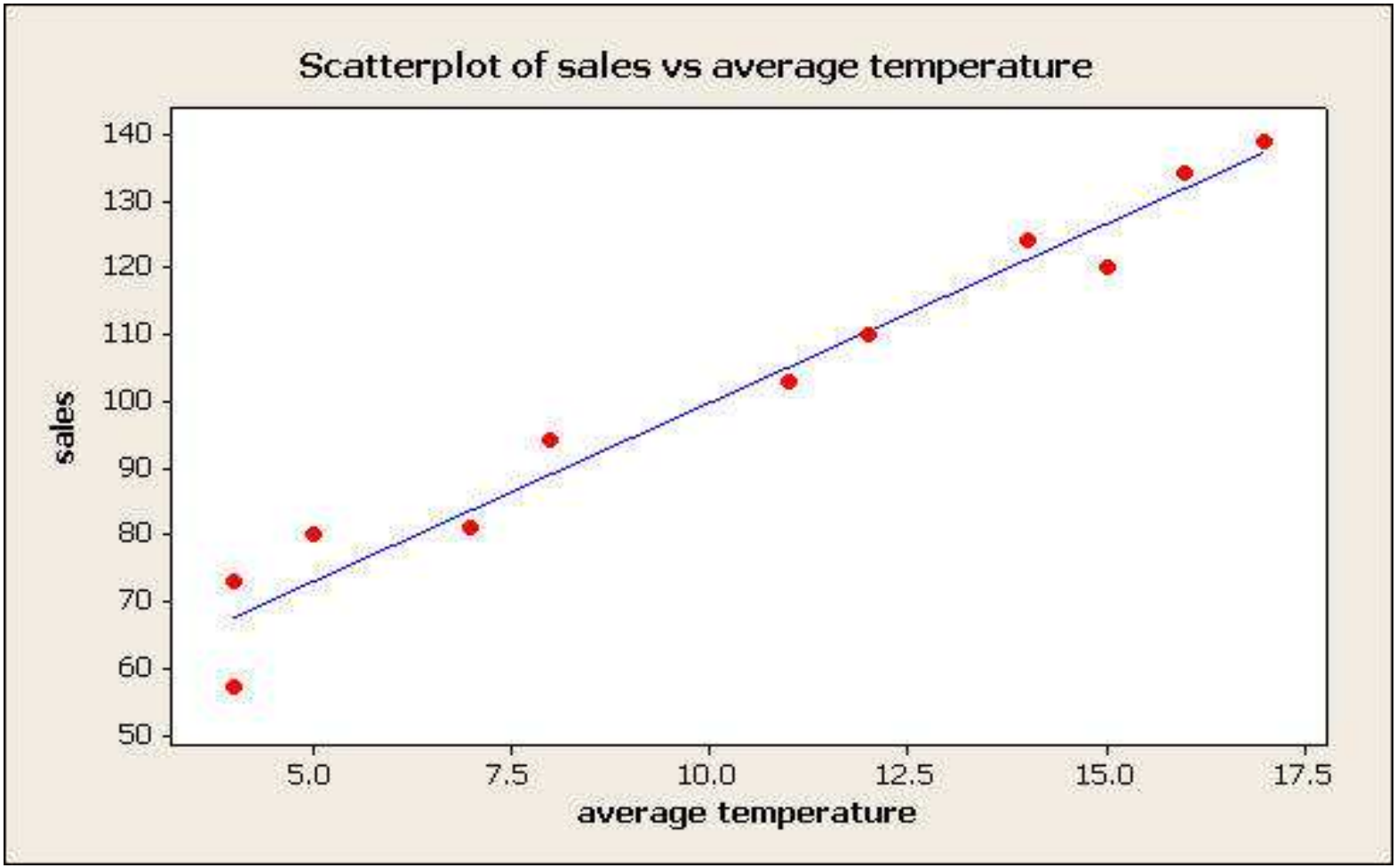
where we can estimate α and β using

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{1362}{250} = 5.448 \quad \text{and}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 100 - 5.448 \times 10 = 100 - 54.48 = 45.52.$$

Thus, the regression equation is

$$Y = 45.52 + 5.448X + \epsilon.$$



Fitted line:

$$\begin{aligned} Y &= \alpha + \beta X \\ &= 45.52 + 5.448X. \end{aligned}$$

Predictions

We can use our regression equation to predict ice cream sales for a given temperature.
For example, if we want to predict sales if the monthly average temperature is 10°C, we can simply substitute 10 into the regression equation and solve for Y.

$$\begin{aligned} Y &= 45.52 + 5.448 \times 10 \\ &= 100, \end{aligned}$$

i.e. the **prediction** of the sales is £100,000 if the monthly average temperature is 10°C.

Assumptions

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where ϵ_i or its estimation is called **residual**.

The key assumptions underlying any simple linear regression analysis are:

- The residuals are **independent**
- The residuals are **Normally distributed**
- The residuals have **common variance** (*heteroscedasticity*)

These can all be checked in `Minitab`. The following slides show a full regression analysis on the ice cream sales data in `Minitab`, including the checking of assumptions.

Regression analysis in Minitab

1. Checking for an association

We have already checked to see if there is an association between average temperature and ice cream sales via a scatterplot.

We have also calculated the sample correlation coefficient $r = 0.983$. Let's see how to do this in Minitab.

If the two samples are in columns C1 and C2 of a Minitab worksheet, then click on Stat – Basic Statistics – Correlation.

Enter the two columns in the Variables box and then hit OK.

Doing so gives the following output:

Correlations: Av. temp., Sales

```
Pearson correlation of Av. temp. and Sales = 0.983
P-Value = 0.000
```

- It is exactly the same as when we did this by hand!
- Notice that `Minitab` also gives a p–value for the correlation coefficient. This is for a hypothesis test where

$$H_0 \quad : \quad \rho = 0, \ v.s. \ H_1 : \rho \neq 0,$$

  and we interpret the ***p*–value** in exactly the same way as before.
- Thus, our correlation coefficient is **significantly different from zero**.

2. Regression analysis

Now that we've established that there's a (significant) linear association between average temperature and ice cream sales, we can perform a linear regression analysis.

In `Minitab`, click on `Stat` – `Regression` – `Regression`. Enter `C2` in `Response` and `C1` in `Predictors` and hit `OK`.

Doing so, gives:

Regression Analysis: Sales versus Av. temp.

```
The regression equation is
Sales = 45.5 + 5.45 Av.  temp.

 Predictor      Coef   SE Coef       T       P
 Constant     45.520     3.503   13.00   0.000
 Av.  temp.   5.4480    0.3186   17.10   0.000

S=5.03809 R-Sq = 96.7% R-Sq(adj) = 96.4%
```

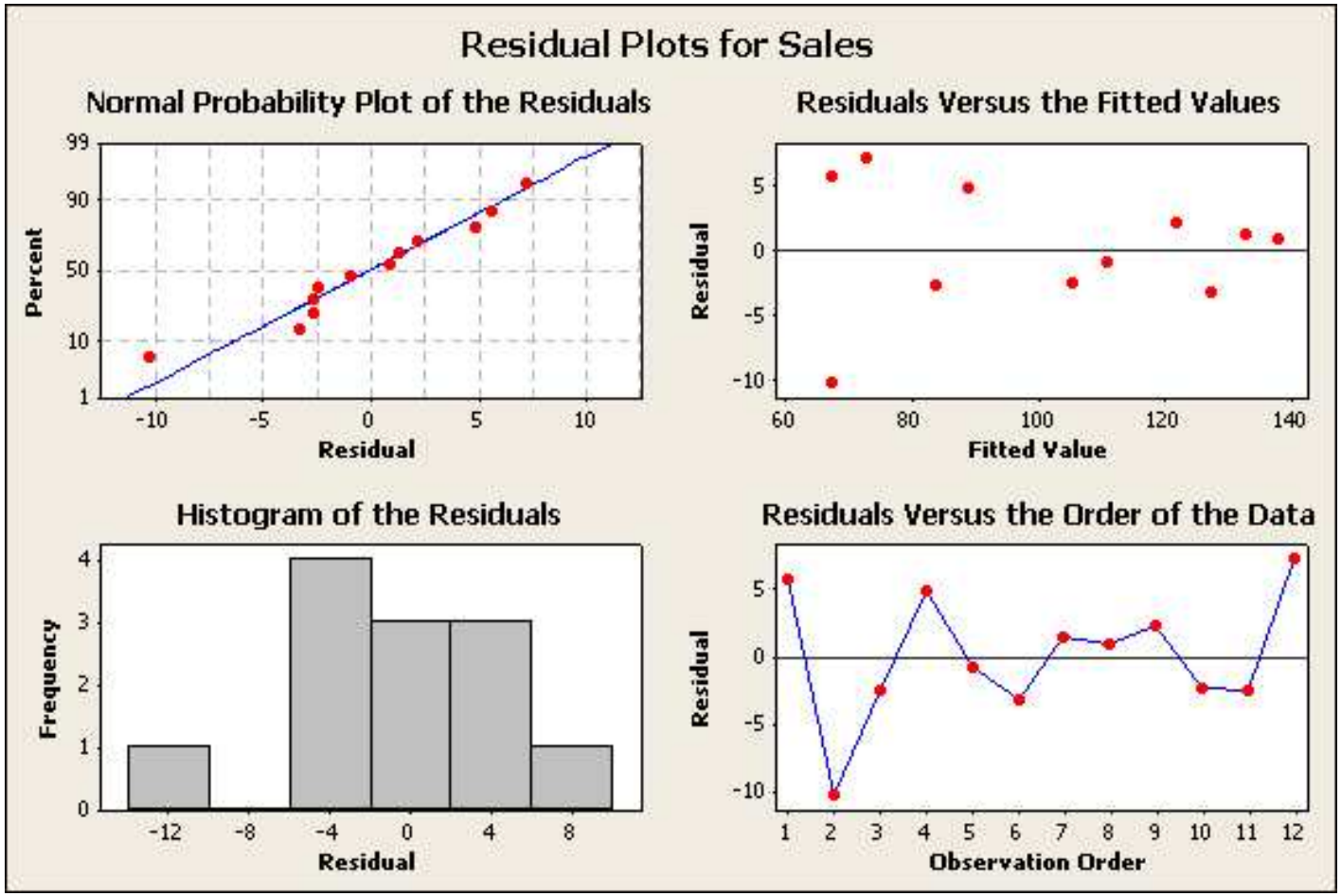
- Again, Minitab gives p–values for each of the model coefficients.
- The significance of the slope value, β, is often tested. The **p–value** is associated with $H_0 : \beta = 0$ *v.s.* $H_1 : \beta \neq 0$.
- Since our p–value is very small, we **Reject** H_0. Thus, the slope parameter β **is significantly different from zero**
- R-Sq = 96.7% suggests that the model is good fit to the data!

3. Checking assumptions

The **residual assumptions** can be checked quite readily in Minitab.

Click Stat – Regression – Regression, and enter the Response and Predictor variables as before. Click Graphs and select Four in one, and hit OK twice.

Doing so will give you the same output as before, along with the following panel of graphs.


Residual Plots for Sales
Normal Probability Plot of the Residuals
Percent
99
90
50
10
1
-10
-5
0
5
10
Residual
Residuals Versus the Fitted Values
Residual
5
0
-5
-10
60
80
100
120
140
Fitted Value
Histogram of the Residuals
Frequency
4
3
2
1
0
-12
-8
-4
0
4
8
Residual
Residuals Versus the Order of the Data
Residual
5
0
-5
-10
1
2
3
4
5
6
7
8
9
10
11
12
Observation Order

The two left–hand plots indicate the **Normality** assumption for the residuals.

- In the **Normal probability plot**, most of the points lie close to the diagonal line, indicating a Normal distribution for our residuals
- The fit to the Normal distribution can also be checked by examining the **histogram of residuals**
- Normal assumption seems acceptable.

The top right–hand plot shows the **residual plots**. It has **constant variance** if

- the residulas are allocated uniformly, and
- there is no any particular pattens.
- The **constant variance** assumption seems doubtful for this example although we should be careful to give any conclusion due to small sample size.

Extensions

- Other correlation coefficients, such as **Spearman's rank correlation coefficient** are also available.
- The followings are some other regression models:
  - **multiple regression** for the case with more than one **explanatory** variables;
  - **Ordinal logistic regression** for survey data or categorical data;
  - **Non–linear regression** for a nonlinear system, e.g.
    - Quadratic regression equation, or
    - Cubic regression equation.