4 Correlation and simple linear regression

4.1 Introduction

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

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

which they are 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 to analyze such data is *always* to draw a scatter diagram.

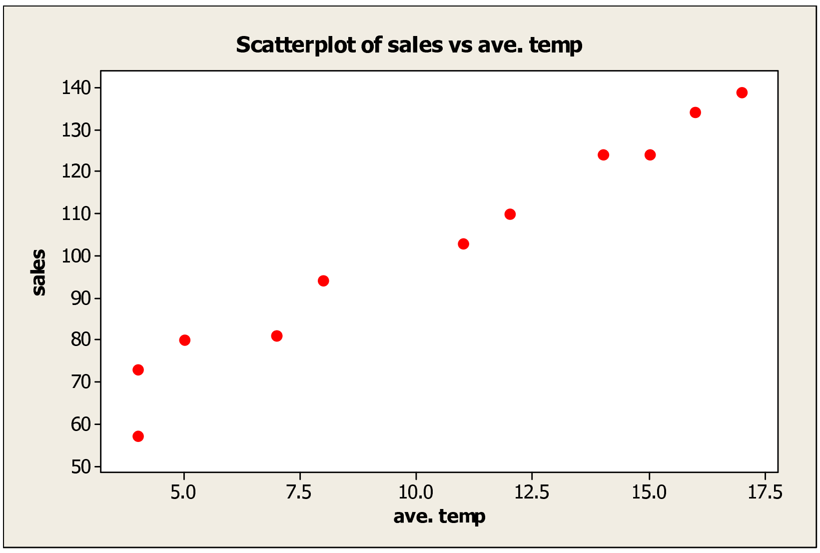
Example: ice cream sales. Consider the following data for ice cream sales at Luigi Minchella's ice cream parlour.

Month	***Average Temp*** (°C)	***Sales*** (£000's)
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

For this data set, we are interested in the following questions.

- 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 between 'sales' and 'ave. temp'.
- 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

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

It has the following properties.

- $-1 \leq \rho \leq 1$.
- $\rho = \pm 1$ 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.

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

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

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

We have a sample size of $n = 12$. Thus,

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

Similarly,

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\ &= 13362 - 12000 \\ &= 1362, \end{aligned}$$

$$\begin{aligned} S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2 \\ &= 1450 - 1200 \\ &= 250 \qquad \text{and} \end{aligned}$$

$$\begin{aligned} S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2 \\ &= 127674 - 120000 \\ &= 7674. \end{aligned}$$

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.

Spurious correlations. Correlation is a useful tool, but it can easily mislead.

- A high correlation does not necessarily imply a ***causal*** link.
  Example. For 1945 – 1964, let

  $$\begin{aligned} x_i &= \text{number of TV licenses taken out in year } i, \qquad \text{and} \\ y_i &= \text{number of convictions of juvenile delinquents in year } i. \end{aligned}$$

  The calculated value of r turns out to be significant and positive, so we are tempted to argue that TV causes increased delinquency!

- A low correlation can hide a strong but ***non–linear relationship*** between two variables – a scatterplot should always be drawn before the correlation coefficient is calculated.

- X and Y may appear related, but might both be related to a ***third variable*** instead
  Example. X might be patients' blood pressure, and Y might be their heart–rate. X and Y might be related numerically, but only because both are related to Z, the patients' weight.

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.

We assume that $\epsilon_i \sim N(0, \sigma^2)$, 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

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

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

where S_{XY} and S_{XX} are as before. There are called *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

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \qquad \text{and}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Thus,

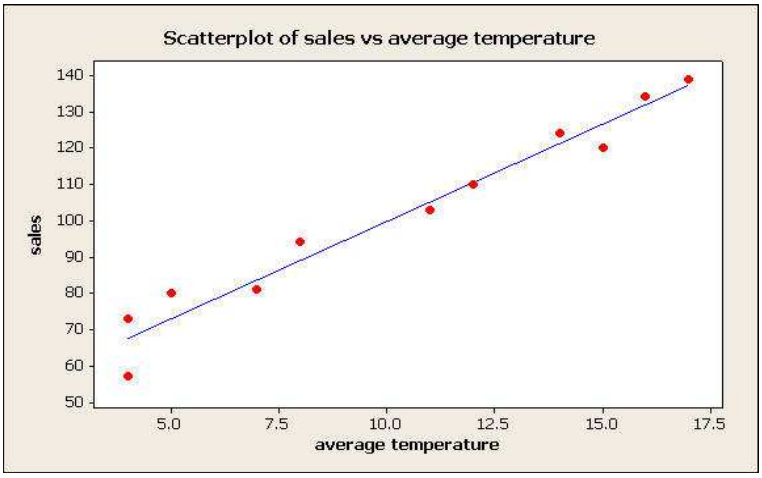
$$\hat{\beta} = \frac{1362}{250}$$

$$= 5.448 \qquad \text{and}$$

$$\hat{\alpha} = 100 - 5.448 \times 10$$
$$= 100 - 54.48$$
$$= 45.52.$$

Thus, the regression equation is

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



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 either (i) take a reading from the graph, or (ii) substitute 10 into our regression equation and solve for Y.

The second approach is probably better! Thus,

$$\begin{aligned} Y &= 45.52 + 5.448 \times 10 \\ &= 45.52 + 54.48 \\ &= 100, \end{aligned}$$

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

Remarks. You should only use your regression line to make predictions ***within the range of the observed data***. We cannot be certain that an association between the two variables will continue in the future, and even if it does, it ***might not be linear***. Making predictions which are outside the range of the given data is known as ***extrapolation***.

Assumptions: The key assumptions underlying any simple linear regression analysis are:

- The residuals, ϵ_i's, are ***independent***;
- The residuals are ***Normally distributed***;
- The residuals have ***common variance*** (*heteroscedasticity*).

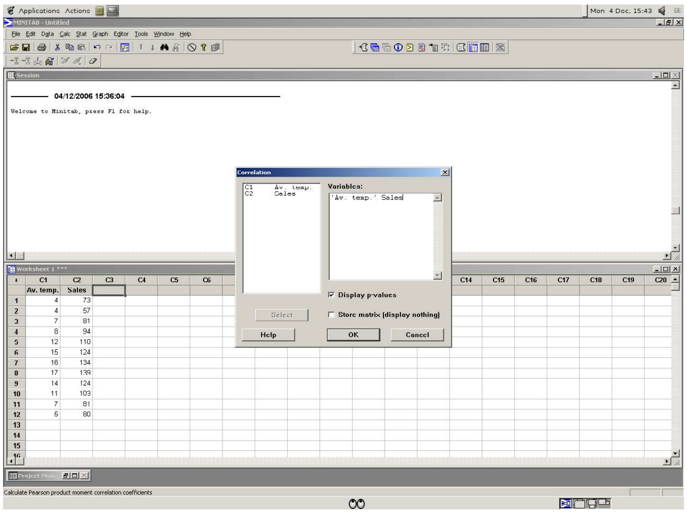
These can all be checked in `Minitab`. The followings show a full regression analysis on the ice cream sales data in `Minitab`, including the checking of assumptions.

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 scatter plot. 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
```

Which 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

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

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 the null hypothesis

$$H_0 \;:\; \beta = 0 \; v.s. \; H_1 : \; \beta \neq 0.$$

Since our p–value is very small, we ***Reject*** H_0. Thus, the slope parameter β ***is significantly different from zero***. It means that 'sales' depends on 'ave. temp' significantly.

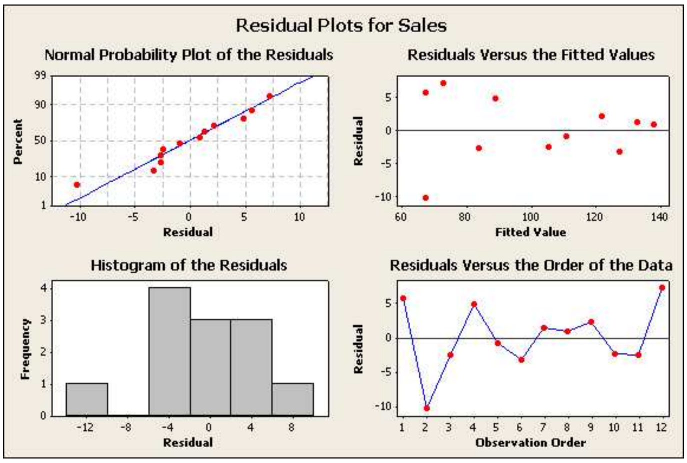
If we had *retained* H_0, then $\beta = 0$, and so the predictor variable X would have been redundant.

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.



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.***

The top right–hand plot shows random scatter, which indicates that the residuals have ***constant variance***.

4.4 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.