Introduction

Email: catherine.palmer@cit.ie
Office: B222L

Aims:

- Establish protocol for analysing large data sets
- Understand and write statistical reports and publications
- Highlight issues particular to large data sets
- Use R to perform statistical analyses

Focal techniques:

- Hypothesis Testing
- Experimental Design
- ANOVA
- PCA
- Logistic Regression

Assessment Breakdown:

Week 7:	In Class Test	25%
Week 12:	In Class Test	25%
Week 12:	Project	50%

Data Analysis Protocol

The aim of this section is to suggest some guidelines for successful data analysis. We break the analysis down to five stages:

1. Define the research problem and collect the data
2. Initial data analysis
3. Model Fitting
4. Model Interpretation
5. Model Validation

1. Define the research problem and collect the data

The starting point for any data analysis is to define the research problem and identify the fundamental relationships to be investigated. Once this has been done it should be possible to write down testable hypotheses and specify which variables are of interest. At this stage the model used to analyse the data is selected. The data is then collected, bearing in mind key issues such as bias and sample size as well as the fact that it must be suitable for use in the selected model. Figure 1 shows how statistical analysis is part of the scientific method.

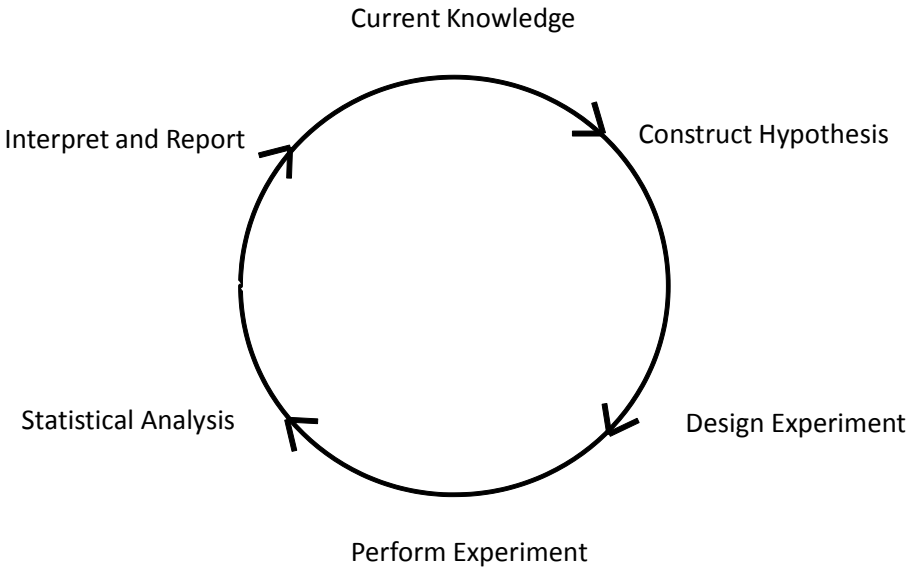


Fig. 1

An important concept in the scientific process is the idea of reproducibility: the ability to repeat a study under the same conditions and obtain the same results. Any discovery of a new effect should depend on multiple experiments under varied, but controlled, conditions. At present, there is much debate on the topic of reproducibility, with reference to a *reproducibility crisis* (Fig. 2).

Fig. 2

Current problems are thought to include:

- data dredging/p-hacking (the practice of analysing data in search of statistically significant relationships without pre-defined hypotheses)
- publication bias (the outcome of an experiment or research study influences the likelihood that it is published)
- piecemeal data analysis

Suggestions on how to improve reproducibility in science include:

- pre-registration of research
- publishing null findings
- reproducible data analysis workflow

The concept of pre-registration of research highlights the importance of defining clear, testable hypotheses at the start of any study.

2. Initial data analysis

Initial data analysis occurs before you implement the selected model. This is a **very important** step where we get to understand the data and check that it is suitable to answer the research question of interest. Initial data analysis can involve the following:

- Summarising and visualising data
- Univariate profiling
- Bivariate profiling
- Missing data
- Outliers
- Check model assumptions
- Variable transformation

Summarising and visualising data

It is important to understand a data set and relationships between variables before analysis with the chosen model.

Univariate profiling

Start by summarising each variable individually and inspecting its distribution with for example a histogram or boxplot see Fig. 3.

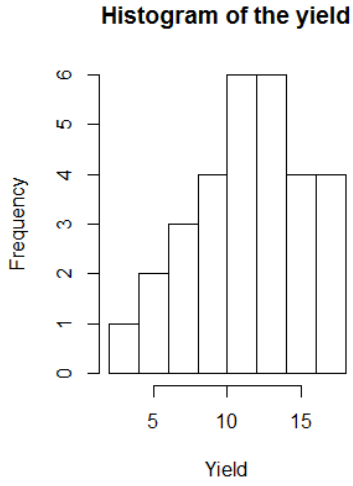


Figure 3

Bivariate profiling

To examine the relationship between two numerical variables we can use scatterplots, this determines whether the relationship is linear (Fig 4(a)), nonlinear (Fig 4(b)) or if there is no relationship (Fig 4(c)).

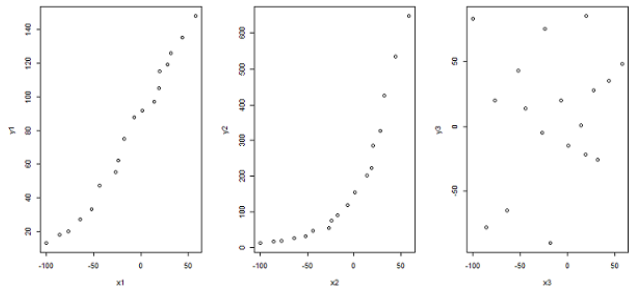


Figure 4

A scatterplot matrix is very useful if there are many variables. A scatterplot matrix shows scatterplots of all possible combinations of the variables. The scatterplot shown in figure 5 was created using the `pairs()` function in R. It shows all possible combinations of scatterplots for a data set containing four different environmental measurements, solar radiation "rad", temperature "temp", wind speed "wind" and ozone levels "ozone".

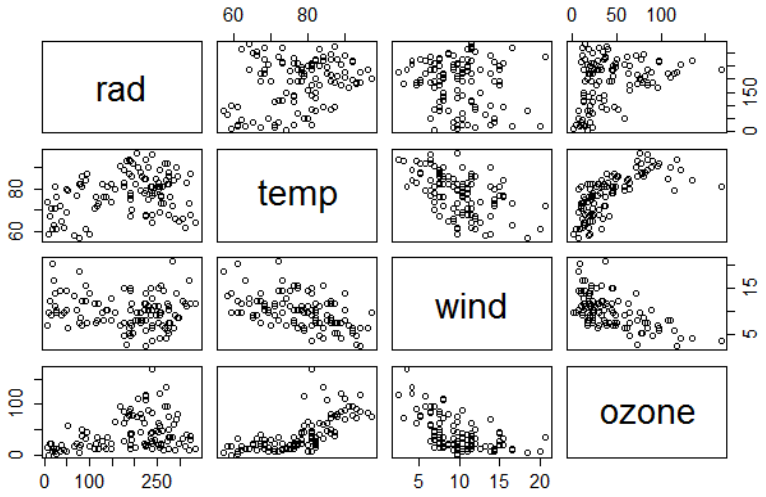


Figure 5

The response variables are named in the rows and the explanatory variables are named in the columns. Thus, in the upper row, labelled "rad", the response variable (on the y axis)

is solar radiation. In the bottom row the response variable , "ozone", is on the y axis of all three panels. There appears to be a strong negative non-linear relationship between ozone and wind speed, a positive nonlinear relationship between air temperature and ozone (middle panel, bottom row) and an indistinct perhaps humped relationship between ozone and solar radiation (left-most panel in the bottom row). We can include correlations in the lower panel and histograms along the diagonal see Fig. 6.

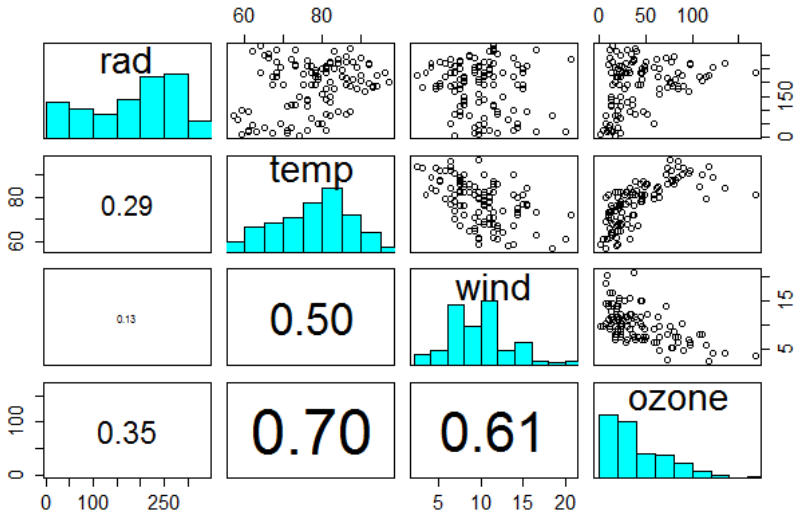


Figure 6

Conditioning plots

When you have many variables, the relationship between two variables may be obscured by the effects of other processes. In the simplest case, we have one response variable (say ozone from the example above) and just two explanatory variables (say wind speed and air temperature). To see whether the relationship between ozone and wind speed changes with temperature we can use a **conditioning plot**. A conditioning plot examines the pairwise relationship between two variables conditional on a third variable (called the conditioning variable). The conditioning variable may be either a variable that takes on only a few discrete values or a continuous variable that is divided into a limited number of subsets. In our example, we are examining the pairwise relationship between ozone and wind speed, conditional on the third variable, air temperature. Since air temperature is a continuous variable, it is divided into subsets see Fig. 7.

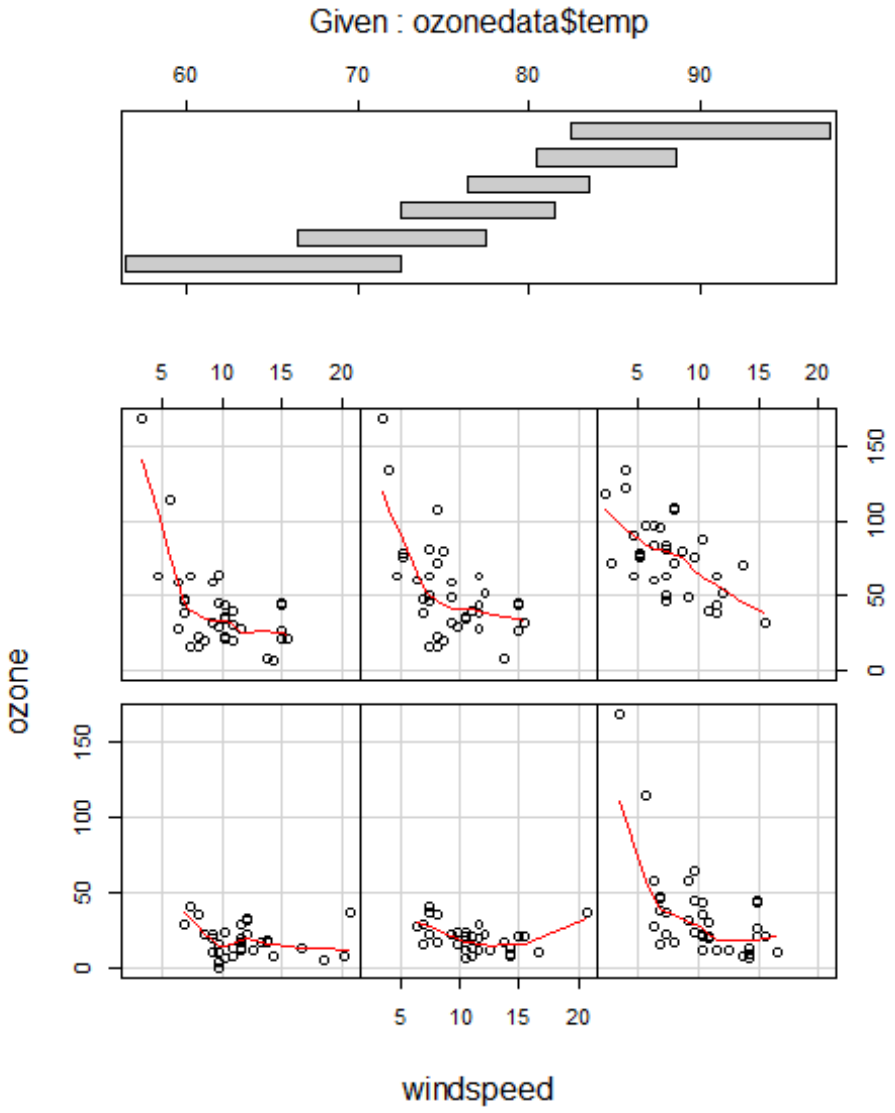


Figure 7

The plot panels are ordered from lower left to upper right, associated with the values of the conditioning variable shown in the upper panel from left to right. Thus, the lower-left plot is for the lowest temperatures (56–72 degrees F) and the upper right plot is for the highest temperatures (82–96 degrees F). This conditioning plot highlights an interesting interaction. At the two lowest levels of the conditioning variable, temperature, there is little or no relationship between ozone and wind speed, but in the four remaining panels (at higher temperatures) there is a distinct negative relationship between wind speed and ozone concentration.

Missing Data

Missing data can be a result of difficulties encountered during an experiment, errors during data collection, data entry errors or omission of answers by respondents. It is important to identify the patterns and relationships underlying the missing data so that when a method is used to deal with the missing data, the modified data remains as close as possible to the original distribution of values Alternatively, we can say that it is important to identify the patterns and relationships underlying the missing data so that we can avoid using a biased data set. Often data is missing due to a **missing data process** whereby data is omitted in a systematic (or non random) way, for example in surveys, questions about income are often unanswered. As well as looking for patterns it is important to examine the extent of the missing data, it is easier to overcome a few random missing values than it is to deal with large systematic chunks of missing data. A data analyst must understand the process leading to the missing data in order to select the appropriate course of action.

Why worry about missing data? Missing data can reduce the sample size available for the analysis. If missing data is not remedied in some way then any observation with missing values on any of the variables will be excluded from the analysis (i.e. a whole row will be excluded). Missing data may eliminate so many observations that what was an adequate sample is reduced to an inadequate sample. For example it has been shown that if 10% of the data is randomly missing in a set of 5 variables, on average almost 60% of the subjects will have at least one missing value. Thus, when complete data are required, the sample is reduced to 40% of the original size.
If data is missing due to a missing data process whereby data is omitted in a systematic way then the data set could be biased, and any conclusions drawn from statistical analysis will not be representative of the population .

Example 1 *The table below shows a simple example of missing data among 20 subjects. The number of missing data varies widely among both subjects and variables.*

						Missing Data by Subject	
Subject ID	V_1	V_2	V_3	V_4	V_5	*Number*	*Percent*
1	1.3	9.9	6.7	3.0	2.6	*0*	*0*
2	4.1	5.7			2.9	*2*	*40*
3		9.9		3.0		*3*	*60*
4	0.9	8.6		2.1	1.8	*1*	*20*
5	0.4	8.3		1.2	1.7	*1*	*20*
6	1.5	6.7	4.8		2.5	*1*	*20*
7	0.2	8.8	4.5	3.0	2.4	*0*	*0*
8	2.1	8.0	3.0	3.8	1.4	*0*	*0*
9	1.8	7.6		3.2	2.5	*1*	*20*
10	4.5	8.0		3.63	2.2	*1*	*20*
11	2.5	9.2		3.3	3.9	*1*	*20*
12	4.5	6.4	5.3	3.0	2.5	*0*	*0*
13					2.7	*4*	*80*
14	2.8	6.1	6.4		3.8	*1*	*20*
15	3.7			3.0		*3*	*60*
16	1.6	6.4	5.0		2.1	*1*	*20*
17	0.5	9.2		3.3	2.8	*1*	*20*
18	2.8	5.2	5.0		2.7	*1*	*20*
19	2.2	6.7		2.6	2.9	*1*	*20*
20	1.8	9.0	5.0	2.2	3.0	*0*	*0*
Missing data by variable						*Total Missing Values*	
Number	2	2	11	6	2	*Number*	23
Percent	10	10	55	30	10	*Percent*	23

In this example, we can see that all of the variables (V_1 to V_5) have some missing data, with V_3 missing 55% of all values. Three subjects (3, 13 and 15) have more than 50% *missing data and only five subjects have complete data. Overall, 23 percent of the data values are missing.*

The missing data in this example can become quite problematic in terms of reducing the sample size. For example, if an analysis was performed that required complete data on all five variables, the sample would be reduced to only the five subjects with no missing data

(subjects 1, 7, 8, 12 and 20). This sample size is too few for any type of analysis. One way to deal with missing data is to delete specific variables or subjects. In our example, assuming that the research is not altered substantially by the elimination of a variable, eliminating V_3 is one approach to reducing the number of missing data.

By eliminating V_3, seven additional subjects, for a total of 12, now have complete data. If the three subjects (3, 13, 15) with exceptionally high numbers of missing data are also eliminated, the total number of missing data is now reduced to only five instances, or 7.4% of all values. This may seem like a reasonable way to deal with missing data, however, if we look at the remaining data we can see that a pattern has emerged. The remaining data is shown in the table below:

					Missing Data by Subject	
Subject ID	V_1	V_2	V_4	V_5	Number	Percent
1	1.3	9.9	3.0	2.6	0	0
2	4.1	5.7		2.9	1	25
4	0.9	8.6	2.1	1.8	0	0
5	0.4	8.3	1.2	1.7	0	0
6	1.5	6.7		2.5	1	25
7	0.2	8.8	3.0	2.4	0	0
8	2.1	8.0	3.8	1.4	0	0
9	1.8	7.6	3.2	2.5	0	0
10	4.5	8.0	3.63	2.2	0	0
11	2.5	9.2	3.3	3.9	0	0
12	4.5	6.4	3.0	2.5	0	0
14	2.8	6.1		3.8	1	25
16	1.6	6.4		2.1	1	25
17	0.5	9.2	3.3	2.8	0	0
18	2.8	5.2		2.7	1	25
19	2.2	6.7	2.6	2.9	0	0
20	1.8	9.0	2.2	3.0	0	0
Missing data by variable					Total Missing Values	
Number	0	0	5	0	Number	5
Percent	0	0	29.4	0	Percent	7.4

Compare the values of V_2 for the remaining 5 subjects with missing data for V_4 (subjects 2, 6, 14, 16 and 18) versus those subjects having valid V_4 values, a distinct pattern emerges.

The five subjects with missing values for V_4 have the five lowest values for V_2, indicating that missing data for V_4 are strongly associated with lower scores on V_2. This systematic association between missing and valid data directly effects any analysis in which V_4 and V_2 are both included. For example, the mean score for V_2 will be higher if the five subjects with missing data on V_4 are excluded (mean = 8.4) than if those five subjects are included (mean = 7.8). In this example, eliminating subjects and variables was not necessarily the best way to address the problem of missing values. It is important to try to identify any possible missing data processes before taking action.

Steps to Identify and Address Missing Data

Step 1 Determine the type of missing data

Ignorable missing data Often, missing data are expected and are part of the research design, in which case they are referred to as ignorable missing data and no action is necessary e.g. in a customer satisfaction survey, there is a section on complaints, these questions can only be answered if a customer has made a complaint.

Missing data that are not ignorable Non ignorable missing data can be a result of either a **known** missing data process or an **unknown** missing data process. Many missing data processes are known to the researcher in that they can be identified due to procedural/experimental factors, such as errors in data entry that create invalid codes, confidentiality, death, dropping out, known experimental difficulties. Unknown missing data processes are less easily identified e.g. a subject may refuse to respond to certain questions, unknown experimental difficulties (gremlins!)

Step 2 Determine the extent of the missing data

If the missing data is non ignorable then the next step is to examine the patterns of the missing data and determine the extent of the missing data for individual variables, individual subjects/observations and overall. The aim is to determine whether the amount of missing data is low enough to not affect the results, even if it operates in a non random manner. To assess the extent of missing data we may tabulate

1. the percentage of variables with missing data for each subject
2. the number of subjects with missing data for each variable.

We should also calculate the number of subjects with no missing data on any of the variables since this will give the sample size if subjects with missing data are eliminated.

Once we have determined the extent of the missing data we can decide how to proceed. If the percentage of missing data is high and occurs in a non random way then ignoring the missing data or substituting some values in for the missing data can create bias in the data that will affect the results. In general, missing data under 10% for an individual subject or observation can generally be ignored except when the missing data occurs in a specific non random way. If the simple method of deleting subjects with missing data is used then the number of subjects with no missing data must be sufficient to carry out the analysis. If deletion of variables or subjects is used then a compromise must be found between deleting variables or subjects with missing values verses the reduction in sample size and variables. Variables or subjects with 50% or more missing values should be eliminated but as the level of missing data goes down, decisions should be based on judgement and trial and error (apply more than one method and compare the results).

Step 3 Diagnose the randomness of the missing data

If the extent of the missing data is enough to impact the results of the analysis, the next step is to ascertain the degree of randomness present in the missing data. There are three levels of randomness:

- MCAR missing completely at random
- MAR missing at random
- MNAR not missing at random

To illustrate what the three levels of randomness mean, consider two variables, X and Y where X has no missing data but Y has some missing data.

Definition 1 *Missing data is deemed to be MCAR if the missing values of Y are independent of both X and Y.*

In simple terms, if missing data is MCAR then the cases with missing data are completely indistinguishable from cases with complete data.

Example 2 *In a questionnaire, assume we know the gender of respondents (the X variable) and are asking about household income (the Y) variable. If we find that the missing data for household income were randomly missing in equal proportions for both males and females then the missing data is MCAR.*

Definition 2 *Missing data is deemed to be MAR if the missing values of Y depend on X but not on Y.*

Example 3 *In the same questionnaire, where we know the gender of respondents (the X variable) and are asking about household income (the Y) variable. If we find that the missing data for household income occurred more frequently for males than females (or vice versa) then the missing data is MAR.*

Definition 3 *Missing data is deemed to be MNAR if the missing values of Y depend on Y*

Example 4 *In the same questionnaire, If it is more likely that a respondent with high income is less likely to report income then the missing data is MNAR*

If the data set is small then it may be possible to ascertain the degree of randomness present in the missing data visually or with very simple calculations as in example 1 above. As the sample size and the number of variables increase it is necessary to use diagnostic tests. To find out whether the missing values from a variable, say Y, depend on the values of another variables, say $X_1, X_2, ...$ we can split the subjects/observations for Y into two groups: subjects/observations of Y that have missing data and subjects/observations of Y that have valid data. The same grouping can be applied to the other variables and statistical tests (e.g. t-test, chi-squared test) can be performed. Significant differences indicate the possibility of a nonrandom data process.

Example 5 *Let's return to the questionnaire where we know the gender of respondents (the X variable) and are asking about household income (the Y) variable. We can split the respondents into two groups: those that have not answered the household income question and those that have. We would then calculate the proportion of gender for each group. If one gender (e.g. males) was found in greater proportion in the missing data group we would conclude that the missing data is MAR.*

Step 4 Select the Imputation Method

Imputation is the process of estimating a missing value based on valid values of other variables an/or subjects/observations in the sample. The objective is to use known relationships that can be identified in the valid data to assist in estimating missing values. All of the methods discussed below are used with metric variables; non metric variables are left as missing unless a specific modeling approach is employed (e.g. logistic regression).

Imputation of a MCAR Missing Data Process If the missing data process can be classified as MCAR then, either of two basic approaches can be used: use only valid data or define replacement values.

Use only valid data If only valid data is used then the data can be analysed using:

(a) the complete case approach

(b) all available approach

The complete case approach eliminates any subject/observation with missing values. To use this method there must be no doubt that the missing data really is MCAR otherwise deletion of incomplete subjects will result in the remaining data set being biased. This approach reduces the sample size and should only be performed if the remaining sample will be large enough to perform the analysis. This is the default method used by **R** for linear regression.
The all available approach does not impute missing values, it uses all the valid data available to calculate the distribution characteristics (e.g. mean or standard deviation) or relationships (e.g. correlations or regressions) using every valid value. For example, consider the case where there are three variables, V_1, V_2 and V_3. To estimate the mean of each variable, all of the valid values are used for each subject. If a subject is missing data for variable V_3 the values for V_1 and V_2 are still used to calculate the means.

Define replacement values Imputation using replacement values is where missing values are replaced with estimated values based on other information available in the sample. There are many ways to define replacement values including:

- hot deck imputation - where the imputed value comes from another subject/observation in the sample that is deemed similar
- cold deck imputation - where the value is derived from an external source e.g. prior studies, other samples)
- case substitution - where entire observations with missing data are replaced by choosing another non sampled observation
- mean substitution - where the missing value of a variable is replaced with the mean value of that variable calculated from all valid values

- regression imputation - here regression analysis based on the valid data is used to predict the missing values of a variable.

Imputation of a MAR Missing Data Process If a MAR data process is found, then a specifically designed modelling approach must be applied, either maximum likelihood estimation (MLE) or multiple imputation. Maximum likelihood techniques attempt to model the processes underlying the missing data and to make the most accurate and reasonable estimates possible. A maximum likelihood estimate is the value that is most likely to have resulted in the observed data. Multiple imputation is a method where a range of possible values is defined for each missing value (e.g. a distribution or a regression model is defined). These distributions or regressions are used to impute (fill in) values for the missing data to give m different, complete data sets. Each imputed data set is analyzed separately using the model of choice and the m different parameter estimates are then averaged to give the final model. The standard errors associated with the parameter estimates used in the final model account for variation due to imputation as well as the residual variance of the model.

Outliers

Outliers are observations with a *unique combination of characteristics identifiable as distinctly different from the other observations*. Typically, an outlier is an observation that has an unusually high or low value on a variable or a unique combination of values across several variables that make the observation stand out from the others. In assessing the impact of outliers, we must consider:

- Outliers can have a large effect on any type of empirical analysis. For example, assume that we sample 20 individuals to determine the average household income. In our sample we gather responses that range between €20,000 and €100,000, so that the average is €45,000. But assume that the 21st person has an income of €1 million. If we include this value in the analysis, the average income increases to more than €90,000. The outlier is a valid observation, but what is the better estimate of household income: €45,000 or €90,000? The researcher must assess whether the outlying value is retained or eliminated due to its undue influence on the result.

- How representative of the population is the outlier? Using the example of household income, how representative of the more wealthy segment is the millionaire? If the researcher feels that it is a small but viable segment of the entire population then

perhaps it should be retained. If, however, this millionaire is the only one in the population and truly far above everyone else (i.e. unique) and represents an extreme value then it may be deleted.

Outliers cannot easily be categorized as either problematic (not representative of the population) or beneficial (although different from the majority of the sample they do represent part of the population). Instead outliers should be considered within the context of the analysis. Why do outliers occur? Outliers can be classified into one of four classes based on the source of their uniqueness:

1. Outliers can occur due to **procedural error**, such as data entry error, or a mistake in coding. If possible these outliers should be corrected but if not they should be recorded as missing values.

2. Outliers can occur due to an **extraordinary event**. For example, assume we are tracking average daily rainfall, when we have a storm that lasts for several days and records extremely high rainfall levels. The researcher should decide whether the extraordinary event fits the objectives of the research. If so, the outlier should be retained in the analysis. If not, it should be deleted.

3. If the researcher has no explanation for the outlier, it is classified as an **extraordinary observation**. These outliers are likely to be deleted or recorded as a missing value but they may be retained if the researcher feels that they are representative of the population. Perhaps they represent an emerging pattern or a previously unidentified pattern.

4. An outlier may fall within the ordinary range of values for each of the variables. These observations are not particularly high or low on the variables, but are unique in their combination of values across the variables. In this situation the outlier is usually retained.

Methods for Detecting Outliers

Outliers can be identified from a univariate, bivariate, or multivariate perspective based on the number of variables considered.

Univariate Detection The univariate identification of outliers examines the distribution of observations for each variable in the analysis and selects as outliers those cases falling at the outer ranges (high or low) of the distribution. The primary issue is establishing the threshold for designation of an outlier. The typical approach first converts the data values to standard scores, which have a mean of 0 and a standard deviation of 1. Because the values are expressed in standard form, comparisons across variables can be made easily. Tests include: Chauvenet's criterion and Grubb's test for outliers.

Bivariate Detection Pairs of variables can be assessed jointly through a scatterplot. Observations that fall outside the range of the other observations will be seen as isolated points on the scatter plot. To assist in determining the expected range of observations in the scatterplot, an ellipse representing a bivariate normal distribution's confidence interval (typically set to 95% level) is superimposed on the scatter plot. A drawback of the bivariate method is the potentially large number of scatterplots that arise as the number of variables increases. For three variables, it is only 3 graphs but for 10 variables it is 45 graphs.

Multivariate Detection When more than two variables are considered, the researcher needs a way to measure the multidimensional position of each observation relative to some common point. The Mahalanobis D^2 measure is a multidimensional version of a z-score. It measures the distance of an observation from the centroid (multidimensional mean) of a distribution. Higher D^2 scores represent observations farther moved from the general distribution of observations in this multidimensional space. An observation is a multivariate outlier if the probability associated with its D^2 score is 0.001 or less. Mahalanobis D^2 requires that the variables be metric. This method only provides an overall assessment, it provides no insight into which variables might lead to a high D^2 value. Multivatate outliers can also be detected by performing Principal Components Analysis on the data and examining the resulting scores plot.

Testing Model Assumptions

It is important to test the data to ensure that it satisfies the statistical assumptions underlying the model of choice. Some models are less affected by violating certain assumptions, which is termed **robustness**.
For linear models, of the form

$$y_i = a_0 + a_1 x_i + e_i$$

the following assumptions are made:

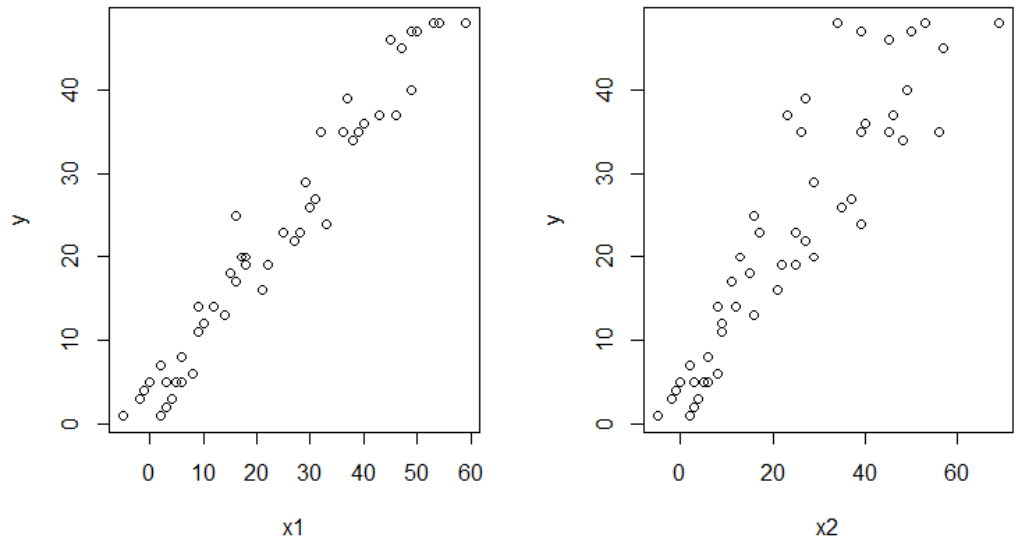
- Linearity - the relationship between x and y is linear.
- Homoscedasticity - the errors, e_i have constant variance
- Normality - the errors e_i are normally distributed

Linearity

Many of the most commonly used statistical models assume a linear relationship between the response variable and the explanatory variables. Linearity can be checked using scatterplots or by running a linear regression analysis and examining the residuals. The residuals represent the unexplained variation in the data; thus any nonlinear part of the relationship, unexplained by the linear model will show up in the residuals. If a nonlinear relationship is detected then one or both variables can be transformed to achieve linearity. Alternatively a nonlinear model can be used.

Homoscedasticity

Homoscedasticity refers to the assumption that explanatory variables have equal levels of variance across the range of the response variable see Fig 1(a). When carrying out the ANOVA on the yield data, we checked for homoscedasticity using the Fligner-Killeen homogeneity of variance test. Another way of describing homoscedasticity is to say that the error terms (residuals) have constant variance (see Fig 2 (a)). If the variance of the explanatory variables is not constant with respect to the response variable (see Fig 1(b)) then the relationship is said to be heteroscedastic and while regression coefficient estimates are not affected, the accompanying standard errors can be biased, possibly above or below the population standard errors. If the relationship is heteroscedastic then the error terms (residuals) will not have constant variance (see Fig 2(b)).



Figs 1(a) and 1(b)

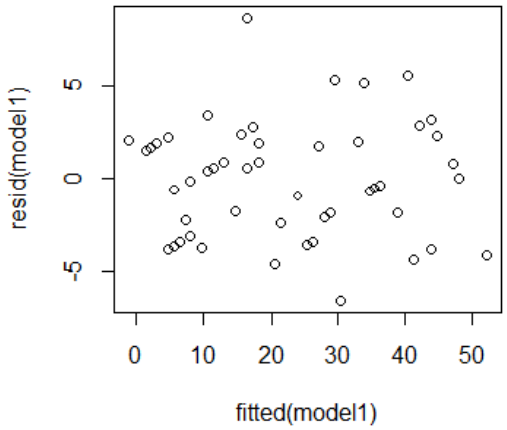


Figure 2(a)

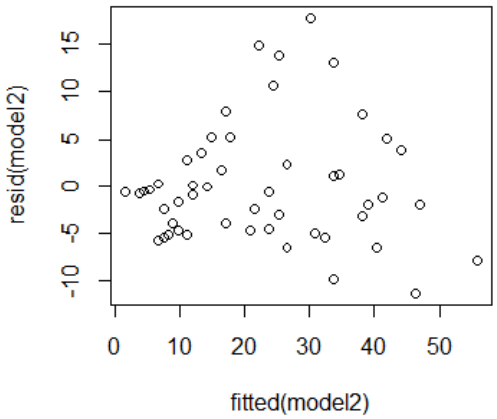


Figure 2(b)

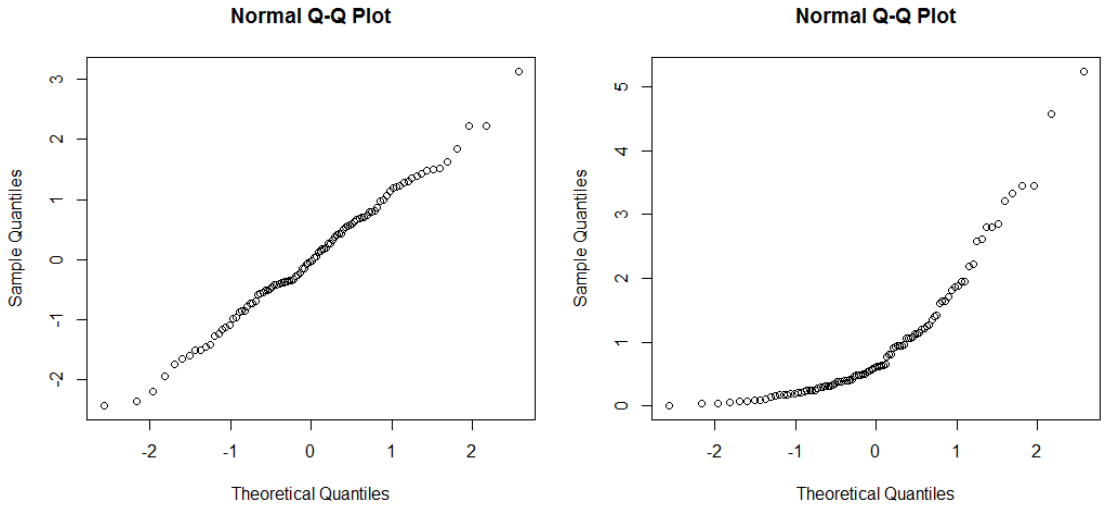
Often heteroscedasticity is a result of one of the variables having a nonnormal distribution, and correction of nonnormality via a transformation can also correct heteroscedasticity.

Normality of the variables

The most fundamental assumption in many statistical analyses is **normality**, referring to how similar the shape of the distribution of an individual metric variable is to the normal distribution. A researcher should always assess the normality for all metric variables included in the analysis. Two terms used to describe the way in which a distribution can differ from the normal distribution are the **kurtosis** and the **skewness**. Kurtosis refers to the "peakedness"

or the "flatness" of the distribution compared to the normal distribution. Distributions that are taller or more peaked than the normal distribution are termed leptokurtic whereas a distribution that is flatter is termed platykurtic. Skewness is used to measure the asymmetry of a distribution. A positive skew denotes a distribution shifted to the left of the mean whereas a negative skew reflects a shift to the right.

Graphical Analyses of Normality The simplest plot for assessing normality is the histogram however this can be problematic for small samples where the shape of the histogram is dependent on the number/width of categories. The normal probability plot (normal Q-Q) plot compares the quantiles of actual data values with the quantiles of a normal distribution. The normal distribution forms a straight diagonal line, and the plotted data values are compared with the diagonal. If a distribution is normal, the line representing the actual data distribution closely follows the diagonal.



The Q-Q plots above have been calculated for two different samples. What can we say about the normality of the data in each of the samples?

Statistical Tests of Normality A z-value can be calculated from the data for both the kurtosis (z-kurtosis) and the skewness (z-skewness), the distribution is said to be nonnormal if either of the z values calculated are greater than a specified critical value. (e.g. 1.96 corresponds to the 0.05 significance level) Two common statistical tests for normality are

the Shapiro-Wilks test and the Kolmogorov-Smirnov test. Each calculates the level of significance for the differences from a normal distribution. **It is important to remember that tests of significance are dependent on sample size.**

Nonnormality can be corrected using a transformation.

Data Transformation

If we find that our data violates one or more of the assumptions then it can be modified by applying a transformation.

A transformation on the data set $x_1, x_2, x_3, ..., x_n$ is a function T that replaces each x_i value by a new value $T(x_i)$, so that the transformed values of the data set are $T(x_1), T(x_2), ..., T(x_n)$.

The transformations used should not change the relative ordering of the values but can alter the distance between successively ordered values to change the overall shape of the distribution. The choice of transformation depends on the nature of the data. The data can violate the assumption(s) of:

- Normality
- Heteroscedasticity
- Linearity

Standardising data An important example of transforming data is standardisation, whereby values are adjusted for differing mean and variance. Standardised values have mean 0, variance1 and have no units: hence standardisation is useful for comparing variables expressed in different units. Most commonly a standard score is calculated using the mean and standard deviation (S.D.) of a variable:

$$z = \frac{x - \overline{x}}{\sigma}$$

Standardisation makes no difference to the shape of a distribution but the coefficients of a model where the data has been standardised will reflect the meaningful relative influence of each explanatory variable (i.e., a positive coefficient will mean that the variable acts positively towards the response variable, and vice versa, plus a large coefficient versus a small coefficient will reflect the degree to which that variable influences the response variable).

3 & 4 Model Fitting and Model Interpretation ....see specific models

5. Model Validation

After fitting a model, we need to validate whether it correctly describes:

(a) the relationships between variables present in the data set

(b) the relationships between variables present in the population.

The details of the validation process depend on the model that has been used. For linear models of the form

$$y_i = a_0 + a_1 x_i + \varepsilon_i$$

the following assumptions are made:

- Linearity - the relationship between x and y is linear.
- Homoscedasticity - the errors, ε_i have constant variance
- Normality - the errors ε_i are normally distributed

Residuals (vs) Fitted Values

Simple scatter plots of x (vs) y during the initial data analysis phase will give an indication of whether the relationship between x and y is linear and whether the variance is constant. After fitting the model, this can be checked by examining a plot of the residuals $(\widehat{\varepsilon}_i)$ (vs) the fitted values $(\widehat{y}_i)$. The residuals are calculated by subtracting the fitted values $\widehat{y}_i$ from the observed values y_i $(\widehat{\varepsilon}_i = y_i - \widehat{y}_i)$. Each residual is then plotted against the fitted value to see if there are any patterns.

Note that we denote, the errors by ε_i, the errors represent the difference between the observed values and the unknown true model.
Note that we denote, the residuals by $\widehat{\varepsilon}_i$, the residuals represent the difference between the observed values and the fitted model.

Example 6 *The data in the table below shows how plant height (cm) varies with the age of the plant (weeks):*

Age (weeks)	*2*	*2*	*3*	*4*	*4*	*5*	*6*	*7*	*8*	*8*
Height (cm)	*4*	*5*	*9*	*11*	*12*	*14*	*17*	*21*	*22*	*24*

The following linear model was fitted to the data: $y = -1.0299 + 3.0469x$, *The fitted values and the residual values are:*

Fitted	*5.06*	*5.06*	*8.11*	*11.16*	*11.16*	*14.20*	*17.25*	*20.30*	*23.34*	*23.34*
Residual	*-1.06*	*-0.06*	*0.89*	*-0.16*	*0.84*	*-0.20*	*-0.25*	*0.70*	*-1.34*	*0.66*

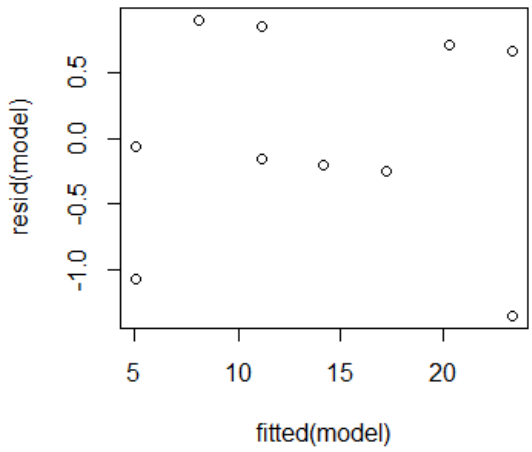


Figure 1

Plotting the residuals (vs) fitted values will reveal any **correlated error terms** or **heteroscedasticity**. Correlated error terms mean that there is some sort of pattern to the error terms caused by an unexplained relationship between the explanatory variable(s) and the response variable (i.e. nonlinearities). In Fig. 1 above, the residuals look reasonably random over the whole plotting region so we conclude that the error terms are uncorrelated.

The data below shows a squared relationship between the explanatory variable x and the response variable y, nevertheless we will model the relationship using a linear model, see Fig. 2. The plot of the residuals (vs) fitted values is shown in Fig. 3. Are the error terms uncorrelated?

x	1	2	3	4	5	6	7	8	9	10
y	1	4	9	16	25	36	49	64	81	100

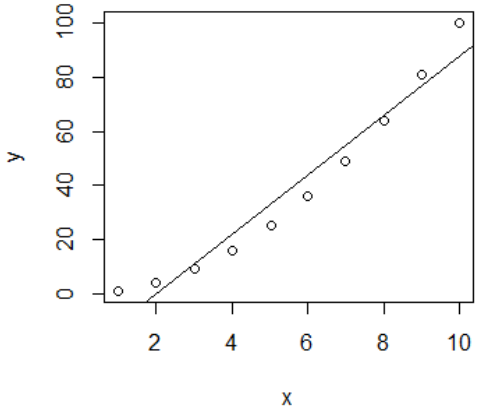


Figure 2

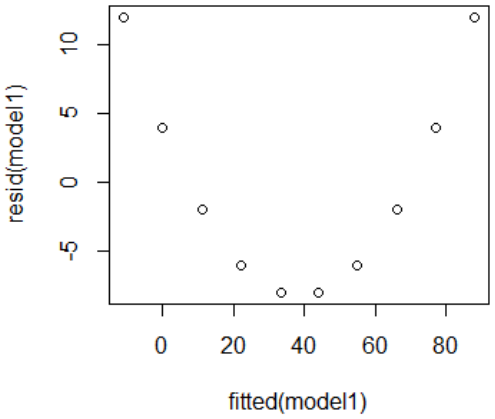


Figure 3

Heteroscedasticity occurs when the variance of the response variable is not constant with respect to the explanatory variable(s). If the relationship is heteroscedastic then the error terms (residuals) will not have constant variance (see Fig 4(b)).

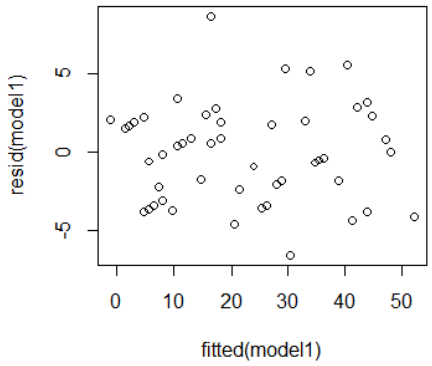


Figure 4(a)

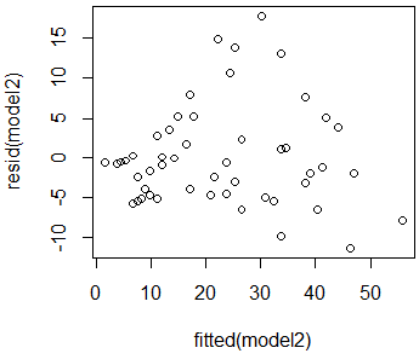


Figure 4(b)

Q-Q Plots and Histograms To check whether the error terms are normally distributed we can look at a Q-Q plot of the residuals and a histogram of the residuals.

Residuals and Influence An influential observation is one which greatly affects the regression line, e.g. see Figs 5(a), 5(b) and Fig 6(a), 6(b).

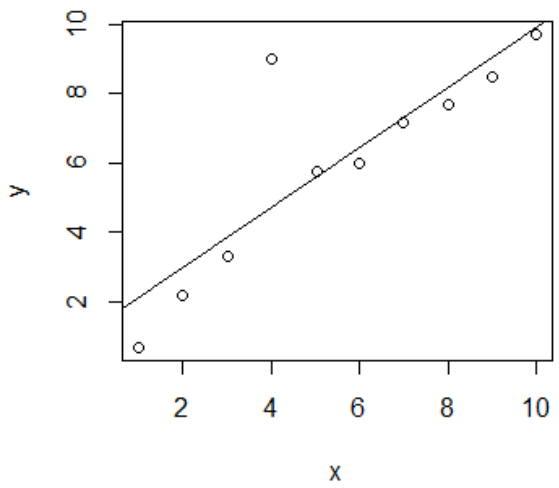


Figure 5(a)

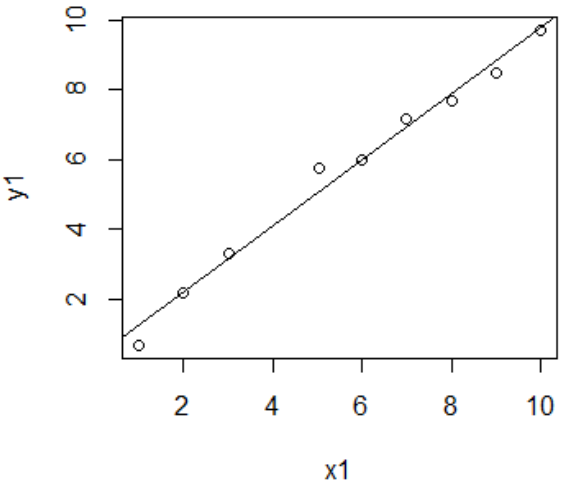
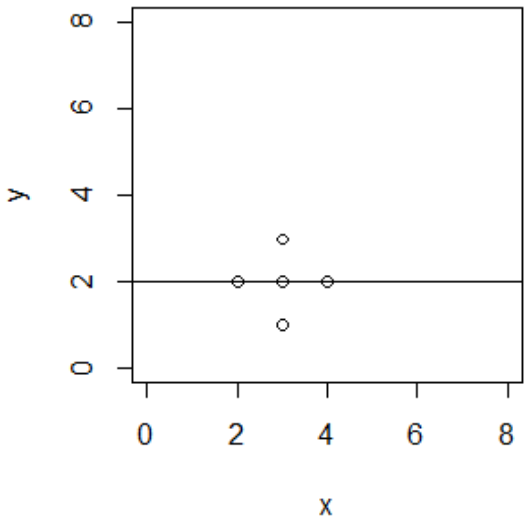
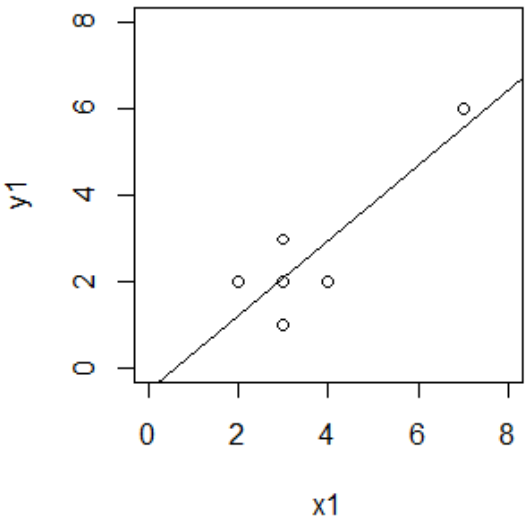


Figure 5(b)

Analysis of the residuals (vs) fitted values is a poor way of looking for influence. Precisely because a point is highly influential, it forces the regression line close to it, and hence the influential point may have a very small residual see Fig 6(a) and (b).





Figures 6(a) and 6(b)

In Fig 6(a) there is no significant relationship between x and y but the inclusion of the outlier observation $(7, 6)$ gives a significant regression between x and y. The outlier is said to be highly influential. This makes the write up more complicated. We must show that the significant relationship depends on the single observation $(7, 6)$ which requires an explanation of both models.

We may not achieve the best model in the first analysis. Unexplained variation, violation of assumptions and influential observations may be revealed in the model checking process and this information can indicate to the researcher how to improve the fit of the model.

Generalise to the Population?

It is possible that the final model will fit the sample data well but will not be generalisable to the population. Ideally, we need a second set of data to test the accuracy of the model, which is not always possible. If there is enough data, the sample can be split into two subsamples, one to estimate the model (the training data set) and one to estimate the accuracy of the model (the test data set). If there is not enough data to test the model on a separate data set, then bootstrapping or cross validation can be used to test the accuracy of the model coefficients. Bootstrapping draws a large number of subsamples (sampling with replacement) from the data set and for each subsample, a model is estimated. The coefficients associated with each of the submodels can then be averaged. Crossvalidation divides the data into m parts, equal (or close to) in size. For each part we use the rest of the data as a training set and that part as the test set.