Contents

1 Inference Procedures

Key Points:

- The two main types of inference procedures are **Hypothesis Tests** and **Confidence Intervals**. You are expected to be familiar with both.
- There are two ways of conducting a hypothesis test. One method is to compute the test statistic, and compare to the critical values.
- The second method is to compute the probability value (i.e. p-value), and compare it to the significance level. Nearly all computer programs use the p-value approach. In this course we will focus on the p-value approach.

1.1 Significance Level

In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis. The significance level is used in hypothesis testing as follows: First, the difference between the results of the experiment and the null hypothesis is determined. Then, assuming the null hypothesis is true, the probability of a difference that large or larger is computed . Finally, this probability is compared to the significance level. If the probability is less than or equal to the significance level, then the null hypothesis is rejected and the outcome is said to be statistically significant.

Traditionally, experimenters have used either the 0.05 level (sometimes called the 5% level) or the 0.01 level (1% level), although the choice of levels is largely subjective. The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 0.01 level is more conservative than the 0.05 level. The Greek letter alpha (α) is sometimes used to indicate the significance level

1.2 The probability value

The probability value (sometimes called the $p-value$) is the probability of obtaining a statistic as different from or more different from the parameter specified in the null hypothesis as the statistic obtained in the experiment.

1.2.1 The precise meaning of the p-value

There is often confusion about the precise meaning of the probability computed in a significance test. The convention in hypothesis testing is that the null hypothesis (Ho) is assumed to be true.

The difference between the statistic computed in the sample and the parameter specified by the null hypothesis is computed and the probability of obtaining a difference this large or large is calculated. This probability value is the probability of obtaining data as extreme or more extreme than the current data (assuming the null hypothesis is true).

It is not the probability of the null hypothesis itself. Thus, if the probability value is 0.005, this does not mean that the probability that the null hypothesis is either true or false is 0.005. It means that the probability of obtaining data as different or more different from the null hypothesis as those obtained in the experiment is 0.005.

The inferential step to conclude that the null hypothesis is false goes as follows: The data (or data more extreme) are very unlikely given that the null hypothesis is true. This means that:

(1) a very unlikely event occurred or

(2) the null hypothesis is false.

The inference usually made is that the null hypothesis is false. (Importantly it doesn't prove the null hypothesis to be false)

1.3 Using p-values to reject the null hypothesis

According to one view of hypothesis testing, the significance level should be specified before any statistical calculations are performed. Then, when the p-value is computed from a significance test, it is compared with the significance level.

The null hypothesis is rejected if p-value is at or below the significance level; it is not rejected if p-value is above the significance level. The degree to which p ends up being above or below the significance level does not matter.The null hypothesis either is or is not rejected at the previously stated significance level.

Thus, if an experimenter originally stated that he or she was using the $= 0.05$ significance level and p-value was subsequently calculated to be 0.042, then the person would reject the null hypothesis at the 0.05 level. If p-value had been 0.0001 instead of 0.042 then the null hypothesis would still be rejected at the 0.05 significance level.

The experimenter would not have any basis to be more confident that the null hypothesis was false with a p-value of 0.0001 than with a p-value of 0.041. Similarly, if the p had been 0.051 then the experimenter would fail to reject the null hypothesis

The experimenter would have no more basis to doubt the validity of the null hypothesis than if p-value had been 0.482. The conclusion would be that the null hypothesis could not be rejected at the 0.05 level.

In short, this approach is to specify the significance level in advance and use p-value only to determine whether or not the null hypothesis can be rejected at the stated significance level.

Many statisticians and researchers find this approach to hypothesis testing not only too rigid, but basically illogical. It is very reasonable to have more confidence that the null hypothesis is false with a p-value of 0.0001 then with a p-value of 0.042?

The less likely the obtained results (or more extreme results) under the null hypothesis, the more confident one should be that the null hypothesis is false.

The null hypothesis should not be rejected once and for all. The possibility that it was falsely rejected is always present, and, all else being equal, the lower the p-value, the lower this possibility.

According to this view, research reports should not contain the p-value, only whether or not the values were significant (at or below the significance level).

However it is much more reasonable to just report the p-values. That way each reader can make up his or her mind about just how convinced they are that the null hypothesis is false.

1.3.1 Guidelines for Data Project

For this module, as a rule of thumb, we will use the threshold of 0.01 for rejecting the null hypothesis. If the p-value is less than 0.01 we reject the null hypothesis. If it is greater than

0.05, we fail to reject the null hypothesis. If between the two, consider it to be a 'grey area'. (i.e. suggest that more data is needed).

If the p-value is greater than 0.1 we would never reject the null hypothesis.

- Greater than 0.05 - Fail to reject Ho
- Less than 0.01 - Reject Ho
- Between 0.01 and 0.05 - advise that it is close to both conclusions.

Many R outputs will give a group of asterisks beside the data to help the user in interpreting the data, depending on how significant the result is.

```
p-value  < 0.0001   ***
p-value  < 0.001 **
p-value  < 0.01 *
p-value  < 0.1
```

1.4 Sample Size

For Student's t distribution, statistical tables such (e.g. Murdoch Barnes and State Examinations Commission tables) only tabulate quantiles with degrees of freedom of less than 30. This restraint has given rise to the convention that a sample of size greater than 30 is a 'large sample' and in this case the standard normal distribution should be used.

However there is a disparity between the Z value and the correct t value. For a sample size of 61 (i.e. degrees of freedom =60), the 97.5% t-quantiles of Student's t distribution is 2.003, and not 1.96.

However, statistical software is free from this restraint. The correct distribution will be automatically used. The Student's t distribution will be used in all appropriate cases. As the sample size increases the Student t distribution converges with the standard normal distribution.

1.5 Commonly Used Inference Procedures

- Hypothesis test for the mean of a single sample
- Hypothesis test for the mean of two independent samples
- Hypothesis test for the proportion of a single group
- Hypothesis test for the proportions of two independent samples

2 Testing The Assumption of Normality

For example, a fundamental assumption of linear models (i.e. regression models) is that the residuals (differences between observed and predicted value) are normally distributed with mean zero.

The null hypothesis of both the 'Anderson-Darling' and 'Shapiro-Wilk' tests is that the population is normally distributed, and the alternative hypothesis is that the data is not normally distributed.

For both tests, the null and alternative hypothesis are :
H_0 : The data set is normally distributed.
H_1 : The data set is **not** normally distributed.

2.1 Anderson-Darling Test

To implement the Anderson-Darling Test for Normality, one must first install the ***nortest*** package.

```
library(nortest)
#Generate 100 normally distributed random numbers
NormDat = rnorm(100)
ad.test(NormDat)
```

2.2 Shapiro-Wilk Test

The Shapiro-Wilk test is directly implementable, without loading any additional packages.

```
#Generate 100 normally distributed random numbers

NormDat = rnorm(100)

shapiro.test(NormDat)
```

Sample output, using the randomly generated `NormDat` data set, is as follows:

```
> shapiro.test(NormDat)

        Shapiro-Wilk normality test

data:  NormDat
W = 0.9864, p-value = 0.4003
```

Here, the p-value is well above the 0.05 threshold. Hence we **fail to reject** the null hypothesis, and may proceed to treat the `NormDat` data set as normally distributed.

2.3 Graphical Procedures for Assessing Normality

There are two useful graphical methods for determining whether a data set was normally distributed. The first is the histogram, which we have seen previously. If the histogram is reasonably bell-shaped, then the data can be assumed to be normally distributed. The relevant R command is `hist()`.

The second is the ***quantile-quantile plot*** (or QQ-plot). For assessing normality, we implement a qq-plot using the `qqnorm()` function.

Additionally the command `qqline()` function adds a trendline to a normal quantile-quantile plot. If the data is normally distributed, then the points on the plot follow the trendline.

```
#Generate 100 normally distributed random numbers

NormDat = rnorm(100)

qqnorm(NormDat)
qqline(NormDat)
```

2.4 Transforming the Data

Sometimes when we get non-normal data, we can change the scale of our data i.e. transform it to get a normal distribution. One transformation that often works for positively skewed data is the natural logarithm (ln) transformation.

In such a case, we work with the natural logarithms of the data set, rather than the data itself.

2.5 Outliers

Another reason that the data may not be normally distributed is the presence of an outlier. We shall look at formal tests for outliers (such as the Grubb's test) next week. Recall that boxplots can be used to detect potential outliers.

3 Single Sample Inference Procedures

While analyzing a (single) sample of data values, we will often want to answer several questions:

- What is the mean value? (i.e. of 100 roles of a die)
- Is the mean value significantly different from some pre-supposed value? (i.e. Hypothesis testing ; is the observed mean reasonably close to our expected value)
- What is the level of uncertainty associated with our estimate of the mean value? (i.e. Confidence interval for the estimates)

We refer to the methods used to answer these questions as ***inference procedures***.

3.1 Hypothesis test for the mean of a single sample

This procedure is used to assess whether the population mean μ has a specified value, based on the sample mean. The hypotheses are conventionally written in a form similar to below (here the hypothesized population mean is simply zero).

Ho : $\mu = 0$

Ha : $\mu \neq 0$

There are two hypothesis test for the mean of a single sample.

1) The sample is of a normally-distributed variable for which the population standard deviation (σ) is known.

2) The sample is of a normally-distributed variable where σ is estimated by the sample standard deviation (s).

In practice, the population parameter values is rarely known. For this reason, we will consider the second case only in this course.

3.2 The t.test() Function

The `t.test( )`function produces a variety of outputs for procedures, hence answering such questions. Let us look at the function first to see what sort of output it gives us. Recall our simulated data from last week, rolling a die 100 times.

```
## Initialize variables
die = 1:6
N=100

## Calculations
x=sample(die,N, replace=TRUE)
t.test(x)
```

The `R` output should look something like this:

```
        One Sample t-test
data:  x
t = 19.8867, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.087767 3.772233
sample estimates:
mean of x :     3.43
```

Working backwards, the t procedure gives us the mean of the data set, and a 95% confidence interval for that mean. (Please refer to previous modules)

The previous statement to these refers to the alternative hypothesis: true mean is not equal to zero. Necessarily the null hypothesis (which proposes the opposite of the alternative hypothesis) states that the true mean is zero. Is this meaningful in such circumstances? No, it is impossible

Lets look at the help file to get a clearer idea about how to use this command.

The default setting of the null value is zero (i.e. `mu = 0`. To assess whether our data set is fair, we must specify `mu = 3.5`.

We can change the confidence level (and by extension, the significance level) by specifying the desired value using the `conf.level = `.

In this instance, we will change it to 0.99 (i.e. 99% confidence). However for the rest of the module, we will use the 95% confidence level. (Remark: Be mindful that I might ask this in the exam.).

```
t.test(x,mu=3.5,conf.level=0.99)
```

```
        One Sample t-test
data:  x
t = -0.4059, df = 99, p-value = 0.6857
alternative hypothesis: true mean is not equal to 3.5
99 percent confidence interval:
 2.977004 3.882996
sample estimates:
mean of x
     3.43
```

The mean value remains constant. We have specified a 99% confidence level, so necessarily a 99% confidence interval is returned. Compare it to the previous output.

The important difference is what is specified as the alternative hypothesis: `true mean is not equal to 3.5`. Necessarily the null hypothesis is that true mean is equal to 3.5 (i.e. a fair dice). This is meaningful in the contest of the dice-roll experiment.

In this instance the p-value is 0.6857. We fail to reject the null hypothesis that the mean is not 3.5. We can use the confidence interval to make an inference. Consider the 95% confidence interval for the mean value (from earlier) : `(3.087767,3.4372233)`

As the expected mean (i.e. the null value) is within this interval, we would fail to reject the null hypothesis. Suppose the 95% confidence interval returned the following limits, with other values being computed accordingly; `(3.087767,3.4372233)`.

What would be the decision in this case? We would reject the null hypothesis that the mean value is 3.5, and surmise that the dice is a crooked dice that favours lower values.

4 Two Sample Inference Procedures

4.1 Hypothesis test for the means of two independent samples

The procedure associated with testing a hypothesis concerning the difference between two population means is similar to that for testing a hypothesis concerning the value of one population mean.

The procedure differs mainly in that the standard error of the difference between the means is used to determine the test statistic associated with the sample result. For two tailed tests, the null hypothesis states that the population means are the same, with the alternative stating that the population means are not equal.

Ho : $\mu_1 = \mu_2$

Ha : $\mu_1 \neq \mu_2$

4.2 Implementation with R

Firstly, lets construct a second data set. In this scenario, it is not possible to score a 6, hence the dice is crooked. (The previous fair dice data set is called x. This crooked dice data is labelled y)

```
## Initialize variables
die2 = 1:5
N=100
y=sample(die2,N, replace=TRUE)
t.test(y)
```

We can perform a two sample test for independent samples. In such a test the null and alternative hypotheses are as follows:
H0: True mean of x is equal to true mean of y.
H1: True mean of x is NOT equal to true mean of y.
An estimate for the difference of sample means, and a confidence interval for that estimate is provided in the output. The expected value under the null hypothesis does not have to be specified in this instance.

```
## Initialize variables
die2 = 1:5
N=100
y=sample(die2,N, replace=TRUE)
t.test(y)
```

To implement a two sample test, simply specify the names of both data sets.

```
t.test(x,y)
```

The output should look something like the output below. Notice the confidence interval for the difference in the means: `( -0.01885674,0.83885674 )`. How do you interpret this output? (Hint: look at the p-value).

```
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 1.8862, df = 183.43, p-value = 0.06084
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01885674  0.83885674
sample estimates:
mean of x mean of y
     3.39      2.98
```

There are in fact two variants of the two sample t-test.

- The Independent Two Sample t-test
- The Welch Two Sample t-test

The Welch Two-Sample test (the procedure from the last segment of Routput) does not require the assumption of equal variance in the two samples. Conversely the Independent Two-Sample test does.

To specify that the assumption of equal variance, the additional argument `var.equal=TRUE` is specified

```
t.test(x,y,var.equal=TRUE)
```

```
> t.test(x,y,var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = 1.8862, df = 198, p-value = 0.06073
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01864727  0.83864727
sample estimates:
mean of x mean of y
     3.39      2.98
```

4.3 Equality of Variances

It is possible to formally test whether or not there is equality of variance in two data sets, using the F-test. The null hypothesis states that there is equal variance between samples. The alternative is that they do not have equal variance.

Ho $\sigma_1^2 = \sigma_2^2$

Ha $\sigma_1^2 \neq \sigma_2^2$

The command is `var.test()`.Variant specifications of the inference procedure, such as confidence level, can be altered as with the `t.test()` procedure.

```
var.test(x,y)
```

```
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 1.7849, num df = 99, denom df = 99, p-value = 0.004299
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.200948 2.652763
sample estimates:
ratio of variances
          1.784889
```

Notice the reference to the ***variance ratio***. The test actually works on the following basis.

Ho $\frac{\sigma_1^2}{\sigma_2^2} = 1$

Ha $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

A variance ratio of 1 is equivalent to equal variance.

4.4 Paired t-test

Two data samples aresaid to be paired (or matched) if they come from repeated observations of the same subject. Here, we assume that the data populations follow the normal distribution.

Using the paired t-test, we can obtain an interval estimate of the difference of the population means. Necessarily there must be equal numbers of elements in both sets.

The `t.test()` function can be used to perform paired t-tests, by making the appropriate specification:`paired=TRUE`.

4.4.1 Example

In the built-in data set named ***immer***, the barley yield in years 1931 and 1932 of the same field are recorded. In the intervening period, fertilizer treatments were applied to each field. The motivation of the study was to determine whether or not the treatment was effective.

The yield data are presented in the data frame columns $Y1$ and $Y2$.

```
> library(MASS)         # load the MASS package
> head(immer)
   Loc  Var    Y1    Y2
1  UF   M      81.0  80.7
2  UF   S      105.4  82.3
   .....
```

Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields.

We apply the t.test function to compute the difference in means of the matched samples. As it is a paired test, we set the "paired" argument as TRUE.

```
attach(immer)
t.test(Y1,Y2, paired=TRUE)
```

```
        Paired t-test

data:  Y1 and Y2
t = 3.324, df = 29, p-value = 0.002413
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.121954 25.704713
sample estimates:
mean of the differences
               15.91333
```

Between years 1931 and 1932 in the data set immer, the 95% confidence interval of the difference in means of the barley yields is the interval between 6.122 and 25.705. One can conclude that the fertilizer treatments were successful in improving the yield of barley.

5 Non-Parametric Inference Procedures

Nonparametric procedures were developed to be used in cases when the distribution of the variable of interest in the population is known to be not-normal, and furthermore the distribution is undetermined (hence the name nonparametric).

Nonparametric tests are also referred to as ***distribution-free*** tests. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and, as a result, if the data have one or two outliers, their influence is negated.

Parametric tests are preferred because, in general, for the same number of observations, they are more likely to lead to the rejection of a false hull hypothesis. That is, they have more power. This greater power stems from the fact that if the data have been collected at an interval or ratio level, information is lost in the conversion to ranked data (i.e., merely ordering the data from the lowest to the highest value).

- Kolmogorov- Smirnov Test (`ks.test()`)
- Wilcoxon test (`wilcox.test()`)

5.1 Kolmogorov-Smirnov Test

For a single sample of data, the Kolmogorov-Smirnov test is used to test whether or not the sample of data is consistent with a specified distribution function. (Not part of this course) When there are two samples of data, it is used to test whether or not these two samples may reasonably be assumed to come from the same distribution. The null and alternative hypotheses are as follows:
H0: The two data sets are from the same distribution
H1: The data sets are not from the same distribution

Consider two sample data sets X and Y that are bothnormally distributed with similar means and variances.

```
> X=rnorm(16,mean=20,sd=5)
> Y=rnorm(18,mean=21,sd=4)
> ks.test(X,Y)

        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.2153, p-value = 0.7348
alternative hypothesis: two-sided
```

Remark: It doesnt not suffice that both datasets are from the same distribution. They must have the same value for the defining parameters. Consider the case of data sets; X and Z. Both are normally distributed, but with different mean values.

```
> X=rnorm(16,mean=20,sd=5)
> Z=rnorm(16,mean=14,sd=5)
> ks.test(X,Z)

        Two-sample Kolmogorov-Smirnov test

data:  X and Z
D = 0.5625, p-value = 0.0112
alternative hypothesis: two-sided
```

5.2 Wilcoxon Mann-Whitney Test

The Wilcoxon Mann-Whitney Test is one of the most powerful of the nonparametric tests for comparing two populations. It is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to ***location*** (i.e. median), if at all.

The Wilcoxon Mann-Whitney test does not require the assumption that the differences between the two samples are normally distributed.

In many applications, the Wilcoxon Mann-Whitney Test is used in place of the two sample t-test when the normality assumption is questionable.

This test can also be applied when the observations in a sample of data are ranks, that is, ordinal data rather than direct measurements.

6 Bivariate data

6.1 What is Bivariate data?

A dataset with two variables contains what is called bivariate data. For example, the heights and weights of people (i.e. for the purposes of determining the extent to which taller people weigh more)

Common bivariate statistical analyses include

- Correlation
- Simple Linear Regression

6.2 Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables.

Scatterplots can be implemented in `R` using the command `plot()`

Exercise: Let us construct scatter-plots for the Immer and Iris data sets.

```
plot(immer$Y1,immer$Y2)

plot(iris[,1],iris[,3])
```

More complex scatterplots, with better visual aesthetics, can be constructed. We will look at this more later on in the semester.

6.3 Correlation

Recall that correlation describes the strength of a relationship between two numeric variables, and that the Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables.

It is referred to as Pearson’s correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson’s correlation is ”ρ” when it is measured in the population and `r` when it is measured in a sample.

As we will be dealing almost exclusively with samples, we will use `r` to to represent Pearson’s correlation unless otherwise noted.

Pearson’s r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an `r` of 0 indicates no linear relationship between variables, and an `r` of 1 indicates a perfect positive relationship between variables.

Importantly it is assumed that the relationship in question is supposed to be linear. Some variables will in fact have a non-linear relationship (more on that latet)

The relevant R command is `cor()`.

```
cor(immer$Y1,immer$Y2)

cor(iris[,1],iris[,3])
```

The strength of the relation is represented in a numeric value known at the correlation coefficient. This coefficient can take a value between -1 and1. Additionally there are no units.

Getting a correlation coefficient is generally only half the story; you will want to know if the relationship is significant. There is a more complex command called `cor.test()`. This command additionally provides a hypothesis test for the correlation estimate.

```
cor.test(immer$Y1,immer$Y2)

cor.test(iris[,1],iris[,3])
```

Ho : The correlation coefficient for the population of values is zero. (i.e. No linear relationship.)

Ha : The coefficient is not zero. (Linear relationship exists.)

A confidence interval for the coefficient is provided for in the `R` output. If the interval includes 0 then we fail to reject the null hypothesis.

Simple linear regression is used to describe the relationship between two variables x and y.

For example, you may want to describe the relationship between age and blood pressure or the relationship between scores in a midterm exam and scores in the final exam, etc.

- x is the independent (i.e. predictor) variable.
- y is the dependent (i.e. response) variable.

That is to say x is said to cause or influence y.

Necessarily both x and y should be of equal length. One of the first steps in a regression analysis is to determine if any kind of relationship exists between x and y.

A scatterplot can created and can initially be used to get an idea about the nature of the relationship between the variables, e.g. if the relationship is linear, curvilinear, or no relationship exists.

To make a simple scatter-plot, we simply use the `plot()` command. The independent variable (the variable to go along the x-axis) is always specified first.

```
X=c(5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97,5.94, 7.32, 6.64, 6.94, 3.51)

Y=c(5.56, 7.80, 6.13, 8.15, 7.95, 7.87, 8.03, 5.67, 7.11, 6.65, 7.02, 3.88)

plot(X,Y)
cor(X,Y)
```

In this case here, we can see from the scatter-plot that there is a linear relationship between x and y. Simple linear regression is only useful when there is evidence of a linear relationship. In other cases, such as quadratic relationships, other types of regression may be more appropriate.

6.4 Linear Regression Model

A linear relationship can be defined by the simple linear regression model

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

The intercept β_0 describes the point at which the line intersects the y axis. The slope β_1 describes the change in y for every unit increase in the predictor variable x.

From the data set, we determine the regression coefficients, i.e. estimates for slope and intercept. (N.B. There are variations on this notation in textbooks).

- b_0 : the intercept estimate.
- b_1 : the slope estimate.

Therefore the fitted model can be expressed as

$$\hat{y} = b_0 + b_1 x$$

Recall $\hat{y}$ denotes the predicted value for y, given some value x.

6.5 Fitting a Model with `R`

The `R` command `lm()` is used to fit linear models. Firstly the response variable y is specified, then the predictor variable x.

The tilde sign is used to denote the dependent relationship (i.e. y depends on x). The regression coefficients are then determined.

```
lm(Y~X) # y depends on X
```

The output will include the formula, and two coefficient terms

- The intercept estimate is recorded under (*Intercept*)

- The slope estimate is recorded under the name of the predictor variable (here : X).

```
Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     0.7812       0.8581
```

A more detailed data output (i.e. more than just the coefficients) is generated in the form of a data object, using the `summary()` command.

We can give a name to the model (e.g. $FIT1$), and view all of the results of the calculation, including the regression coefficients, hypothesis test results and information on the residuals (i.e. the differences between the estimated y values and the observed y values).

In common with all data structures we can use the `names()` function and $ to access components.

```
FIT1 = lm(Y~X)
summary(FIT1)
names(FIT1)
names(summary(FIT1))
FIT1$coefficients
class(FIT1)
```

6.6 Confidence Interval for Regression Estimate

To compute the confidence intervals for both estimates, we use the confint() command, specifying the name of the fitted model.

```
C=c(0,2,4,6,8,10,12)
F=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
Fit1=lm(F˜C)
coef(Fit1)
# (Intercept)        Conc
     1.517857    1.930357

confint(Fit1)
#               2.5 %   97.5 %
# (Intercept) 0.75970 2.276014
# Conc        1.82522 2.035495
```

6.7 The Coefficient of Determination

The coefficient of determination R^2 is the proportion of variability in a data set that is accounted for by the linear model.

Equivalently R^2 provides a measure of how well future outcomes are likely to be predicted by the model.

(For simple linear regression, it canbe computed by squaring the correlation coefficient.)

```
summary(fit1)$r.squared
```

7 Assessing Model Assumptions

7.1 Residuals

The difference between the predicted value (based on the regression equation) and the actual, observed value. In simple linear regression models, the matter of whether or not residuals are normally distributed often arises.

Additionally the expected value of the residuals should be zero.

We have seen previously two methodologies for determining whether or not a data set is normally distributed;

- Shapiro-Wilk tests (or Anderson-Darling test)
- QQ plots

We will explore this more in a forthcoming example.

7.2 Influence Analysis

7.2.1 Outlier

In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

7.2.2 Leverage

An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients.

7.2.3 Influence

An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

7.3 Example

A new hotel is built 15 miles from the location of a prominent annual sporting event. A study of the number of enquiries received by a random sample of 9 established hotels in the area showed that the number of enquiries and the distance in miles between the hotel and event. Here the independent variable is distance (x) and the dependent variable is number of enquiries.

Lets looks at the residuals, and assess whether they are normally distributed.

```
#enquiries
y=c(35,61,74,92,113,159,188,217,328)

#distance from hotel
x=c(28,20,17,12,16,8,2,3,1)
#

#fit the linear model
fit2=lm(y~x)
resid(fit2)
res.fit=resid(fit2)

# test the residuals for normality.
# Normal if p.value is high.
shapiro.test(res.fit)

qqnorm(res.fit) #QQ plot
qqline(res.fit) #Add Trendline


#Do all your analyses agree?
```

Lets look at the scatterplot of x and y (`plot(x,y)`). Does the first covariate seem to be an outlier, given that a linear model is assumed? Lets omit the first element of both data sets and run the analysis again.

```
fit2=lm(y[-1]~x[-1])
resid(fit2)
res.fit2=resid(fit2)

shapiro.test(res.fit2)

#test the residuals for normality. Normal if p.value is high.
qqnorm(res.fit2);  qqline(res.fit2)

# compare the coefficients of both models.
coef(fit1)
coef(fit2)
```

Does the covariate in question have high leverage or high influence?

Remark: Arguably it is a case that this problem is not best described by a simple linear regression model, and that a non-linear model would be more suitable.

7.4 Diagnostic Plots

Homoscedascity (constant variance) is one of the assumptions required in a regression analysis in order to make valid statistical inferences about population relationships.

Homoscedasticity requires that the variance of the residuals are constant for all fitted values, indicated by a uniform scatter or dispersion of data points about the trend line (i.e. "The Zero Line").

From the above plot, we can conclude that the constant variance assumption is valid. We can see that the mean value of the residuals is zero.

```
plot(fit1)
#Four Diagnostic Plots are printed to screen sequentially.
```

8 Multiple Linear Regression

In your future studies, you will come across multiple linear regression (MLR). This is a linear model uses multiple independent variables to explain a single dependent variable.

The implementation is very similar to simple linear regression (SLR). All that is required is to specify the additional independent variables.

```
Fit.slr =lm(y~x)    # SLR: y explained by predictor x
Fit.mlr=lm(y~x+z)   # MLR: y explained by predictors x and z
```

For this case, a linear relationship can be defined by the regression model

$$y = \beta_0 + \beta + 1x + \beta_2 z + \epsilon$$

.

Again, we determine the regression coefficients, i.e. estimates for slopes and intercept. (N.B. There are variations on this notation).

- b_0 : the intercept estimate.
- b_1 : the slope estimate for X
- b_2 : the slope estimate for z

In many project datasets it is possible to implement a MLR model. For the moment, we will just look at slope and intercept estimates, their p-values and the coefficient of determination.

Let try this out using the ***iris*** data set. (This is not be a valid statistical analysis in practice. However we are focussing on the mechanics, so we shall proceed nonetheless).

```
lm(Sepal.Length ~ Sepal.Width + Petal.Width)
```

8.1 Model Selection

There are many important methodologies for determining which combination of predictor variables bests describes a response variable. You will meet this in future modules. We will use two simple ones for this module only.

- Adjusted Rsquared value
- The Akaike Information Criterion (AIC)

The adjusted R-square value is found on the summary output for a fitted model. It is called ***adjusted*** because it takes into account the number of predictor variables being used. The law of parsimony states the simplest model that adequately explains the outcomes is the best. The candidate model with the higher adjusted R squared is considered preferable.

The AIC is a model selection metric often used in statistics.It is computed using the R command `AIC()`.The candidate model with the smallest AIC value is considered preferable.

```
fitA = lm(Sepal.Length ~ Sepal.Width + Petal.Width)
fitB = lm(Sepal.Length ~ Sepal.Width + Petal.Length)

summary(fitA)$adj.r.squared
summary(fitB)$adj.r.squared

AIC(fitA)
AIC(fitB)
```