GROUP 1

Analysis of Variance (ANOVA)

*The analysis of variance is a commonly used method to determine differences between several samples. R provides a function to conduct ANOV , aov(model, data).*

*The model can take a variety of forms:*

|  |  |
| --- | --- |
| *Model* | *Explanation* |
| *y ~ x1* | *y is explained by x1 only, a one-way ANOVA* |
| *y ~ x1 + x2* | *y is explained by x1 and x2, a two-way ANOVA* |
| *y ~ x1 + x2 + x3* | *y is explained by x1, x2 and x3, a three-way ANOVA* |
| *y ~ x1 \* x2* | *y is explained by x1, x2 and also by the interaction between them* |

*The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses.*

*a. Independence of observations - this is an assumption of the model that simplifies the statistical analysis*

*b. Normality - the distributions of the residuals are normal*

*c. Homogeneity of variances, called homoscedasticity - the variance of data in groups should be the same.*

***Example:***

*To test the lifetime of batteries, 12 toy drummers are fitted with new batteries of three types: Amazing, Superlong, Endurance.*

*The lengths of time (in hours) that the drummers continue to drum are summarized in the table below:*

*Amazing 4.7, 5.1, 5.2*

*Superlong 4.8, 5.1, 5.4, 5.4*

*Endurance 5.1, 5.2, 5.2, 5.4, 5.6*

*Determine whether there is significant evidence, at the 5 % level, of a difference between the mean lifetimes of the three types of batteries:*

*Summarize the findings in the* ***ANOVA*** *table.*

*We will start with a simple function which requires no other packages than stats.*

*The function for the analysis of variance in stats is* ***aov****.*

*Description: Fit an analysis of variance model by a call to lm for each stratum.*

*Syntax aov(formula, data = NULL)*

*where*

* *formula Indicates a formula specifying the model*
* *data indicates a data frame in which the variables specified in the formula will be found*

*A formula is typically of the form: y ~ model, where y is the analyzed response and model is a set of terms for which some parameters are to be estimated.*

*The function summary returns detailed results.*

1. Open RGui.
2. Try to run the program group1\_example\_1.R which uses the data file, dataset\_batterylife.xlsx to solve this problem.
3. Try more examples in the document “R Programming\_Lessons\_7.pdf” from pages 41 to 47.

GROUP 2 Student t Tests

Refer to pages 15 to 21 in the PDF file R\_Programming\_Lessons\_7.

A t-test is used to test hypotheses about the mean value of a population from which a sample is drawn. A t-test is suitable if the data is believed to be drawn from a normal distribution, or if the sample size is large.

The function t.test() can be used to perform both one and two sample t-tests on vectors of data. The function contains a variety of options and can be called as follows:

t.test(x,y = NULL, alternative = c("two sided","less","greater"), mu= 0, paired= FALSE, var.equal= FALSE, conf.level = 0.95)

Here, x is a numeric vector of data values and y is an optional numeric vector of data values. If y is excluded, the function performs a one-sample t-test on the data contained in x, if it is included it performs a two-sample t-tests using both x and y.

The option mu provides a number indicating the true value of the mean (or difference in means if you are performing a two-sample test) under the null hypothesis. The option alternative is a character string specifying the alternative hypothesis, and must be one of the following: "two sided" (which is default), "greater" or "less" depending on whether the alternative hypothesis is that the mean is different than, greater than or less than mu, respectively.

The option paired indicates whether or not you want a paired t-test (TRUE= yes and FALSE = no). If you leave this option out it defaults to FALSE.

The option var.equal is a logical variable indicating whether or not to assume the two variances as being equal when performing a two-sample t-test. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degree of freedom is used. If you leave this option out it defaults to FALSE.

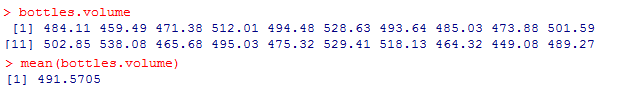
Finally, the option conf.level determines the confidence level of the reported confidence interval for µ0 in the one-sample case and µ1- µ2 in the two-sample case.



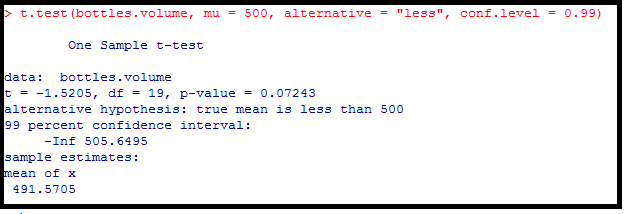
## Example: One-tailed, one-sample t-test

A bottle filling machine is set to fill bottles with drinking water to a volume of 500 ml. The actual volume is known to follow a normal distribution. The manufacturer believes the machine is under filling bottles. A sample of 20 bottles is taken and the volume of liquid is measured. The results are given in the bottles dataset, which is available here.

bottles.volume <- c(484.11, 459.49, 471.38, 512.01, 494.48, 528.63, 493.64, 485.03, 473.88, 501.59, 502.85, 538.08, 465.68, 495.03, 475.32, 529.41, 518.13, 464.32, 449.08, 489.27)



Suppose you want to use a one-sample t-test to determine whether the bottles are being consistently under-filled, or whether the low mean volume for the sample is purely the result of random variation. A one-sided test is suitable because the manufacturer is specifically interested in knowing whether the volume is less than 500 ml. The test has the null hypothesis that the mean filling volume is equal to 500 ml, and the alternative hypothesis that the mean filling volume is less than 500 ml. A significance level of 0.01 is to be used.



From the output, we can see that the mean bottle volume for sample is 491.6 ml. The one-sided 99 % confidence interval tells us that the mean filling volume is likely to be less than 505.6 ml. The p-value of 0.07243 tells us that if the mean filling volume of the machine were 500 ml, the probability of selecting a sample with a mean volume less than or equal to this one would be approximately 7 %.

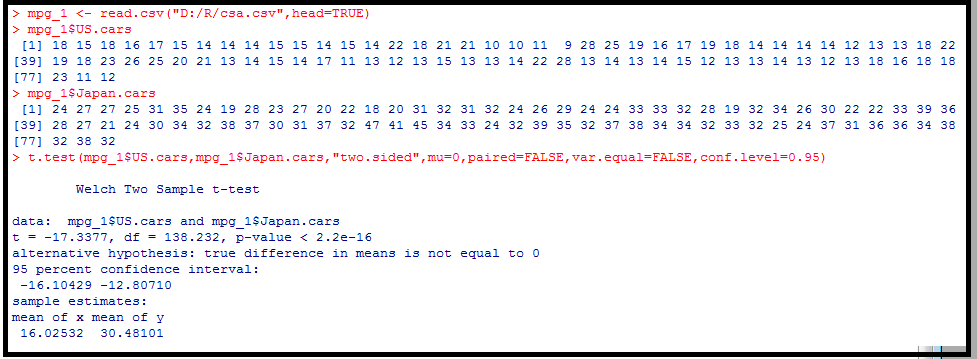
Since the p-value is not less than the significance level of 0.01, we cannot reject the null hypothesis that the mean filling volume is equal to 500 ml. This means that there is no evidence that the bottles are being under-filled.

## Example: two-sample t-test

The following is the data used for the two-sample t-test example. The first column is miles per gallon for US cars and the second column is miles per gallon for Japanese cars.

We are testing the hypothesis that the population means are equal for the two samples. We assume that the variances for the two samples are equal.

| **#** | **Mpg of US Cars** | **Mpg of Japan cars** |
| --- | --- | --- |
| 1 | 18 | 24 |
| 2 | 15 | 27 |
| 3 | 18 | 27 |
| 4 | 16 | 25 |
| 5 | 17 | 31 |
| 6 | 15 | 35 |
| 7 | 14 | 24 |
| 8 | 14 | 19 |
| 9 | 14 | 28 |
| 10 | 15 | 23 |
| 11 | 15 | 27 |
| 12 | 14 | 20 |
| 13 | 15 | 22 |
| 14 | 14 | 18 |
| 15 | 22 | 20 |
| 16 | 18 | 31 |
| 17 | 21 | 32 |
| 18 | 21 | 31 |
| 19 | 10 | 32 |
| 20 | 10 | 24 |
| 21 | 11 | 26 |
| 22 | 9 | 29 |
| 23 | 28 | 24 |
| 24 | 25 | 24 |
| 25 | 19 | 33 |
| 26 | 16 | 33 |
| 27 | 17 | 32 |
| 28 | 19 | 28 |
| 29 | 18 | 19 |
| 30 | 14 | 32 |
| 31 | 14 | 34 |
| 32 | 14 | 26 |
| 33 | 14 | 30 |
| 34 | 12 | 22 |
| 35 | 13 | 22 |
| 36 | 13 | 33 |
| 37 | 18 | 39 |
| 38 | 22 | 36 |
| 39 | 19 | 28 |
| 40 | 18 | 27 |
| 41 | 23 | 21 |
| 42 | 26 | 24 |
| 43 | 25 | 30 |
| 44 | 20 | 34 |
| 45 | 21 | 32 |
| 46 | 13 | 38 |
| 47 | 14 | 37 |
| 48 | 15 | 30 |
| 49 | 14 | 31 |
| 50 | 17 | 37 |
| 51 | 11 | 32 |
| 52 | 13 | 47 |
| 53 | 12 | 41 |
| 54 | 13 | 45 |
| 55 | 15 | 34 |
| 56 | 13 | 33 |
| 57 | 13 | 24 |
| 58 | 14 | 32 |
| 59 | 22 | 39 |
| 60 | 28 | 35 |
| 61 | 13 | 32 |
| 62 | 14 | 37 |
| 63 | 13 | 38 |
| 64 | 14 | 34 |
| 65 | 15 | 34 |
| 66 | 12 | 32 |
| 67 | 13 | 33 |
| 68 | 13 | 32 |
| 69 | 14 | 25 |
| 70 | 13 | 24 |
| 71 | 12 | 37 |
| 72 | 13 | 31 |
| 73 | 18 | 36 |
| 74 | 16 | 36 |
| 75 | 18 | 34 |
| 76 | 18 | 38 |
| 77 | 23 | 32 |
| 78 | 11 | 38 |
| 79 | 12 | 32 |



H0: µ1 = µ2

H1: µ1 != µ2

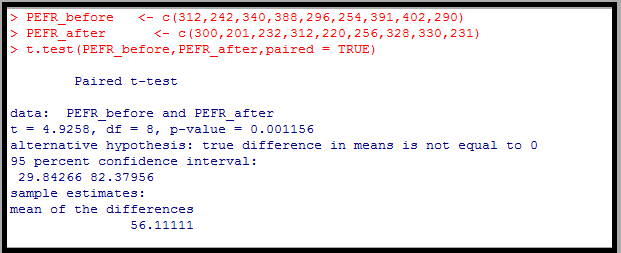
*Since the p-value is very much less than the significance level of 0.05, we* ***reject*** *the null hypothesis and conclude that the* ***two population means are different*** *at the 0.05 significance level.*

## Example: Paired t-test

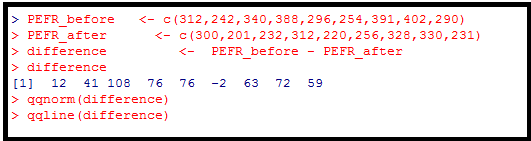
Nine asthmatic patients are randomly selected for a walk on a cold winter day. Comparison of peak expiratory flow rate before and after a walk on a cold winter's day is made.

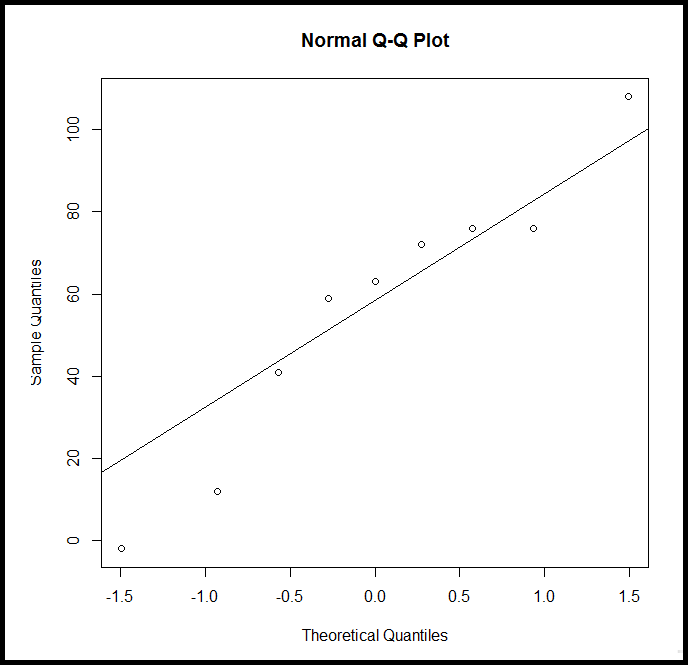
|  |  |  |
| --- | --- | --- |
| Patient | Before | After |
| 1 | 312 | 300 |
| 2 | 242 | 201 |
| 3 | 340 | 232 |
| 4 | 388 | 312 |
| 5 | 296 | 220 |
| 6 | 254 | 256 |
| 7 | 391 | 328 |
| 8 | 402 | 330 |
| 9 | 290 | 231 |

Check if there is any significance difference between the means.



*Since the p-value 0.001156 is less than the 5 % significance level, we reject the null hypothesis of no difference between the means is clearly rejected.*

**

**

GROUP 3

Monte Carlo simulation

Refer to pages 48 to 58 in the PDF file R\_Programming\_Lessons\_7.

* Monte Carlo simulation is a statistical approach which is concerned with experiments employing random numbers. The technique is used by professionals with applications in a variety of fields including Operations Research, Physics, Finance, Chemistry, Biology and Medicine.
* Monte Carlo methods are used to handle both probabilistic and deterministic problems according to whether or not they are directly concerned with the behavior and outcome of a random process. In the case of a probabilistic problem a simple Monte Carlo approach is to observe random numbers, chosen in such a way that they directly simulate the physical random process of the original problem, and to infer the desired solution from the behavior of these random numbers.
* Monte Carlo simulation has wide application in performing risk analysis by building models of possible results by substituting a range of values (a probability distribution) for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. Depending on the number of uncertainties and the ranges specified for them, a Monte Carlo simulation produces distributions of possible outcome values. By using probability distributions, variables can have different probabilities of different outcomes occurring. Probability distributions are a realistic way of describing uncertainty in variables of a risk analysis.

## Steps used in Monte Carlo methods

*1 Define some domain of inputs. This just means we have some set of variables and what values they can take on, or we have some observations that are part of a dataset.*

*2 Generate inputs (the values of the variables or sets of observations) randomly, governed by some probability distribution.*

*3 Perform some computation on these inputs.*

*4 Repeat 2 and 3 over and over either an infinite number of times ( a very large number of times usually >= 10000), or until convergence.*

*5 Aggregate the results from the previous step into some final computation.*

## Business Planning sample

There is an interesting article in the website “Frontline Solvers – Developers of the Excel Solver” (<http://www.solver.com/monte-carlo-simulation-example>) about a Business Planning sample using Monte Carlo Simulation.

You, as a marketing manager of a firm are planning to introduce a new product. You need to estimate the first year profit from this product, which will depend on:

* + Sales volume in units
  + Price per unit
  + Unit Cost
  + Fixed Cost

Net profit will be calculated as Net Profit = Sales Volume \* (Selling Price – Unit Cost) – Fixed Cost.

Fixed costs (for overhead, advertising, etc.) are known to be $120,000. But other factors all involve some uncertainty. Sales volume (in units) can over quite a range, and the selling price per unit will depend on competitor actions. Unit costs will also vary depending on vendor prices and production experience.

**Uncertain variables**

To build a risk analysis model, you must identify the uncertain variables – also called random variables. While there’s some uncertainty in almost all variables in a business model, we want to focus on variables where the range of values is significant.

**Sales and Price**

Based on the market research, you believe that there are equal chances that the market will be Slow, OK or Hot.

* In the “Slow market” scenario, you expect to sell 50,000 units at an average selling price of $11 per unit.
* In the “OK market” scenario, you expect to sell 75,000 units at an average selling price of $10 per unit.
* In the “Hot market” scenario, you expect to sell 100,000 units at an average selling price of $8 per unit. In this scenario, your competitors will push your price down.

**Unit Cost**

Another uncertain variable is unit cost. Your firm’s production manager advises you that unit costs may vary anywhere between $5.50 to $7.50, with a most likely cost of $6.50. In this case, the most likely cost is also the average cost.

**Uncertain Functions**

**Net Profit**

We now identify uncertain functions – also called functions of a random variable. Since Sales volume and Selling Price and Unit cost are all uncertain variables, the net profit calculated based on these variables is an uncertain function.

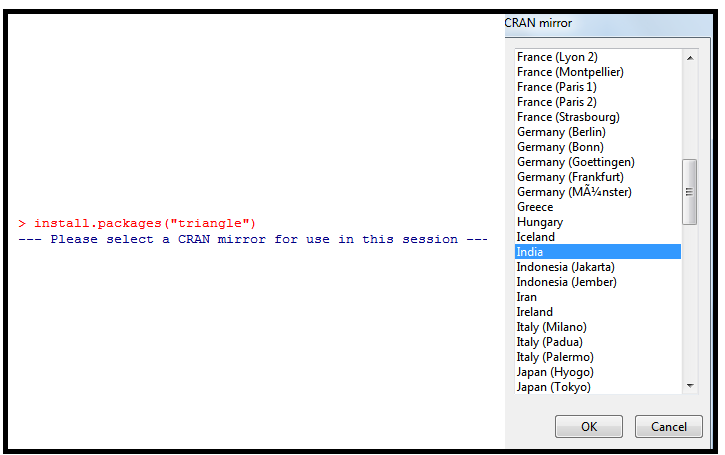
## Solution using R

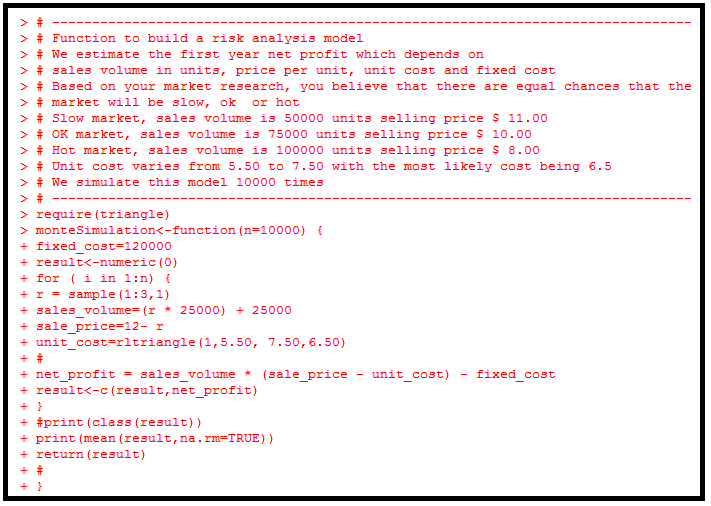
Since there are equal chances that the market will be Slow, OK, or Hot, we want to create an uncertain variable that selects among these three possibilities by generating a random number – say 1 or 2 or 3 – with equal probability. We associate 1 with “Slow Market” state, 2 with “OK market” state and 3 with “Hot market” state. We generate this number easily by using the R function – “sample” and then base the Sales Volume and Selling Price of this uncertain variable.

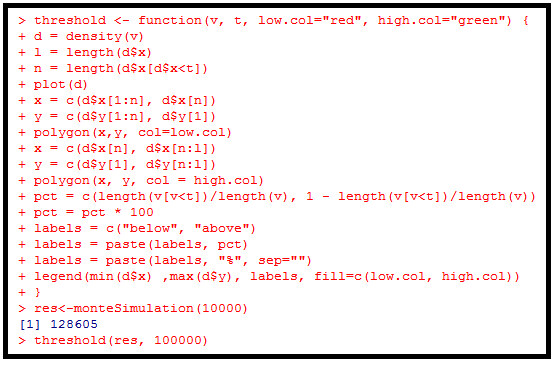
Unit cost may vary anywhere from $5.50 to $7.50, with a most likely cost of $6.50. We use triangular distribution to generate this variable.

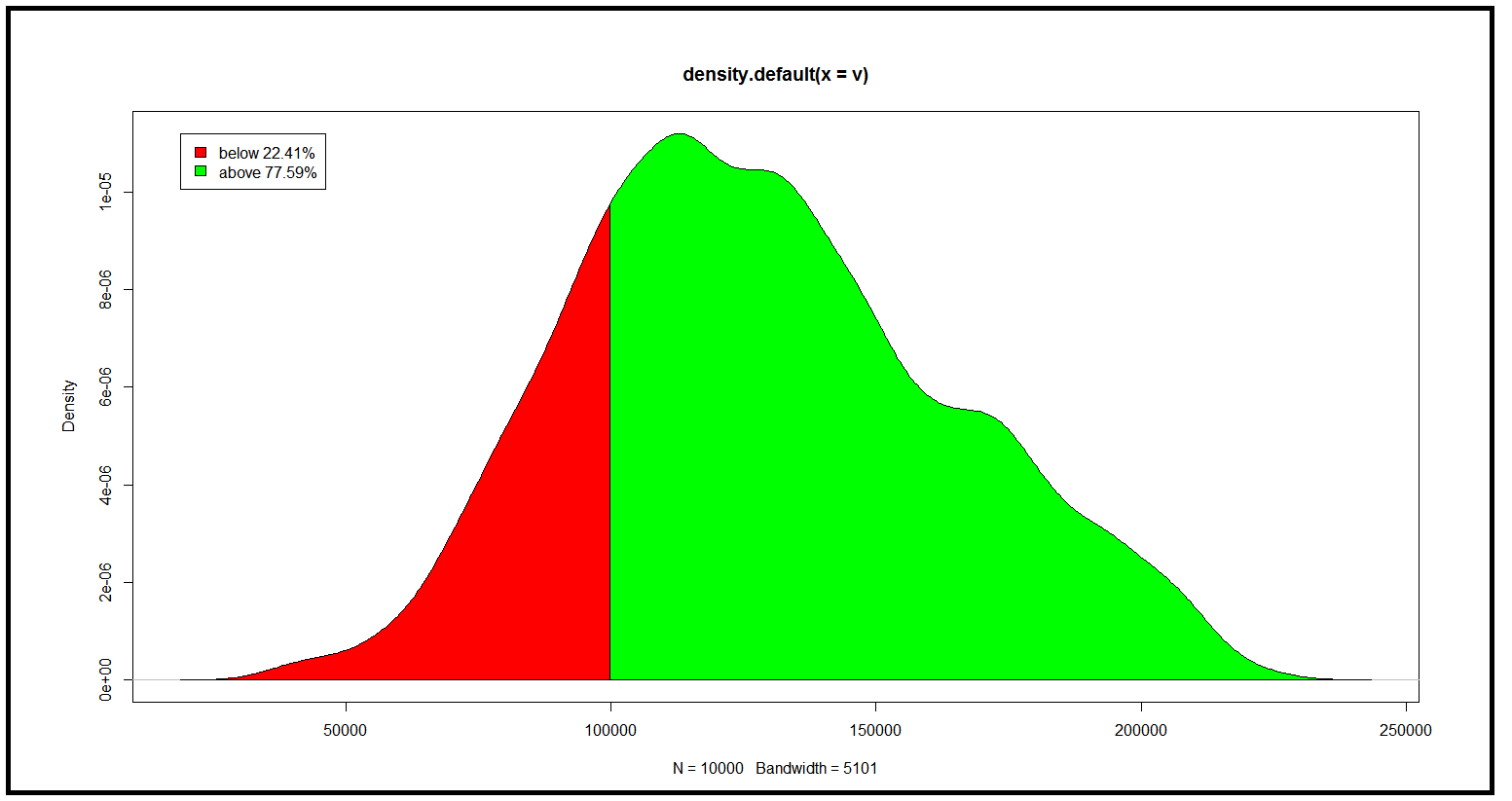
*In probability theory and statistics, the triangular distribution is a continuous probability distribution with a lower limit a, upper limit b and mode c, where a < b and a ≤ c ≤ b.*

We install “triangle” package first before using the R to solve this problem.









*We have observed that the average net profit after 10000 trials is $128,605. From this graph, we see that 77.59% of the trials resulted in the net profit above $100,00 and 22.41% of the trials resulted in the net profit below $100,000.*

GROUP 4

Regression

Refer to pages 65 to 77 in the PDF file R\_Programming\_Lessons\_7

Correlation denotes the association between two quantitative variables. Regression involves estimating the best equation to summarize the association.

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of correlation must be used.

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related, it means that when one changes by a certain amount the other changes on an average by a certain amount. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards them left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. First one is its distance from the baseline and the other being its slope. This is explained in the following regression equation.

y = a + bx

Regression equation enables us to predict y from x and gives us a better summary of the relationship between the two variables.

# Various Regression models

1. **Linear Regression**

The two basic types of regression are simple linear regression and multiple linear regression.

* Simple linear regression uses one independent variable to explain and / or predict the value of a dependent variable, Y
* Multiple linear regression uses two or more independent variables are used to predict the value of a dependent variable.

The difference between the two is the number of independent variables. In both cases, there is only a single dependent variable.

Linear Regression: Y = α + β X + ε

Multiple Regression: Y = α + β1 X1 + β2 X2 + β3X3 +…...+ βn Xn + ε

Where Y = the variable that we are trying to predict

X = the variable that we are using to predict Y

α = the intercept

β = the slope

ε = the regression residual

* This is the oldest type of regression, designed 250 years ago
* Computations (On small data) could be easily be carried out by a human being or computer.
* This can be used for interpolation, but not suitable for predictive analytics
* This has many drawbacks when applied to modern data, e.g., sensitivity to both outliers and cross-correlations (both in variable and observation domains) and subject to over-fitting. A better solution is piecewise-linear regression, in particular for time series.

1. **Polynomial Regression**

Polynomial regression uses one independent variable x and a dependent variable y.

Y = α + β1 X + β2 X2 + β3X3 +…...+ βn Xn + ε

* In principle this is no different from fitting multiple regression model except that the powers of x play the role of different independent variables. But polynomial regression has special features.
* We fit a polynomial to smooth out fluctuations in the data caused by random or uncontrolled errors, not because it is thought to represent the relationship.
* While fitting the polynomial regression the form of the null hypothesis takes is that polynomial regression being fitted represents certain relationship and secondly, whether terms of higher degree contributes significantly to the relationship.

1. **Logistic Regression**

We use the logistic regression equation to predict the probability of a dependent variable taking the dichotomy values 0 or 1. Suppose x1,x2,x3,..xp are the independent variables, α and βk (k = 1,2..p) are the parameters, and E(y) is the expected value of the dependent variable y, then the logistic regression equation is

E(y) = 1/ (1 + e-(α + ∑βk xk) )

* This is used extensively in clinical trials, scoring and fraud detection, when the response is binary (chance of succeeding or failing, e.g. for a new tested drug or credit card transaction).
* This also suffers same drawbacks as linear regression (not robust, model-dependent), and computing regression coefficients involves using complex iterative, numerically unstable algorithm.
* This can be well approximated by linear regression after transforming the response (ligit transform).
* Some versions (Poisson or Cox regression) have been designed for a non-binary response, for categorical data (classification), ordered integer response (age groups), and even continuous response (regression trees)

# Statistical Models in R

R includes a variety of tools for complex modeling, among them:

* glm() for generalized linear models
* gam() for generalized additive models
* lme() and lmer() for linear mixed-effects models
* nls() and nlme() for nonlinear models

R functions such as aov(), lm() use a formula interface to specify the variables to be included in the analysis. The formula determines the model that will be built and tested by the R procedure. The basic format of such a formula is

Response variable ~ explanatory variables

The tilde should be read “is modeled by” or “is modeled as a function of”.

A basic regression analysis would be formulated this way:

y ~ x

… where “x” is the explanatory variable or IV and “y” is the response variable or DV. Additional variables would be added in as follows:

y ~ x + z which would make this z multiple regression with two predictors. This raises a critical issue that must be understood to get model formulae correct.

Symbols used as mathematical operators in other contexts do not have their usual mathematical meaning inside model formulae.

The following table lists the meaning of these symbols when used in a formula.

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Example** | **Meaning** |
| + | +x | Include this variable |
| - | -x | Delete this variable |
| : | x:z | Include interaction between these variables |
| \* | x\*z | Include these variables and the interactions between them |
| / | x/z | Nesting: include z nested within x |
| | | x | z | Conditioning: include x given z |
| ^ | (u + v +w)^3 | Include these variables and all interactions up to three way |
| poly | poly(x,3) | Polynomial regression: Orthogonal polynomials |
| Error | Error(a/b) | Specify the error term |
| I | I(x\*z) | As is: include a new variable consisting of these variables multiplied |
| I | -I | Intercept: Delete the intercept (regress through the origin) |

* Some formula structures can be specified in more than one way..
  + y ~ u + v + w + u : v + u :w + u: v: w
  + y ~ u \* v \* w – u :v :w
  + y ~ (u + v + w) ^ 2
* The nature of variables – binary, categorical (factors), numerical – will determine the nature of the analysis.
* For example. If u and v are factors
  + y ~ u + v dictates an analysis of variance (without the interaction term).
* If u and v are numerical, the same formula would dictate a multiple regression.
* If u is numerical and v is a factor, then an analysis of covariance is dictated.

# Least square method – Line of best fit

**A line of best fit is a straight line that is the best possible approximation of the given set of data.**

It is used to study the nature of relation between two variables.

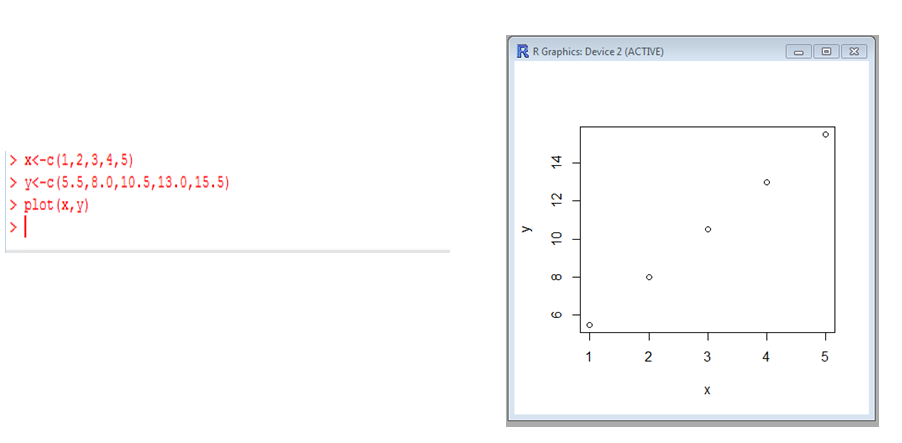
Let us assume y = 2.5 x + 3

|  |  |
| --- | --- |
| **x** | **y** |
| **1** | **5.5** |
| **2** | **8** |
| **3** | **10.5** |
| **4** | **13** |
| **5** | **15.5** |
|  | |

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal

(Note that the line passes through as many points as possible).

Plot the points on a coordinate plane.



A more accurate way of finding the line of best fit is the least square method.

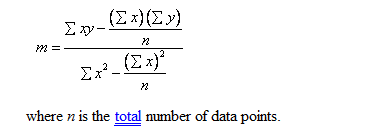
Use the following steps to find the equation of line of best fit for a set of ordered pairs.

Step 1: Calculate the mean of the x-values and the mean of y-values.

Step 2: Compute the sum of squares of the x-values.

Step 3: Compute the sum of each x-value multiplied by its corresponding y-value.

Step 4: Calculate the slope of the line using the formula:

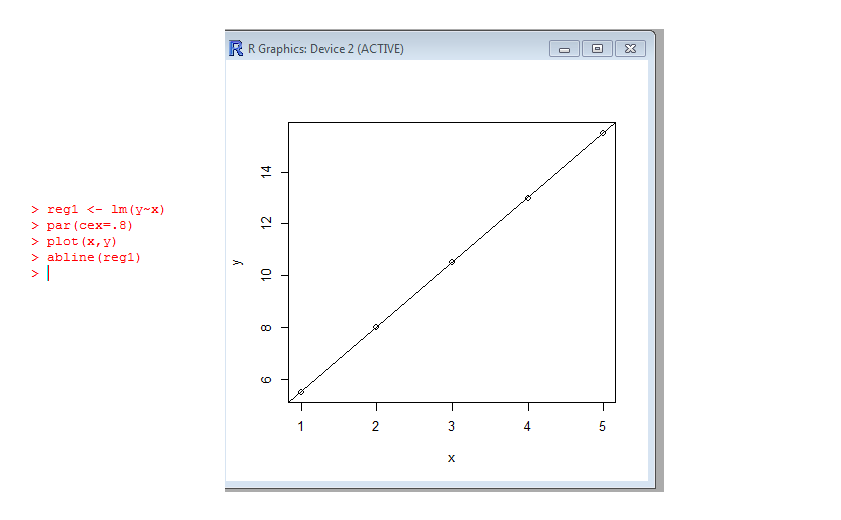


Step 5: Compute the y-intercept of the line using the formula:



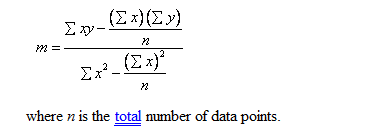
Step 6: Use the slope and the y-intercept to form the equation of the line.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | 1 | 2 | 3 | 4 | 5 |
| y | 5.5 | 8.0 | 10.5 | 13.0 | 15.5 |



Calculate the slope of the line:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | |  |
| x | 1 | 2 | 3 | 4 | 5 | 15 |
| y | 5.5 | 8.0 | 10.5 | 13.0 | 15.5 | 52.5 |
| xy | 5.5 | 16.0 | 31.5 | 52.0 | 77.5 | 182.5 |
| x2 | 1 | 4 | 9 | 16 | 25 | 55 |



m = (182.5 – (15 X 52.5)/5) / (55 – (152 / 5) = 25 / 10 = 2.5

Compute the y-intercept of the line using the formula:



b = (52.5 / 5) - 2.5 \*(15/5) = 10.5 – 7.5 = 3

The regression equation is y = 2.5 x + 3.0.



# Simple Linear Regression

A simple linear regression model that describes the relationship between two variables x and y can be expressed by the following equation. The numbers **α** and **β** are called parameters, and **ε** is the error term.

**y = α + βx + ε**

Using R, we can compute the slope and intercept as follows:

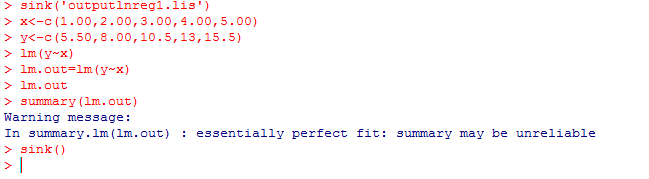
lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

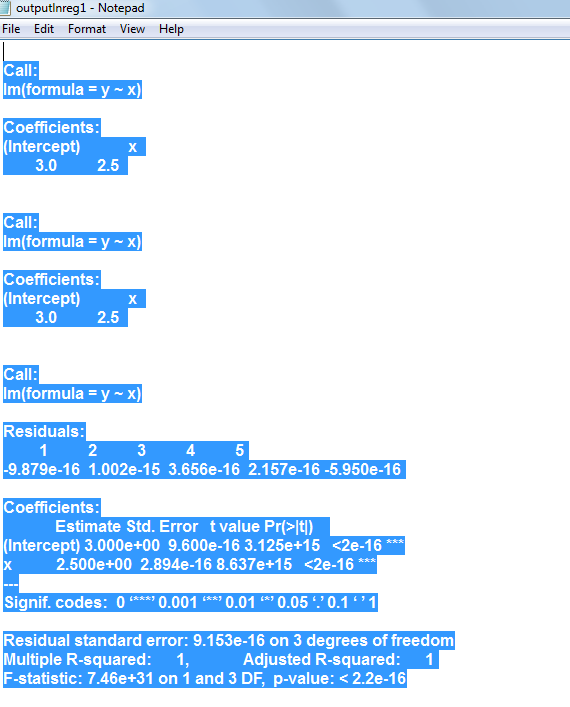
lm (formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)

**Arguments**

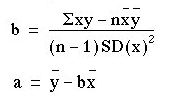
|  |  |
| --- | --- |
| formula | an object of class "[formula](http://127.0.0.1:13120/library/stats/help/formula)" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under ‘Details’. |
| data | an optional data frame, list or environment (or object coercible by [as.data.frame](http://127.0.0.1:13120/library/stats/help/as.data.frame) to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called. |
| subset | an optional vector specifying a subset of observations to be used in the fitting process. |
| weights | an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. If non-NULL, weighted least squares is used with weights weights (that is, minimizing sum(w\*e^2)); otherwise ordinary least squares is used. See also ‘Details’, |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of [options](http://127.0.0.1:13120/library/stats/help/options), and is [na.fail](http://127.0.0.1:13120/library/stats/help/na.fail) if that is unset. The ‘factory-fresh’ default is [na.omit](http://127.0.0.1:13120/library/stats/help/na.omit). Another possible value is NULL, no action. Value [na.exclude](http://127.0.0.1:13120/library/stats/help/na.exclude) can be useful. |
| method | the method to be used; for fitting, currently only method = "qr" is supported; method = "model.frame" returns the model frame (the same as with model = TRUE, see below). |
| model, x, y, qr | logicals. If TRUE the corresponding components of the fit (the model frame, the model matrix, the response, the QR decomposition) are returned. |
| singular.ok | logical. If FALSE (the default in S but not in **R**) a singular fit is an error. |
| contrasts | an optional list. See the contrasts.arg of [model.matrix.default](http://127.0.0.1:13120/library/stats/help/model.matrix.default). |
| offset | this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length equal to the number of cases. One or more [offset](http://127.0.0.1:13120/library/stats/help/offset) terms can be included in the formula instead or as well, and if more than one are specified their sum is used. See [model.offset](http://127.0.0.1:13120/library/stats/help/model.offset). |
| ... | additional arguments to be passed to the low level regression fitting functions (see below). |

Models for lm are specified symbolically. A typical model has the form response ~ terms where response is the (numeric) response vector and terms is a series of terms which specifies a linear predictor for response.

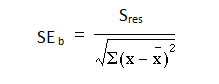




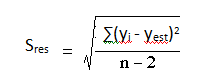
If, for xi the regression equation predicts a value of yest, the regression error is (yi - yest). It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error ∑(yi - yest) of zero because the positive and negative terms exactly cancel. To remove the negative signs we square the differences and the regression equation chosen to minimize the sum of squares of the prediction errors, S2 = ∑(yi - yest)2

It can be shown that the one straight line that minimizes S2, the least squares estimate gives:

The standard error of the slope SEb is given by:



where Sres is the residual standard deviation, is given by:





## Estimated Simple Regression Equation

If we choose the parameters α and β in the simple linear regression model so as to minimize the sum of squares of the error term ε, we will have the so called estimated simple regression equation. It allows us to compute the fitted values of y based on values of x.

ŷ = a + b x

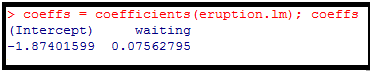
Apply the simple linear regression model for the data set faithful, and estimate the next eruption duration if the waiting time since the last eruption has been 90 minutes.

**Solution**

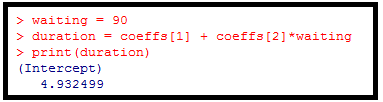
We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.



Then we extract the parameters of the estimated regression equation with the coefficients function.



Now, we fit the eruption duration using the estimated regression equation

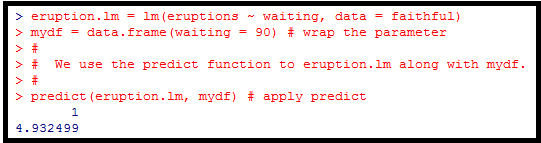


**Conclusion**

Based on the simple linear regression model, if the waiting time since the last eruption has been 90 minutes, we expect the next one to last 4.932499 minutes.

**Alternative Solution**

We include the waiting parameter inside a new data frame named as mydf.



The coefficient of determination of a linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable.

We denote y, as the observed value of the dependent variable, as its mean, and ŷi- as the fitted value, then the coefficient of determination is:



## Coefficient of Determination, r2 or R2

The coefficient of determination, r2 is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model / graph.

The coefficient of determination is the ratio of the explained variation to the total variation.

The coefficient of determination is such that 0 ≤ r2 ≤ 1, and denotes the strength of the linear association between x and y.

The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if r = 0.922, then r2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% if the total variation in y remains unexplained.

The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation, the further the line is away from the points, the less is it able to explain.

**Problem**

Find the coefficient of determination for the simple linear regression model of the data set faithful.

**Solution**

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.



Then we extract the coefficient of determination from the ***r.squared*** attribute of its summary.

****

**Answer**

The coefficient of determination of the simple linear regression for the data faithful is 0.8114608. That means 81% of the total variation in eruptions can be explained by the linear relationship between waiting and eruptions (as described by the regression equation).

## Significance Test for Linear Regression

Assume that the error ε in the linear regression model is independent of x and is normally distributed with zero mean and constant variance. We can decide whether there is any significant relationship between x and y by testing the null hypothesis that β = 0.

**Problem**

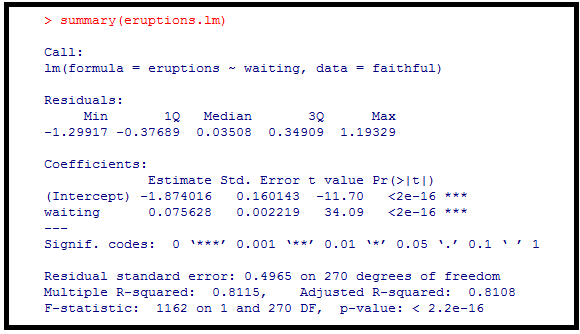
Decide whether there is a significant relationship between the variables in the linear regression model of the data set faithful at 0.05 significance level.

**Solution**

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruptions.lm.



Then we print out the F-statistics of the significance test with the summary function.



**Answer**

As the p-value is much less than 0.05, we reject the null hypothesis that β = 0. Hence there is a significant relationship between the variables in the linear regression model of the data set faithful.

## Confidence Interval for Linear Regression

Let us assume that the error term **ε** in the linear regression model is independent of x, and is normally distributed, with mean zero and constant variance. For a given value of x, the interval estimate for the mean of the independent variable, y is called the confidence interval.

**Problem**

In the dataset faithful, develop a 95 % confidence interval of the mean eruption for the waiting time of 80 minutes.

**Solution**

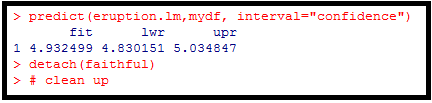
We apply the ***lm*** function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable **eruption.lm**.



Then we create a new data frame that set the waiting time value.



We now apply the predict function and set the predictor variable in the mydf argument. We also set the interval type as “confidence”, and use the default 0.95 confidence level.



**Conclusion**

*The 95 % confidence level of the mean eruption duration for the waiting time of 90 minutes is between 4.830151 and 5.034847.*

## Prediction Interval for Linear Regression

Let us assume that the error term **ε** in the simple linear regression model is independent of x, and is normally distributed, with mean zero and constant variance. For a given value of x, the interval estimate for the mean of the independent variable, y is called the prediction interval.

**Problem**

In the data set faithful, develop a 95 % prediction interval of the eruption duration for the waiting time of 90 minutes.

**Solution**

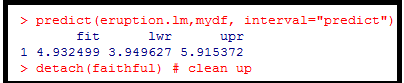
We apply the ***lm*** function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable **eruption.lm**.



Then we create a new data frame that set the waiting time value.



We now apply the predict function and set the predictor variable in the mydf argument. We also set the interval type as “predict”, and use the default 0.95 confidence level.



**Conclusion**

*The 95% prediction level of the mean eruption duration for the waiting time of 90 minutes is between 3.949627 and 5915372.*

## Residual Plot

The residual data of the simple linear regression model is the difference between the observed data of the dependent variable y and the fitted values ŷ.

Residual = y – ŷ

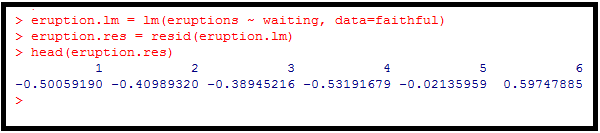
**Problem**

Plot the residual of the simple linear regression model of the data set **faithful** against the independent variable **waiting**.

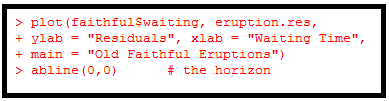
**Solution**

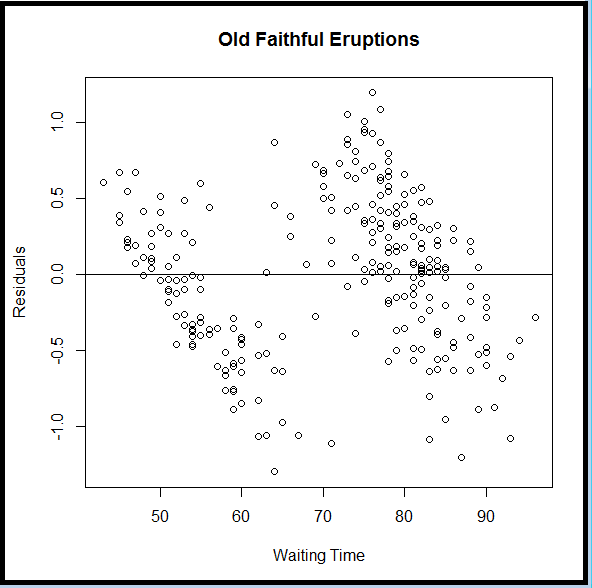
We apply the **lm** function to a formula that describes the variable eruptions by the variable **waiting**, and save the linear regression model in a new variable **eruption.lm**.

Then we compute the residual with the **resid** function.



We now plot the residual against the observed values of the variable **waiting**.





## Standard Residual

The standard residual is the residual divided by its standard deviation.

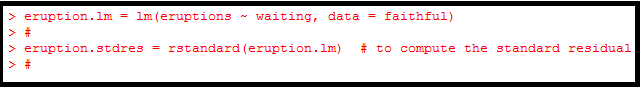
***Standard Residual i = Residual i / Standard Deviation of Residual i***

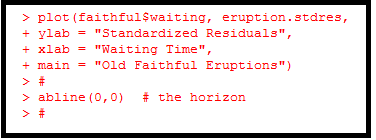
**Problem**

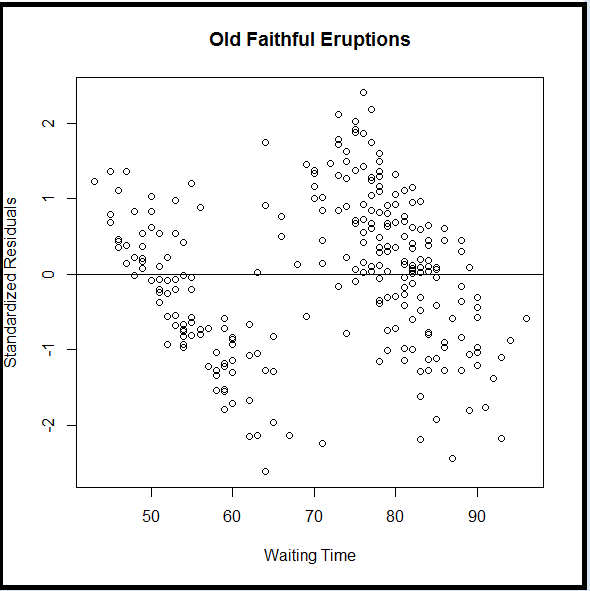
Plot the standardized residual of the simple linear regression model of the data set **faithful** against the independent variable **waiting**.

**Solution**

We apply the **lm** function to a formula that describes the variable eruptions by the variable **waiting**, and save the linear regression model in a new variable **eruption.lm**. Then we compute the standardized residual with the ***rstandard*** function.

We now plot the standardized residual against the observed values of the variable **waiting**.





## Normal Probability Plot of Residuals

The normal probability plot is a graphical tool for comparing a data set with the normal distribution. We can use it with the standard residual of the linear regression model and see if the error term e is actually normally distributed.

**Problem**

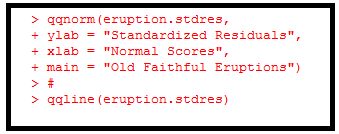
Create a normal probability plot for the standardized residual of the data set faithful.

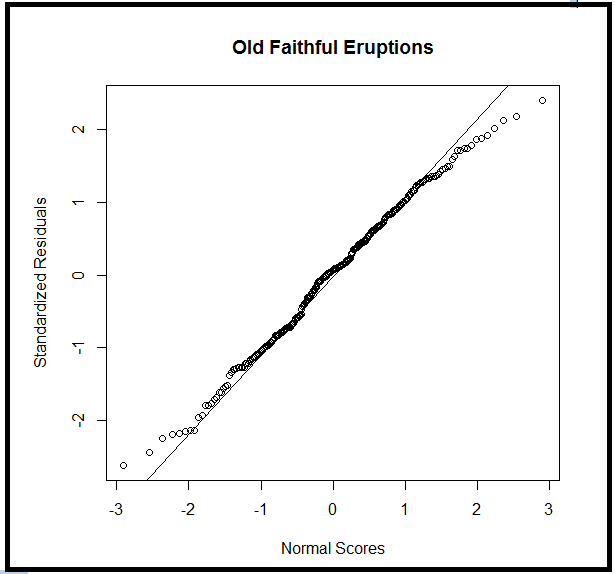
**Solution**

We apply the **lm** function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable **eruption.lm**. Then we compute the standardized residual with the ***rstandard*** function.



We now create the normal probability plot with the **qqnorm** function, and add the **qqline** for further comparison.





**About Quantile-Quantile Plots**

qqnorm is a generic function the default method of which produces a normal QQ plot of the values in y. qqline adds a line to a “theoretical”, by default normal, quantile-quantile plot which passes through the probs quantiles, by default the first and third quartiles.

qqplot produces a QQ plot of two datasets.

Graphical parameters may be given as arguments to qqnorm, qqplot and qqline.

**Usage**

qqnorm(y, ...)

#

## Default S3 method:

#

qqnorm(y, ylim, main = "Normal Q-Q Plot", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", plot.it = TRUE, datax = FALSE, ...)

qqline(y, datax = FALSE, ...)

qqplot(x, y, plot.it = TRUE, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)), ...)

**Arguments**

x The first sample for qqplot.

y The second or only data sample.

xlab, ylab, main plot labels.

plot.it logical. Should the result be plotted?

datax logical. Should data values be on the x-axis?

ylim, ... graphical parameters.

**Value**

**For qqnorm and qqplot, a list with components**

x The x coordinates of the points that were/would be plotted

y The original y vector, i.e., the corresponding y coordinates including NAs.

**References**

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.



# Multiple Linear Regression

A multiple linear regression (MLR) model that describes a dependent variable y by independent variables x1,x2,x3,...,xp (p > 1) is expressed by the equation as follows:

y = α + ∑βkxk + ε where the α, βk (k =1,2,..p) are the parameters and ε is the error term.



## Estimated Multiple Regression Equation

If we choose the parameters α and βk (k =1,2,..p) in the multiple linear regression model so as to minimize the sum of squares of the error term ε, we will have the so called estimated multiple regression equation. It allows us to compute fitted values of y based on a set of values of xk (k =1,2,..p).

ŷ= a + ∑bkxk

**Problem**

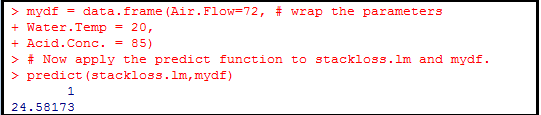
Apply the multiple linear regression model for the built-in data set stackloss, and predict the stackloss if the air flow is 72, water temperature is 20 and acid concentration is 85.

**Solution**

We apply the **lm** function to a formula that describes the variable **stack.loss** by the variables **Air.flow**, **Water.Temp** and **Acid.Conc**. Save the linear regression model in a new variable **stackloss.lm**.

****

We also wrap the parameters inside a new data frame named mydf.

****

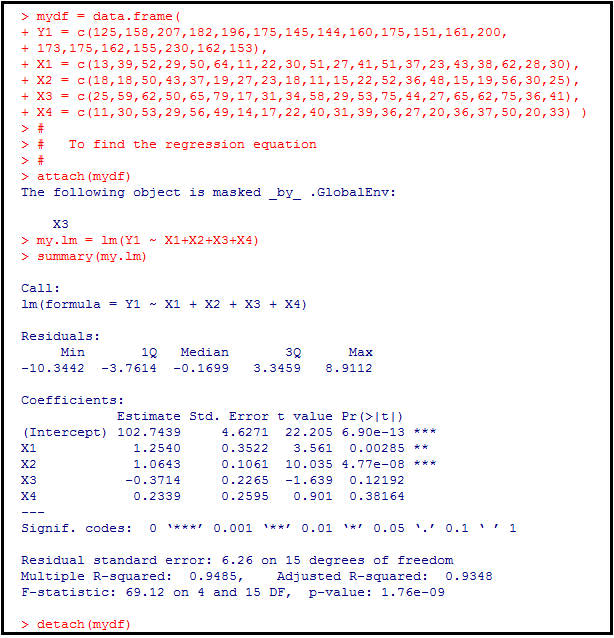
**Conclusion**

Based on the multiple linear regression model and the given parameters, the predicted stackloss is 24.58173.

Example - Find the regression equation between Y1 and X1,X2,X3,X4

* Y1 is a measure of success in graduate school
* X1 is a measure of intellectual ability
* X2 is a measure of work ethic
* X3 is a second measure of intellectual ability
* X4 is a measure of spatial ability

**Solution**



The regression equation is

**Y1 = 102.7439 + 1.254\* X1 + 1.0643 \* X2 – 0.3714 \* X3 + 0.2339 \* X4**

## Multiple Coefficient of Determination

The coefficinet of determination of a multiple linear regression model is the quotient of the variances of the fitted values and observed values of the depedent variable. If we denote yi as the observed values of the dependent variable, as its mean, and Ŷi as the fitted value, then the coefficient of determination is:

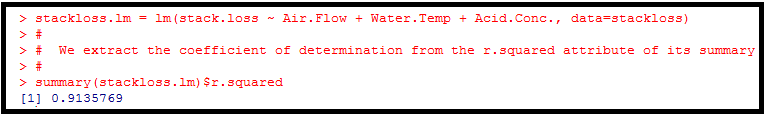


**Problem 1**

Find the coefficient of determination for the multiple linear regression model of the data set stackloss.

**Solution**

We apply the **lm** function to a formula that describes the variable **stack.loss** by the variables **Air.flow**, **Water.Temp** and **Acid.Conc**. Save the linear regression model in a new variable **stackloss.lm**.



**Conclusion**

The coefficient of determination of the multiple linear regression model for the data set stackloss is 0.9135769. This means that over 91% of the total variation in **stack.loss** can be explained by the linear relationship between **Air.flow**, **Water.Temp** and **Acid.Conc** and **stack.loss** (as described by the regression equation). The other 9% of the total variation in **stack.loss** remains unexplained.

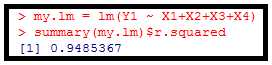
**Problem 2**

Find the coefficient of determination for the multiple linear regression model of the data Y1 and (X1,X2,X3 and X4)

**Solution**

We apply the **lm** function to a formula that describes the variable Y1by the variables and (X1,X2,X3 and X4)

. Save the linear regression model in a new variable **my.lm.**



**Conclusion**

The coefficient of determination of the multiple linear regression model for the above data is 0.9485347. This means that approximately 95% of the total variation in **Y1** can be explained by the linear relationship between (X1,X2,X3 and X4) and Y1 (as described by the regression equation). The other 5% of the total variation in **Y**1 remains unexplained.

## Adjusted Coefficient of Determination

The adjusted coefficient of determination of a multiple linear regression model is defined in terms of the coefficient of determination as follows, where n is the number of observations in the data set, and p is the number of independent variables.

R2 adj = 1 – ( 1 – R2) ((n -1) /( n- p -1))

We extract the adjusted coefficient of determination from the adj.r.squared attribute of the summary of stackloss.lm as determined in the previous page.



The ajusted coefficient of determination of the multiple linear regression model for the data stackloss is 0.8983258.

Similarly, when we extract adjusted coefficient of determination from the adj.r.squared attribute of the summary of my.lm as determined in the earlier pages.



The ajusted coefficient of determination of the multiple linear regression model for the linear relationship between (X1,X2,X3 and X4) and Y1 is 0.9348132.

**Note:**

* The use of an adjusted R2 is an attempt to take account of the phenomenon of the R2 automatically and spuriously increasing when extra explanatory variables are added to the model.
* It is a modification due to Theil of R2 that adjusts for the number of explanatory terms in a model relative to the number of data points.
* The adjusted R2 can be negative, and its value will always be less than or equal to that of R2.
* Unlike R2, the adjusted R2 increases when a new explanator is included only if the new explanator improves the R2 more than would be expected by chance.
* If a set of explanatory variables with a predetermined hierarchy of importance are introduced into a regression one at a time, with the adjusted R2 computed each time, the level at which adjusted R2 reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess/unnecessary terms.

## Significant Test for Multiple Linear Regression

Assume that the error term ε in the mulitple linear regression is independent of xk (k =1,2,...p), and is normally distributed, with zero mean and constant variance. We can decide whether there is any significant relationship between the dependent variable y and any of the dependent variables of xk (k =1,2,...p).

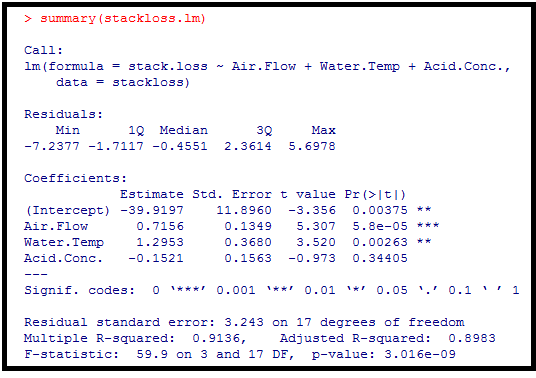
**Problem 1**

Decide which of the independent variables in the multiple linear regression model of the data set stackloss are statistically significant at .05 significance level.

**Solution**



The t values of the independent variables can be found with the summary function.



C**onclusion**

*As the p-value of Aire Flow and Water.Temp are less than 0.05, they are both statistically significant in the multiple linear regression model of stackloss.*

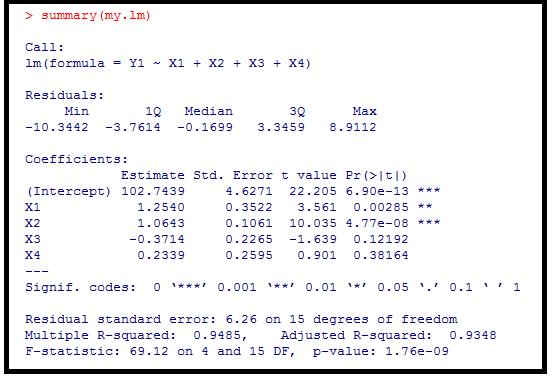
**Problem 2**

Decide which of the independent variables in the multiple linear regression model between (X1,X2,X3 and X4) and Y1 are statistically significant at .05 significance level.

**Solution**



The t values of the independent variables can be found with the summary function.



C**onclusion**

*As the p-value of X1 and X2 are less than 0.05, they are both statistically significant in the multiple linear regression model.*

## Confidence Interval for Multiple Linear Regression

Assume that the error term ε in the mulitple linear regression is independent of xk (k =1,2,...p), and is normally distributed, with zero mean and constant variance. For a given set of values of xk (k =1,2,...p), and the interval estimate for the mean of the dependent variable, is called the confidence interval.

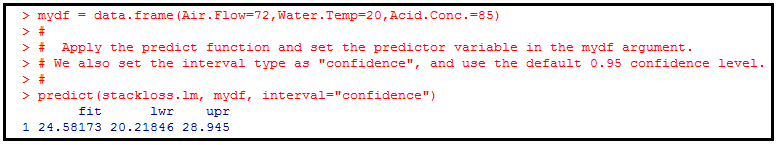
**Problem**

In the data set stackloss, develop a 95% confidence interval of the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

**Solution**



We wrap the parameters inside a new data frame, mydf.



**Conclusion**

*The 95% confidence interval of the stack loss with the given parameters is between 20.21846 and 28.945.*

## Prediction Interval for Multiple Linear Regression

Assume that the error term ε in the mulitple linear regression is independent of xk (k =1,2,...p), and is normally distributed, with zero mean and constant variance. For a given set of values of xk (k =1,2,...p), and the interval estimate for the dependent variable y is called the prediction interval.

**Problem**

In the data set stackloss, develop a 95% prediction interval of the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

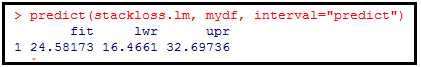
**Solution**



We wrap the parameters inside a new data frame, mydf.



Let us apply the predict function and set the predictor variable in the predictor variable in the mydf argument. We also set the interval type as :predict”, and use the default 0.95 confidence level.



**Conclusion**

*The 95% confidence interval of the stack loss with the given parameters is between 16.466 and 32.69736.*

# Polynomial Regression

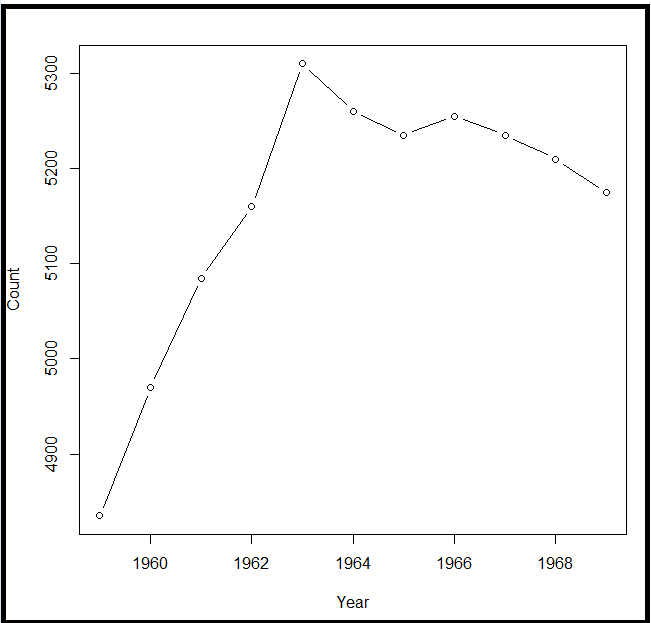
Polynomial regression uses one independent variable x and a dependent variable y.

Y = α + β1 X + β2 X2 + β3X3 +…...+ βn Xn + ε

Assume we want to create a ploynomial that can approximate better the following dataset on the customers vising a popular restaurant in an Indian City in January over 10 years.



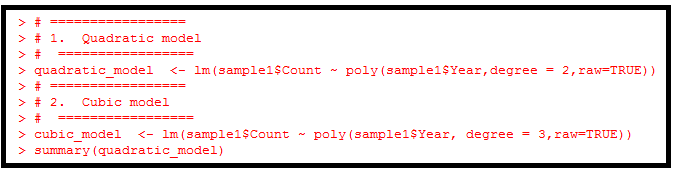
Use the scatter plot to check the nature of data.



From the above scatter plot we find that the data may not be linear.

We will explore the quadratic and cubic model.

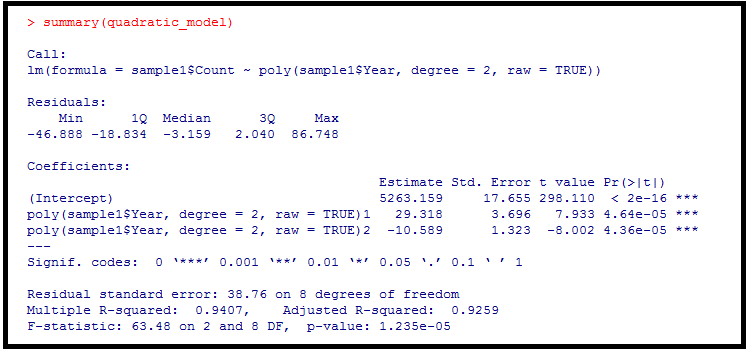
1. **Creating the models**



1. **Evaluating the models**

It is helpful to summarize and compare our potential models using the summary(MODEL) and anova(MODEL1,MODEL2,MODEL3) functions.

1. **Quadratic Model**



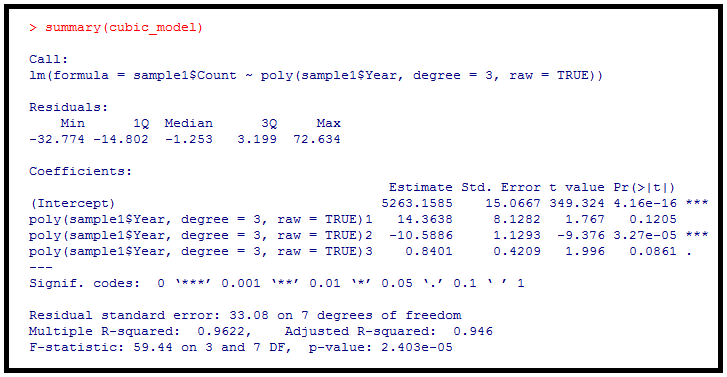
The output of summary(quadratic\_model) is given above.

We observe the values of beta(5263.159), beta1(29.318) and beta2(-10.589), which appear to be significant and all are less than 0.05 significance level.

The equation of polynomial of degree 2 of our model is:

Count = 5263.1597 + 29.318 x – 10.589x2

1. **Cubic Model**



The output of summary(cubic\_model) is given above.

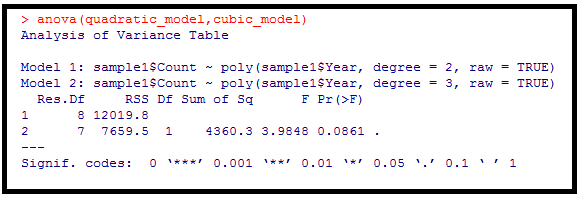
The equation of polynomial of degree 3 of our model is:

Count = 5263.1585 + 14.3638 x – 10.5886x2 + 0.8401 x3

We observe the values of beta(5263.1585), beta1(14.3638), beta2(-10.5886) and beta3(0.8401).

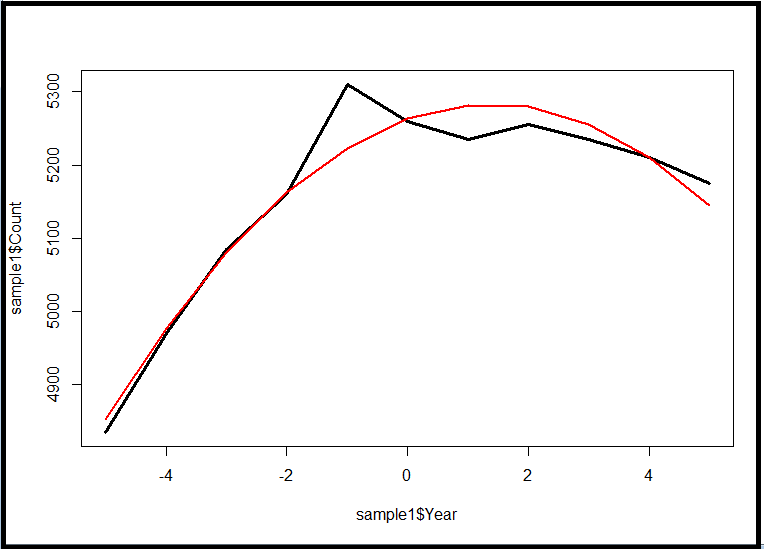
We observe that the coefficients beta1 and beta3 are not significant.

Multiple R –squared in the 2nd degree model is 94.07 % while in the 3rd degree model it is 96.22%. It seems that there has been an increase in accuracy of the model, but it is a significant increase? We can compare the two model with an ANOVA table.



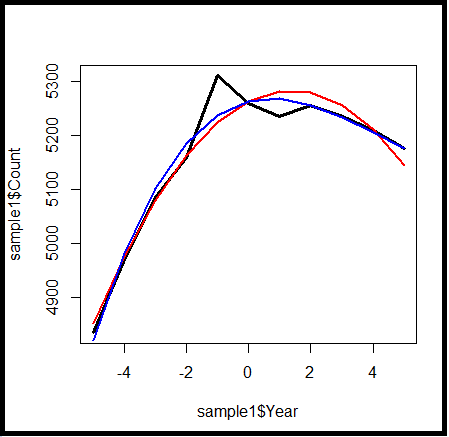
Since the p-value is greater than 0.05, we accept the null hypothesis: there wasn’t a significant improvement of the model.

Now add to the scatter plot -we have already drawn the progress of 2nd degree polynomial.



The function predict() compute the Y values given the X values. The coordinates are linked with the lines. With a few values this method is highly debilitating.

Let us add the graph of the polynomial of 3rd degree.



**The two models have very similar trends.**

# Logistic Regression

We use the logistic regression equation to predict the probability of a dependent variable taking the dichotomy values 0 or 1. Suppose x1,x2,...xp are independent variables, α and βk (k=1,2,...p) are parameters, and E(y) is the expected value of the dependent variable y, then the logistic regression equation is:

E(y) = 1 / ( 1 + e-(α + ∑βkxk))

For example, in the built-in data set mtcars, the data column am represents the transmission type of the automobile model (0 = automatic, 1 = manual). With the logistic regression equation, we can model the probability of a manual transmission in a vehicle based on its engine horsepower and weight data.

P(Manual Transmission) = 1 / ( 1 + e-(α + β1\* horse power + β2\*weight))



## Estimated Logistic Regression Equation

Using the generalized linear model, an estimated logistic regression equation can be formulated as below:

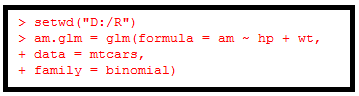
The coefficients a and bk (k=1,2...,p) are determined according to a maximum likelihood approach, and it allows us to estimate the probability of the dependent variable y taking on the value 1 for the given values of xk( k=1,2....p).

**Problem**

By use of the logistic regression equation of vehicle transmission in the data set mtcars, estimate the probability of a vechicle being fitted with a manual transmission if it has a 120 hp engine and weights 2800 lbs.

**Solution**

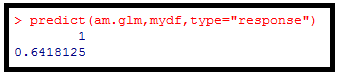
We apply the function glm to a formula that describes the transmission type (am) by the horsepower (hp) and weight (wt). This creates a generalized linear model (GLM) in the binomial family.



We then wrap the test parameters inside a data frame mydf.



Now we apply the function predict to the generalized linear model **am.glm** along with **mydf**. We will have to select *response* prediction type in order to obtain the predicted probability.



**Conclusion**

*For an automobile with 120 hp engine and 2800 lbs weight, the probability of it being fitted with a manual transmission is about 64%.*

## Significance Test for Logistic Regression Equation

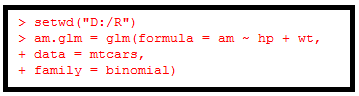
We can decide whether there is any significant relationship between the dependent variable y and the independent variables xk (k = 1,2,...p) in the logistic regression equation. In particular, if any of the null hypothesis that βk (k = 1,2,...p) is valid, then xk is statistically insignificant in the logitic regression model.

**Problem**

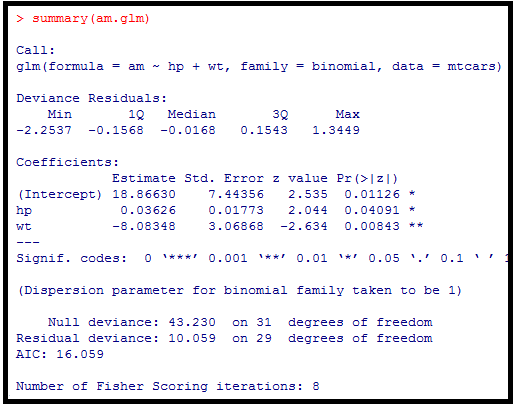
At .05 significance level, decide if any of the independent variables in the logistic regression model of vechicle transmission in data set mtcars is statistically insignificant.

**Solution**

We apply the function glm to a formula that describes the transmission type (am) by the horsepower (hp) and weight (wt). This creates a generalized linear model (GLM) in the binomial family.



We then print out the summary of the generalized linear model and check for the p-values of the hp and wt variables.

****

**Conclusion**

*As the p-values of the hp and wt variables are both less than 0.05, neither hp nor wt is insignificant in the logistic regression model.*

## Another example for Logistic Regression Equation

Ref: <http://www.tatvic.com/blog/logistic-regression-with-r/>

* Logistic regression is one of the type of regression and it is used to predict outcome of the categorical dependent variable (i.e. categorical variable has limited number of categorical values) based on the one or more independent variables.
* In binomial or binary logistic regression, the outcome can have only two possible types of values (e.g. "Yes","No","Success","Failure").
* Multinomial logistic refers to cases where the outcome can have three ot more possible types of values (e.g."Good" Vs "Very Good" Vs "Best"). Generally outcome is coded as "0" or "1" in binary logistic regression.

Suppose you want to predict whether a student will get admission based on his two exam scores. For this problem, we have a historical data from previous applicants which can be used as the dataset to build a model.

The data set contains:

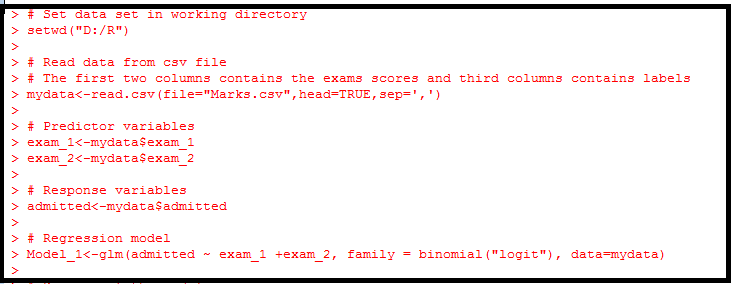
1. exam\_1 - Exam\_1 score

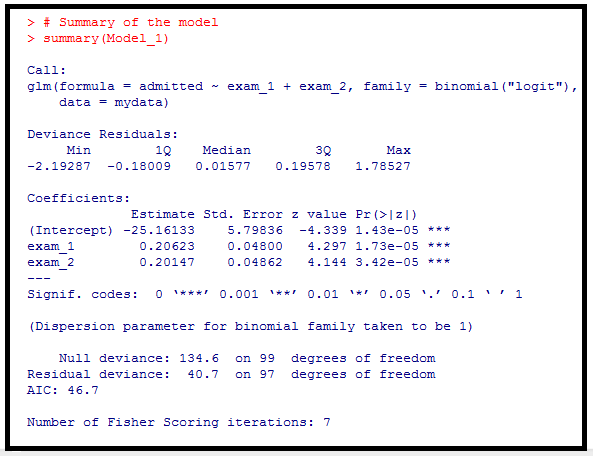
2. exam\_2- Exam\_2 score

3. admitted - 1 if admitted or 0 if not admitted

In the above parametes, parameter admitted has value 1 or 0 for each observation. Now, we will generate a model that can predict, will student get admission based on two exam scores?

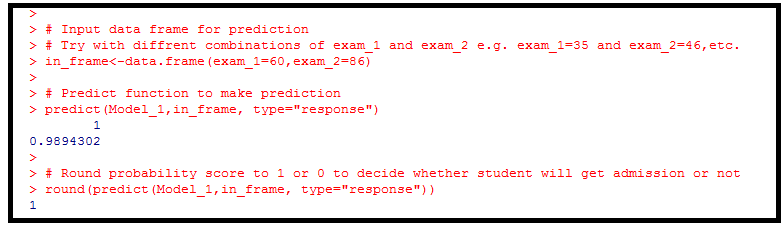
Here, admitted is considered as dependent variable, exam\_1 and exam\_2 are considered as independent variables.





After generating the model. let us try to predict using this model. Suppose we have two exam marks of a student, 60 of exam\_1 and 85 of exam\_2. We will predict that will student get admission?

Following the R code for predicting probability of student get admission.



**Conclusion:**

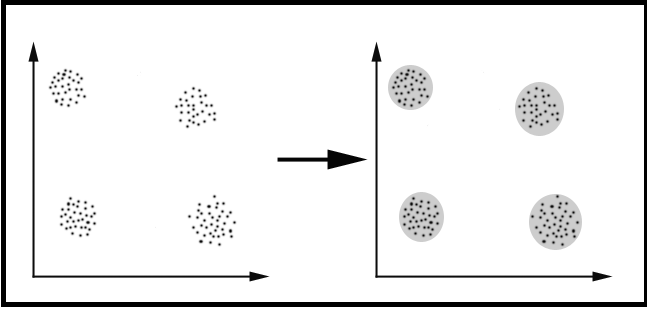
*Since the output is 1, the student will get admission.*.

GROUP 5

Clustering

Refer to pages 78 to 84 in the PDF file R\_Programming\_Lessons\_7

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.



A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

In the diagram, we easily identify the four clusters into which the data can be divided; the similarity criterion is distance; two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering.

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

**The goals of clustering**

*To determine the intrinsic grouping in a set of unlabeled data*

Some of them are:

1. Finding representatives for homogeneous groups (data reduction)
2. Finding natural clusters and describe their unknown properties (natural data types)
3. Finding useful and suitable groupings (useful data classes)
4. Finding unusual data objects (outlier detection)

# Possible applications

Clustering algorithms can be applied in many fields, for instance:

* ***Marketing:*** finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records
* ***Biology:*** Classification of plants and animals given their features
* ***Libraries:*** Book Ordering
* ***Insurance:*** Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds
* ***City-Planning:*** Identifying groups of houses according to their house type, value and geographical location.
* ***Earthquake studies:*** Clustering observed earthquake epicentres to identify dangerous zones

# Clustering algorithms

1. Exclusive clustering
2. Overlapping clustering
3. Hierarchical clustering
4. Probabilistic clustering

**The four most used clustering algorithms:**

1. ***K-means***
2. ***Fuzzy C-means***
3. ***Hierarchical clustering***
4. ***Mixture of Gaussians***

Each of these algorithms belongs to one of the clustering types listed above. K-means is an exclusive clustering algorithm. Fuzzy C-means is an overlapping clustering algorithm. Hierarchical is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm.



## K-Means clustering

This is a prototype-based, partitional clustering technique that attempts find a number of specified clusters (k), which are presented by their centroids (mean).

K-means algorithm proceeds in such a way that the elements are assigned randomly to k clusters and the centroid (mean) is calculated for each cluster. In the next step, the elements are reassigned in such a manner that it belongs to the cluster with closest centroid. This process is iterated until two consecutive steps end up in the same assignment of elements. In R package, k-means clustering is done using the function kmeans().

**Syntax:**

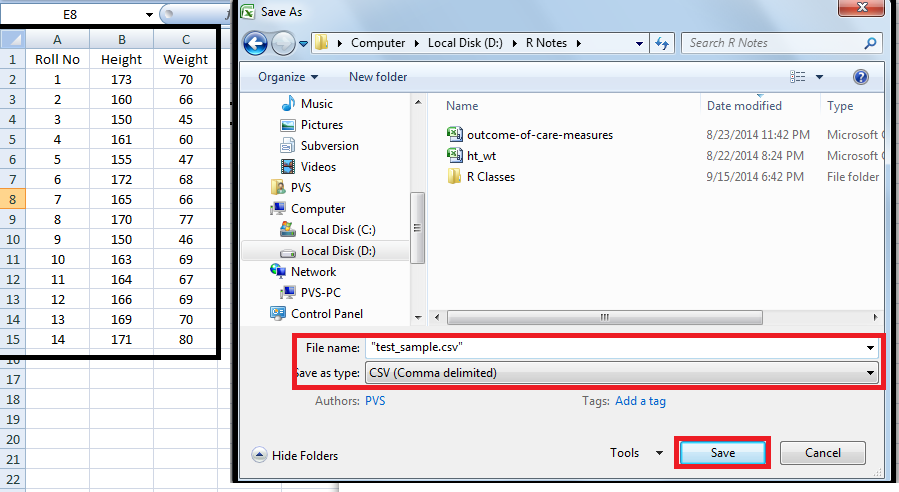
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c(“Hartigan-Wong”,”Lloyd”,”Forgy”,”MacQueen”), trace= FALSE)

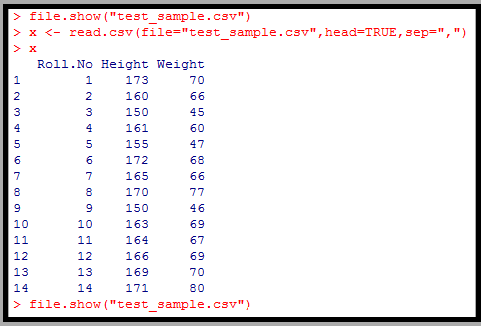
**Arguments**

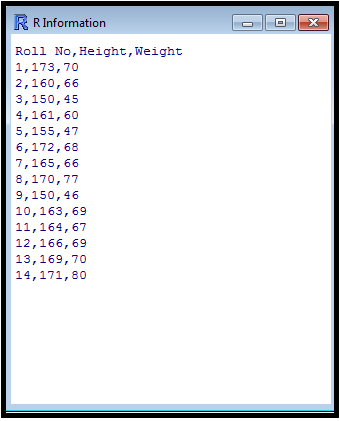
|  |  |
| --- | --- |
| x | Numeric matrix of data, or an object that can be coerced to such a matrix |
| centers | Either the number of clusters, say k or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers. |
| iter.max | The maximum number of iterations allowed |
| nstart | If centers is a number, how many random set should be chosen |
| Algothithm | Determines the algorithm to be used |
| Trace | A logical value which product the tracing information on the progress of the algorithm, if TRUE |

## Example 1

Let us illustrate this with the help of an example. We shall calculate the k-means of the height and weight of fourteen students in the file test\_sample.csv.



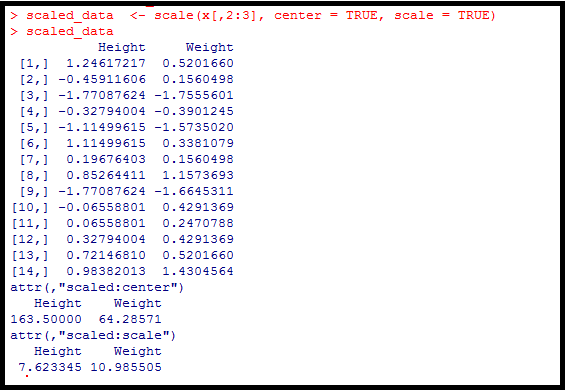


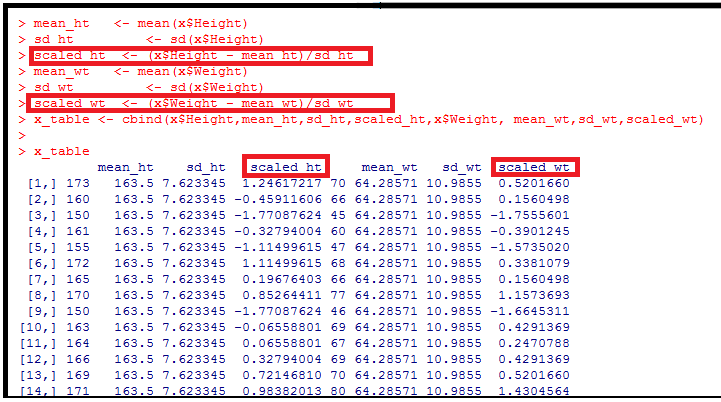


Since the data lies in different range, we normalize it using the function scale().

The function scale() centers and / or scales the columns of a numeric matrix. The attribute center determines whether centering should be done for the data. If center is TRUE, then centering is done. Similarly, the attribute scale determines whether scaling should be done for the centered data. Please note that the scaled height is arrived at by using the formula:

Height – mean(Height) / standard deviation (Height). Similarly, we do it for weight.





The function k\_means() will return an object of class kmeans with the following components:

***cluster***  A vector of integers indicating the cluster to which each point is allocated.

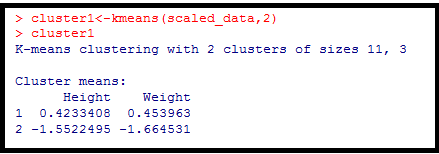
***centers***  A matrix of clusters centers

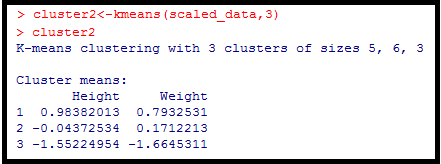
***cluster means*** Coordinates of the centroid (center) for each cluster

***withinss*** The measure of the total variance in our data set explained by the clustering. This is the difference between the squared distances of each data point to the centroid and the sum of squared distances of each cluster to the centroid.

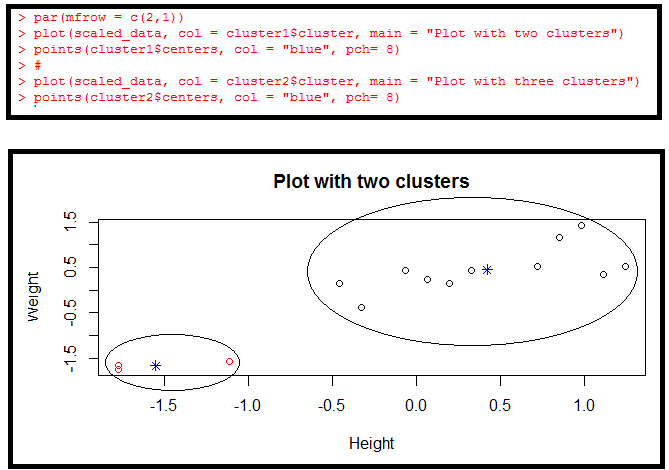
***size*** The number of points in each cluster

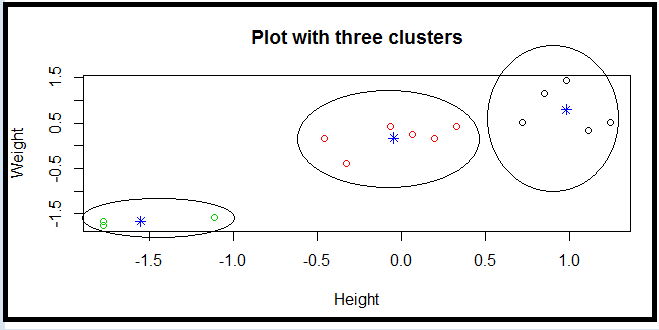
Now, we can apply the k-means() function to our scaled data. The first kmeans() generates two clusters while the second one generates three clusters.





To visualize the results of the k-means, generate plots.





**Note:**

1. The function par() enables to combine the plots of cluster1 and cluster2 in one single graph. Its attribute mfrow with values 2 and 1 enables to create a matrix of 2 rows and 1 column filled by row.
2. The function points() will indicate the location of the centroid in the plot. The attribute pch of the points() function denotes the plotting character to be used. Its value will be an integer code indicating the symbol.
3. The attribute col is meant for colour of the symbol.
4. We have circled the cluster of students having similar height and weight.

## Example 2

Consider the following data set consisting of the scores of two variables on each of seven individuals:

|  |  |  |
| --- | --- | --- |
| **Subject** | **A** | **B** |
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

This data set is to be grouped into two clusters. As a first step in finding an initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector  (Centroid) |
| Group 1 | 1 | (1.0, 1.0) |
| Group 2 | 4 | (5.0, 7.0) |

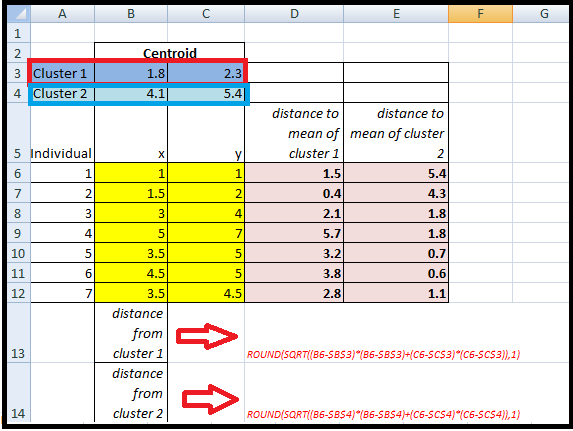
The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. Now, we arrive at the following steps:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 1 | | Cluster 2 | |
| Step | Individual | Mean Vector  (Centroid) | Individual | Mean Vector  (Centroid) |
| 1 | 1 | (1.0, 1.0) | 4 | (5.0, 7.0) |
| 2 | 1,2 | (1.2,1.5) | 4 | (5.0, 7.0) |
| 3 | 1,2,3 | (1.8,2.3) | 4 | (5.0, 7.0) |
| 4 | 1,2,3 | (1.8,2.3) | 4,5 | (4.2,6.0) |
| 5 | 1,2,3 | (1.8,2.3) | 4,5,6 | (4.3,5.7) |
| 6 | 1,2,3 | (1.8,2.3) | 4,5,6,7 | (4.1,5.7) |

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector  (Centroid) |
| Cluster 1 | 1,2,3 | (1.8,2.3) |
| Cluster 2 | 4,5,6,7 | (4.1, 5.4) |

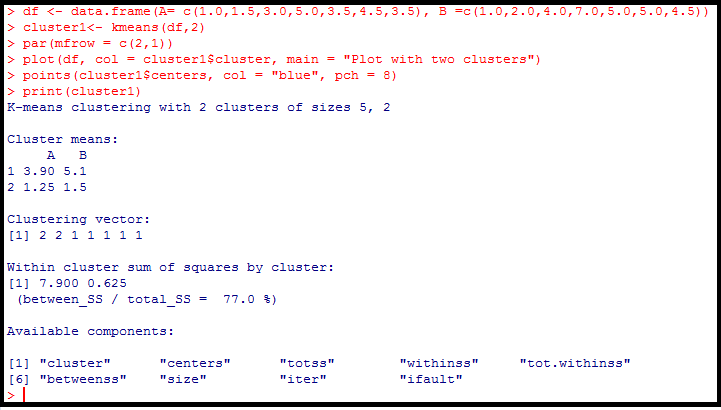
But, we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual’s distance to its own cluster mean and to that of the opposite cluster.

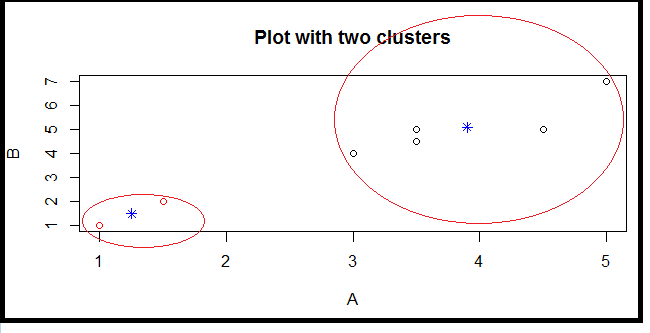


Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1). In other words, each individual’s distance to its own cluster mean should be smaller that the distance to the other cluster’s mean. It is different case for individual 3. Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector  (Centroid) |
| Cluster 1 | 1,2 | (1.3,1.5) |
| Cluster 2 | 3,4,5,6,7 | (3.9, 5.1) |

The iterative relocation would now continue from this new partition until no more relocations occur. However, in this example, each individual is now nearer to its own cluster mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution.





.

GROUP 6

Classification

Refer to pages 48 to 58 in the PDF file R\_Programming\_Lessons\_7.