CORRELATION, REGRESSION AND TEST OF SIGNIFICANCE IN R

B N Mandal
I.A.S.R.I., Library Avenue, New Delhi – 110 012
bnmandal @iasri.res.in

Introduction

This note shows how to perform correlation, linear regression analysis and test of significance in R. Throughout the note we will use some data sets from Design resources server at http://www.iasri.res.in/design.

Correlation analysis

Example**:** The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

sn	pp	ph	ngl	yld
1	142	0.525	8.2	2.47
2	143	0.64	9.5	4.76
3	107	0.66	9.3	3.31
4	78	0.66	7.5	1.97
5	100	0.46	5.9	1.34
6	86.5	0.345	6.4	1.14
7	103.5	0.86	6.4	1.5
8	155.99	0.33	7.5	2.03
9	80.88	0.285	8.4	2.54
10	109.77	0.59	10.6	4.9
11	61.77	0.265	8.3	2.91
12	79.11	0.66	11.6	2.76
13	155.99	0.42	8.1	0.59
14	61.81	0.34	9.4	0.84
15	74.5	0.63	8.4	3.87
16	97	0.705	7.2	4.47
17	93.14	0.68	6.4	3.31
18	37.43	0.665	8.4	1.57
19	36.44	0.275	7.4	0.53
20	51	0.28	7.4	1.15
21	104	0.28	9.8	1.08
22	49	0.49	4.8	1.83
23	54.66	0.385	5.5	0.76
24	55.55	0.265	5	0.43
25	88.44	0.98	5	4.08
26	99.55	0.645	9.6	2.83
27	63.99	0.635	5.6	2.57

28	101.77	0.29	8.2	7.42
29	138.66	0.72	9.9	2.62
30	90.22	0.63	8.4	2
31	76.92	1.25	7.3	1.99
32	126.22	0.58	6.9	1.36
33	80.36	0.605	6.8	0.68
34	150.23	1.19	8.8	5.36
35	56.5	0.355	9.7	2.12
36	136	0.59	10.2	4.16
37	144.5	0.61	9.8	3.12
38	157.33	0.605	8.8	2.07
39	91.99	0.38	7.7	1.17
40	121.5	0.55	7.7	3.62
41	64.5	0.32	5.7	0.67
42	116	0.455	6.8	3.05
43	77.5	0.72	11.8	1.7
44	70.43	0.625	10	1.55
45	133.77	0.535	9.3	3.28
46	89.99	0.49	9.8	2.69

We want to do following analysis on the above data.

1. Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield.
2. Obtain partial correlation between NGL and yield after removing the linear effect of PP and PH.
3. Give a scatter plot of the variable PP and yield.
4. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables. Print the matrices used in the regression computations.
5. Test the significance of the regression coefficients.
6. Obtain the predicted values corresponding to each observation in the data set.
7. Check for the linear relationship among the biometrical characters, i.e., multi-colinearity in the data.
8. Fit the multiple linear regression models without intercept.

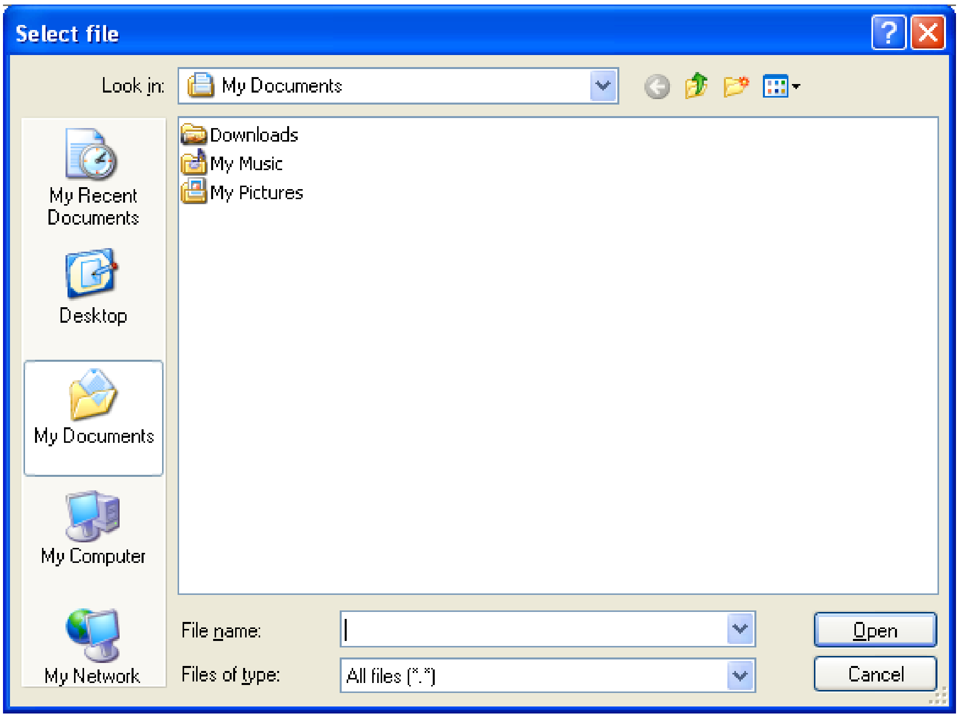
Importing the data:

First save the data file in comma delimited form. For this use “Save as” and then file type as “.csv” in MS Excel.

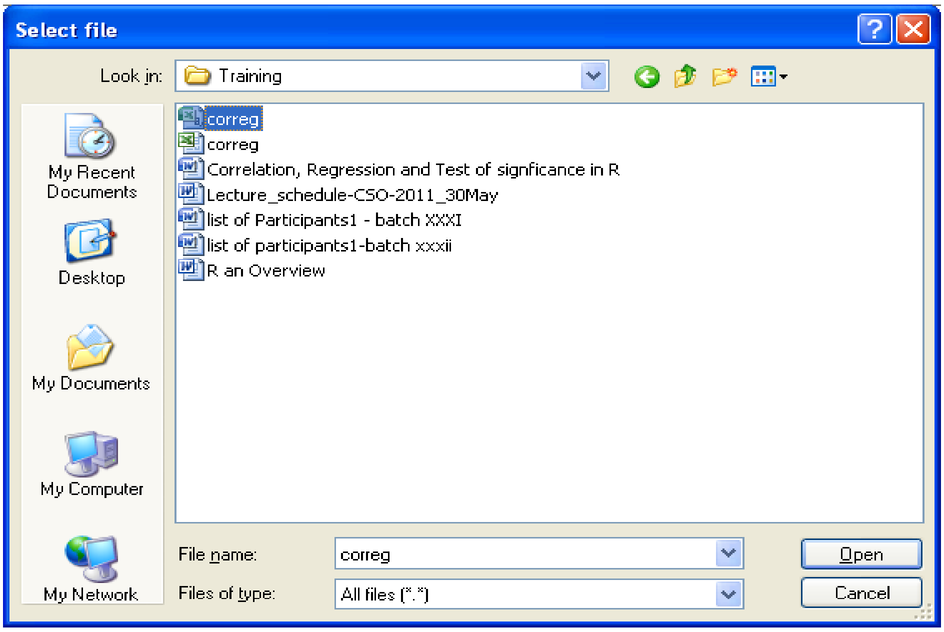
To read the data and name the data as ‘correg’, use the following command.

```
>correg=read.csv(file.choose())
```

Then a pop-up window appears as below.



Then select the ".csv" file and click on 'Open' to load the data into R environment.


Select file
Look in: Training
correg
correg
Correlation, Regression and Test of signficance in R
Lecture_schedule-CSO-2011_30May
list of Participants1 - batch XXXI
list of participants1-batch xxxii
R an Overview
My Recent Documents
Desktop
My Documents
My Computer
My Network
File name: correg
Files of type: All files (*.*)
Open
Cancel

Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,

```
>attach (correg)
```

causes R to directly read all the variables names eg. sn, pp, ph, ngl, yld etc. It is a good practice to use the *attach(dataset)* function immediately after reading the *datafile* into R.

Use the *cor()* function To obtain correlation coefficient between two variables. Three types of correlation coefficients namely Pearson, Kendall and Spearman Rank correlations can be computed through this function.

cor(var1, var2): The default correlation returns the Pearson correlation coefficient
cor(var1, var2, method = "spearman") returns Spearman correlation coefficient and *cor(dataset, method = "kendall"* returns Kendall correlation coefficients

```
> cor(pp,ph)
```

The correlation coefficient between the variables is then shown.

```
[1] 0.2396052
```

To obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield, use the following command.

```
> cor(correg)
```

This shows following result.

```
            sn         pp        ph       ngl        yld
sn  1.000000000 0.03247915 0.1126113 0.1378665 0.004169625
pp  0.032479150 1.00000000 0.2396052 0.2852699 0.385935933
ph  0.112611290 0.23960522 1.0000000 0.0886597 0.332323029
ngl 0.137866548 0.28526994 0.0886597 1.0000000 0.278784234
yld 0.004169625 0.38593593 0.3323230 0.2787842 1.000000000
```

To test the significance of the correlation between two variables, use *cor.test(var1,var2)* function.

```
> cor.test(pp,ph)
```

```
 Pearson's product-moment correlation
data:  pp and ph
t = 1.637, df = 44, p-value = 0.1088
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05448246  0.49544185
sample estimates:
      cor
0.2396052
```

Partial correlation

To obtain partial correlation between two variables given a third variable, we need to install a code for pcor.test which is available on web. After installing that code into R, one can use pcor.test(x,y,z) to obtain partial correlation between x and y given the variables in z matrix.

Therefore, to obtain partial correlation between NGL and yield after removing the linear effect of PP, type the following command.

```
>pcor.test(ngl,yld,pp)
```

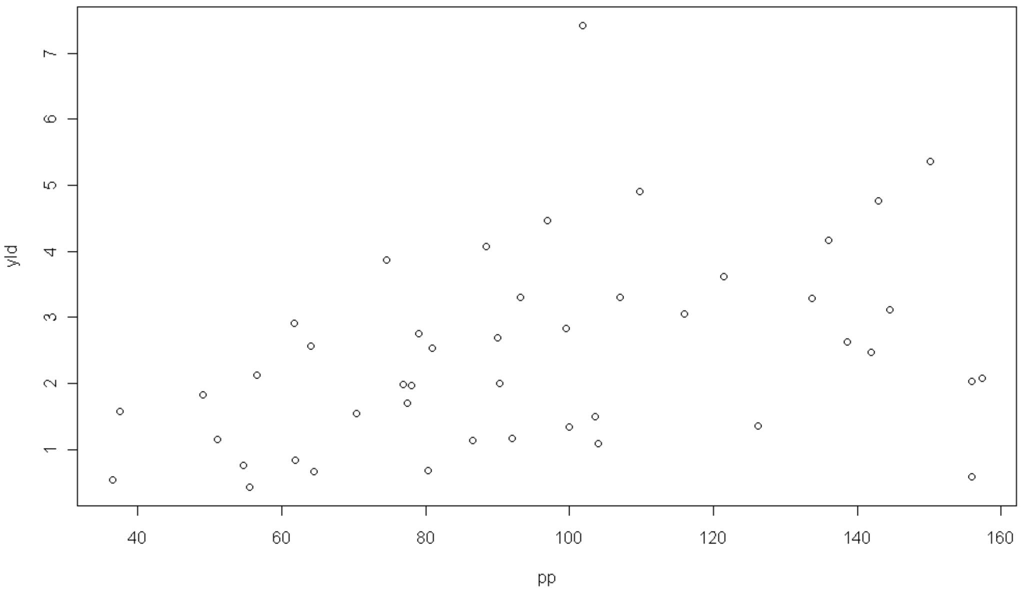
This gives following output.

```
   estimate  p.value statistic  n gn  Method            Use
1 0.1907824 0.202503  1.274453 46  1 Pearson Var-Cov matrix
```

Scatter plot

To have a feel of the correlation between two or more variables visually, scatter plot may be seen. To obtain scatter plot of two variables in R, use *plot(var1, var2)* function.

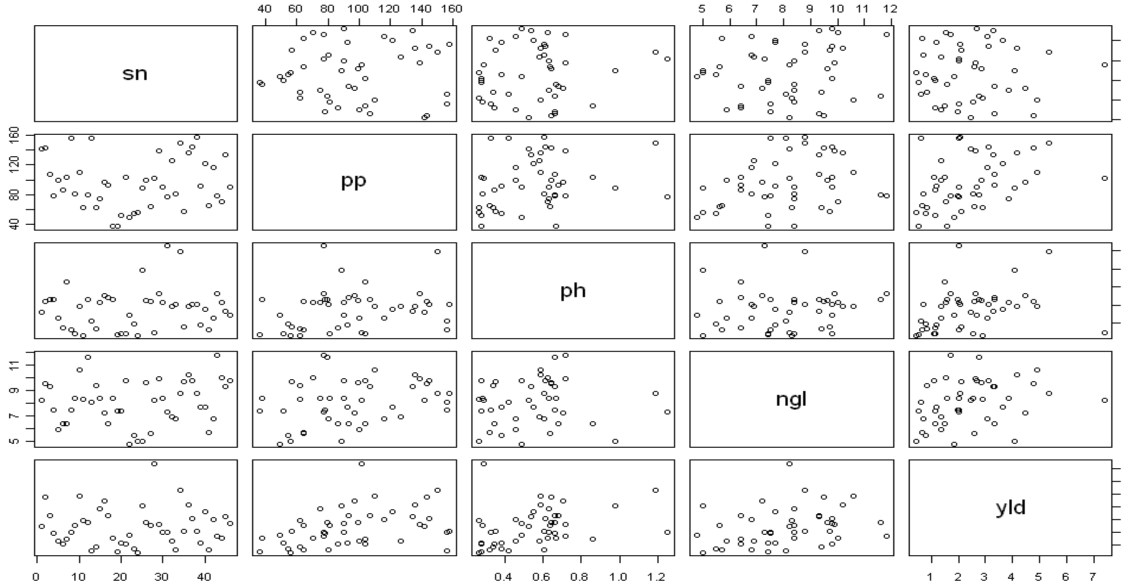
```
>plot(pp,yld)
```



Alternatively, if there are more than two variables, then *pairs()* function produces pairwise scatter plot matrix.

```
>  pairs (correg)
```

This gives following output.



Linear Regression Analysis

The basic form of a linear regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$.

Given data on the variables *Y, X$_1$, X$_2$, …, X$_p$,* estimation of parameters can be done using R. The function used for regression analysis in R is *lm(y ~ x1 + x2 + … +xp)*

There are a number of options available in R, depending upon your data set. To fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables, use following commands.

```
> out=lm(yld~pp+ph+ngl)
> out

Call:
lm(formula = yld ~ pp + ph + ngl)

Coefficients:
(Intercept)         pp          ph          ngl
    -0.8480      0.0120      1.6606      0.1514
```

The results of the analysis are stored in a variable called ‘out’. One can give any name to this variable.

summary() gives other information related to the regression analysis.

```
> summary(out)

> summary(out)

Call:
lm(formula = yld ~ pp + ph + ngl)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3568 -0.8308 -0.3055  0.7415  5.3243

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.848019   1.054337  -0.804   0.4257
pp           0.011995   0.006284   1.909   0.0631 .
ph           1.660605   0.918814   1.807   0.0779 .
ngl          0.151390   0.119406   1.268   0.2118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 42 degrees of freedom
Multiple R-squared: 0.2391,     Adjusted R-squared: 0.1848
F-statistic: 4.399 on 3 and 42 DF,  p-value: 0.008851
```

From the output we see that the intercept is estimated to -0.848019 with a standard error of 1.054337.

The hypothesis of that H_0 : b1 = 0 (slope equal to zero) is tested as a *t*-test. From the output we see that the *t*-statistic is -0.804 and that the (two-sided) *P*-value is 0.4257 ·

Confidence intervals for the regression parameters may be computed by the function confint. As default confint computes 95% confidence intervals but other levels are obtained by the option level.

For example,

```
> confint(out)
                    2.5 %     97.5 %
(Intercept) -2.9757580826 1.27971946
pp          -0.0006867137 0.02467732
ph          -0.1936356077 3.51484604
ngl         -0.0895818297 0.39236142
```

shows that the 95% confidence intervals

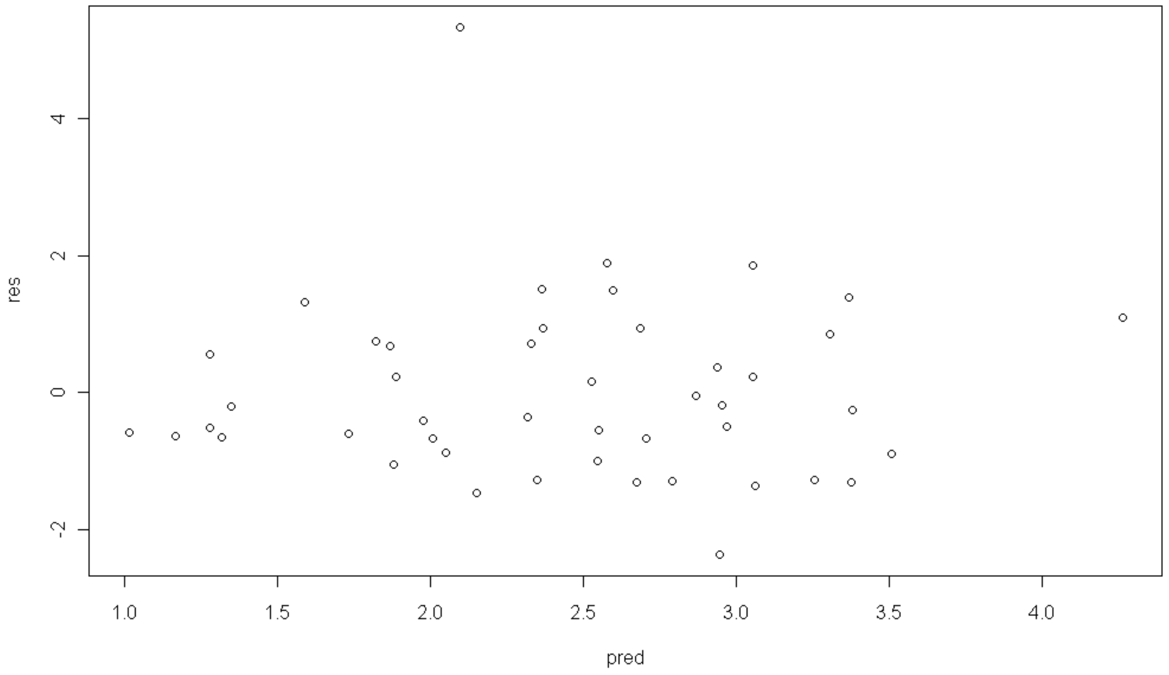
Model control

The predicted values and residuals are extracted from linreg as follows:

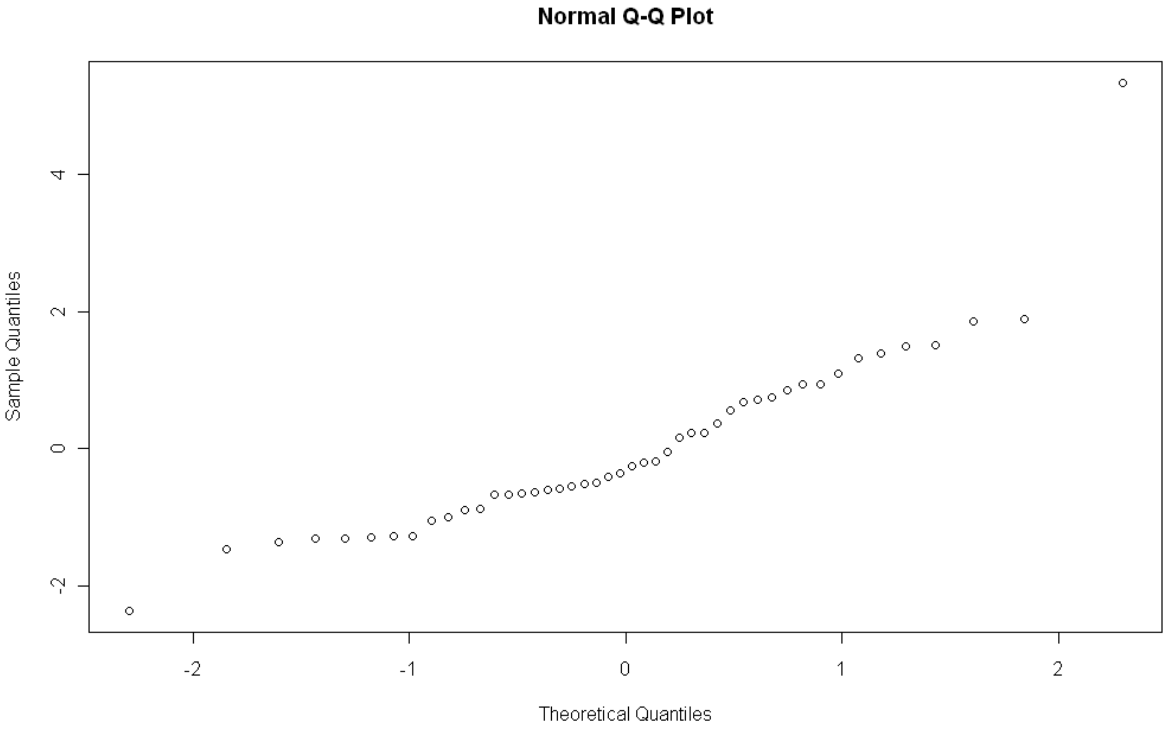
```
> pred=predict(out)
> pred
       1        2        3        4        5        6        7        8        9       10       11       12       13 
2.968528 3.368300 2.939403 2.319037 2.008589 1.731378 2.790510 2.706552 1.867108
3.053194 1.589526 2.953050 2.946840
      14       15       16       17       18       19       20       21       22       23       24       25       26 
1.881080 2.363487 2.576259 2.367330 1.976942 1.166041 1.348995 2.348082 1.280118
1.279621 1.015329 2.597188 2.870546
      27       28       29       30       31       32       33       34       35       36       37       38       39 
1.821827 2.095715 3.509644 2.552053 3.255562 2.673769 2.150040 4.262386 1.887711
3.307275 3.381892 3.376098 2.052160
      40       41       42       43       44       45       46 
2.688445 1.319993 2.328462 3.063652 2.548586 3.052942 2.528755

> res=resid(out)
> res
          1           2           3           4           5           6           7           8           9          10 
-0.49852813  1.39170023  0.37059708 -0.34903743 -0.66858943 -0.59137811 -1.29050999 -
0.67655158  0.67289223  1.84680572
         11          12          13          14          15          16          17          18          19          20 
 1.32047361 -0.19305037 -2.35683993 -1.04108037  1.50651348  1.89374147  0.94267032 -
0.40694172 -0.63604054 -0.19899521
         21          22          23          24          25          26          27          28          29          30 
-1.26808192  0.54988177 -0.51962097 -0.58532927  1.48281240 -0.04054575  0.74817255
5.32428523 -0.88964448 -0.55205272
         31          32          33          34          35          36          37          38          39          40 
-1.26556162 -1.31376877 -1.47004020  1.09761415  0.23228869  0.85272477 -0.26189151 -
1.30609846 -0.88216025  0.93155540
         41          42          43          44          45          46 
-0.64999340  0.72153789 -1.36365220 -0.99858627  0.22705840  0.16124522
```

Hence, the residual plot and the normal probability plot (QQ-plot) for the residuals are constructed
as follows:

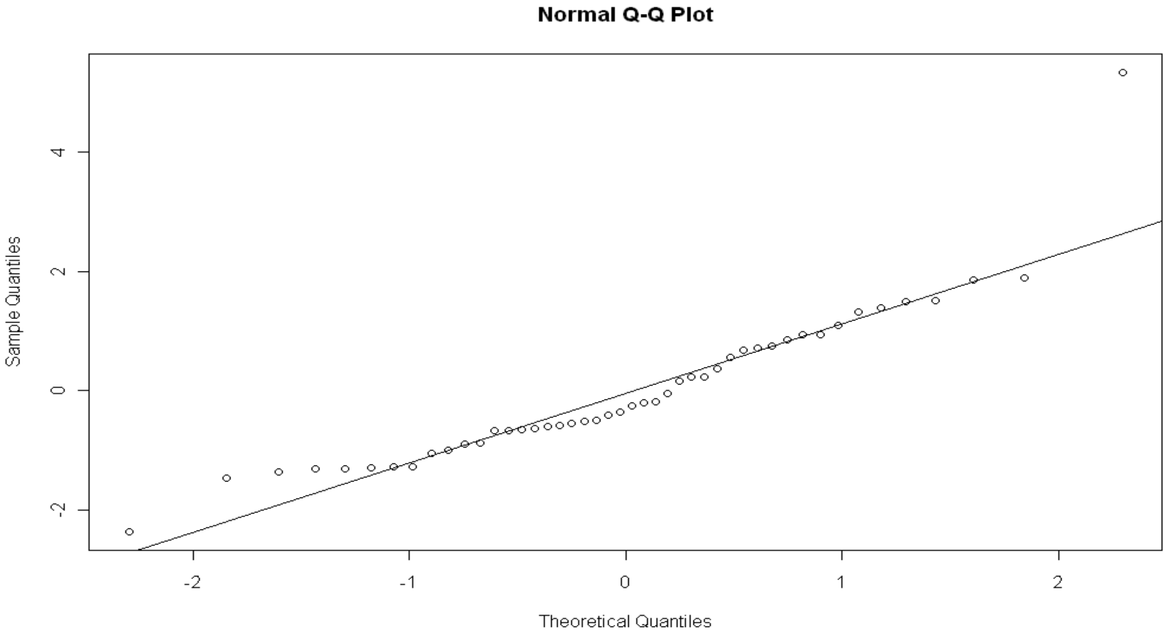
```
> plot(pred,res)
```



```
> qqnorm(res)
```



```
> qqline(res)
```



The plots do give rise to serious objections against the model assumptions for this data set.

To fit a regression line through origin use the following command.

```
> out=lm(yld~-1+pp+ph+ngl)
> out

Call:
lm(formula = yld ~ -1 + pp + ph + ngl)

Coefficients:
     pp       ph      ngl  
0.01093  1.41342  0.07956
```

Test of significance

t-test

We took the following data from design of resources server.

Example 1: An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria (Mol) Standl)* Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination {*the male flowers pinched from the seed parent before the anthesis regularly with utmost care to avoid the chance selfing. The pollination is carried out by the natural pollinating agent*} and hand pollination {*the male flowers also pinched from the seed parent before the anthesis regularly. The female buds are covered with butter paper bag which contain 5-6 tiny hole to felicitate the ventilation and to avoid the built up of high temperature in size the butter paper bag. The butter paper bag is clipped/ stippled. On the same day the male* *bud at pollen parent (male plant) are also covered with butter paper bag. On the next day the male bud are removed and the anthers are rubbed gently over the all three lobes. The female flower is again covered with butter paper bag and label is placed over the peduncle of pollinated female flower (plate 4, 5, 6 &7). The pollination is performed at noon (1-3pm)*} under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

group	nfs45	fw	syp	sl
1	7	1.85	147.7	16.86
1	7	1.86	136.86	16.77
1	6	1.83	149.97	16.35
1	7	1.89	172.33	18.26
1	7	1.8	144.46	17.9
1	6	1.88	138.3	16.95
1	7	1.89	150.58	18.15
1	7	1.79	140.99	18.86
1	6	1.85	140.57	18.39
1	7	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.5	18.07
2	7.3	2.58	230.34	19.07
2	8	2.62	217.05	19
2	8	2.68	233.84	18
2	8	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7	2.45	199.87	19.31
2	7.3	2.44	214.3	19.36

1. Test whether the mean of the population of Seed yield/plant (g) is 200 or not.
2. Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.
3. Test whether hand pollination is better alternative in comparison to natural pollination.

Reading the data

```
> data1=read.csv(file.choose())

> attach(data1)

> names(data1)

[1] "group" "nfs45" "fw"    "syp"   "sl"
```

For one sample t test, use *t.test(var1)* function. By default, the t test result produced is for two sided tests with 5% level of significance.

```
> t.test(syp,mu=200)

        One Sample t-test

data:  syp

t = -2.3009, df = 19, p-value = 0.03289

alternative hypothesis: true mean is not equal to 200

95 percent confidence interval:

 163.3414 198.2656

sample estimates:

mean of x

 180.8035
```

The two sample t-test is used to know whether two samples have same mean or not. To perform a t-test use *t.test(var1, var2)* function.

```
> gr1=subset(data1,group==1)
> gr1
   group nfs45   fw    syp    sl
1      1     7 1.85 147.70 16.86
2      1     7 1.86 136.86 16.77
3      1     6 1.83 149.97 16.35
4      1     7 1.89 172.33 18.26
5      1     7 1.80 144.46 17.90
6      1     6 1.88 138.30 16.95
7      1     7 1.89 150.58 18.15
8      1     7 1.79 140.99 18.86
9      1     6 1.85 140.57 18.39
10     1     7 1.84 138.33 18.58

> gr2=subset(data1,group==2)
> gr2
   group nfs45   fw    syp    sl
11     2   6.3 2.58 224.26 18.18
12     2   6.7 2.74 197.50 18.07
13     2   7.3 2.58 230.34 19.07
```

```
14     2   8.0 2.62 217.05 19.00
15     2   8.0 2.68 233.84 18.00
16     2   8.0 2.56 216.52 18.49
17     2   7.7 2.34 211.93 17.45
18     2   7.7 2.67 210.37 18.97
19     2   7.0 2.45 199.87 19.31
20     2   7.3 2.44 214.30 19.36
> t.test(gr1$syp,gr2$syp)

        Welch Two Sample t-test
data:  gr1$syp and gr2$syp
t = -13.9583, df = 17.771, p-value = 5.136e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -80.07285 -59.10515
sample estimates:
mean of x mean of y
  146.009   215.598

> t.test(gr1$syp,gr2$syp,alternative='greater',var.equal=TRUE)
        Two Sample t-test
data:  gr1$syp and gr2$syp
t = -13.9583, df = 18, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -78.23419       Inf
sample estimates:
mean of x mean of y
  146.009   215.598
```

For paired t test, use *t.test(var1, var2, paired=T)* function.

```
> t.test(x,y, paired=T)
```

Analysis of Variance

The analysis of variance is a commonly used method to determine whether several samples have come from a population with same mean or not. R provides a function to conduct ANOVA so: aov(model, data)

One-way ANOVA

Example: {Nigam, A.K. and Gupta V.K., 1979, *Handbook on Analysis of Agricultural experiments,* First Edition, I.A.S.R.I. Publication, New Delhi, pp16-20}.A feeding trial with 3 feeds namely (i) Pasture(control), (ii) Pasture and Concentrates and (iii) Pasture, Concentrates and Minerals was conducted at the Yellachihalli Sheep Farm, Mysore, to study their effect on wool yield of Sheep. For this purpose twenty-five ewe lambs were allotted at random to each of the three treatments and the three treatments and the weight records of the total wool yield (in gms) of first two clipping were obtained. The data for two lambs for feed 1, three for feed 2 and one for feed 3 are missing. The details of the experiment are given below:

trt	yld	trt	yld	trt	yld
1	850.5	2	510.3	3	850.5
1	453.6	2	963.9	3	1474.2
1	878.85	2	652.05	3	510.3
1	623.7	2	1020.6	3	850.5
1	510.3	2	878.85	3	793.8
1	765.45	2	567	3	453.6
1	680.4	2	680.4	3	935.55
1	595.35	2	538.65	3	1190.7
1	538.65	2	567	3	481.95
1	850.5	2	510.3	3	623.7
1	850.5	2	425.25	3	878.85
1	793.8	2	567	3	1077.3
1	1020.6	2	623.7	3	850.5
1	708.75	2	538.65	3	680.4
1	652.05	2	737.1	3	737.1
1	623.7	2	453.6	3	737.1
1	396.9	2	481.95	3	708.75
1	822.15	2	368.55	3	708.75
1	680.4	2	567	3	652.05
1	652.05	2	595.35	3	567
1	538.65	2	567	3	453.6
1	850.5	2	595.35	3	652.05
1	680.4	3	992.25	3	567

where Feed 1- Pasture (control),
Feed 2- Pasture and Concentrates and
Feed 3- Pasture, Concentrates and Minerals.

1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.
2. Perform all possible pair wise treatment comparisons and identify the best treatment i.e. the treatment giving highest yield.

As usual first read and attach the data in R. Then check whether the ‘trt’ variable is numeric or factor. To check this, use the following command.

```
> is.factor(trt)

[1] FALSE
```

```
> is.numeric(trt)

[1] TRUE
```

This shows that ‘trt’ is a numeric variable and hence we need to convert this to a factor variable. Otherwise, analysis will be wrong. To convert the variable ‘trt’ from numeric to a factor variable, use *as.factor()* function.

```
> trt=as.factor(trt)

> trt

 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3

[61] 3 3 3 3 3 3 3 3 3

Levels: 1 2 3

> is.factor(trt)

[1] TRUE
```

Now, use *aov()* function with model specification *y~Group* and store the results in anova_output.

```
> anova_output=aov(yld~trt)
> anova_output
Call:
   aov(formula = yld ~ trt)

Terms:
                      trt Residuals
Sum of Squares   287872.4 2460182.9
Deg. of Freedom         2        66

Residual standard error: 193.0686
Estimated effects may be unbalanced
```

To get more information on the analysis done, use *summary()* function.

```
> summary(anova_output)

            Df  Sum Sq Mean Sq F value  Pr(>F)
trt          2  287872  143936  3.8614 0.02595 *
Residuals   66 2460183   37275
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Post Hoct Tests

To conduct post-hoc tests, R provides a simple function to carry out the Tukey HSD test.

```
> TukeyHSD(anova_output)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yld ~ trt)

$trt
         diff        lwr       upr     p adj
2-1 -86.89891 -224.94944  51.15161 0.2931780
3-1  71.38859  -63.68998 206.46716 0.4186377
3-2 158.28750   21.65028 294.92472 0.0192552
```

To get all the entries stored in ‘parameters’ data frame, use *names()* function.

```
> names(anova_output)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"        "qr"
 [8] "df.residual"   "contrasts"     "xlevels"       "call"          "terms"         "model"
```

To carry out two-way and other complex analysis of variance, the model may be modified as per situations.

Model	Interpreation
y~x	Dependent variable y, explained by one factor x (one way ANOVA)
y~x_1 + x_2	y is explained by two factors x_1 and x_2 (two way ANOVA)
y~x_1*x_2	y is explained by two factors x_1 and x_2 as well as their interactions
y~ x_1 + x_2 + x_1:x_2	-Do-

References:

Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed on 12.05.2011)

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0,
URL http://www.R-project.org/.