Chapter V

Correlation and Regression Analysis

B. Nirmala[1], Santosh Patil[2] and Santosha Rathod[1]

[1]ICAR- Indian Institute of Rice Research, Hyderabad-500030
[2]AEC & RI, Tamil Nadu Agricultural University,
Coimbatore - 641 003.
Email: nirmala.b@icar.gov.in

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. Scatter plot of two variables will visualize the relationship among them.

Pearson correlation coefficient

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables

The formula for the correlation is:

$$r = \frac{\mathrm{N}\sum \mathrm{xy} - \sum \mathrm{x}\sum \mathrm{y}}{\sqrt{(N\sum x2 - (\sum x)2)(N\sum y2 - (\sum y)2)}}$$

where:

N is the number of pairs of scores

∑xy is the sum of the products of paired scores

∑x is the sum of x scores

∑y is the sum of y scores

$\sum x^2$ is the sum of squared x scores

$\sum y^2$ is the sum of squared y scores

The Pearson correlation coefficient is used when (1) the relationship is linear and (2) both variables are quantitative and (3) normally distributed and (4) have no outliers (Gupta & Kapoor, 2020).

The **cor()** function is used to calculate the Pearson correlation coefficient in R. To test the significance of the correlation, the **cor.test()** function is used (Loftus, 2021). We use the symbol r to stand for the correlation. R will always be between -1.0 and + 1.0. If the correlation is negative, we have a negative relationship; if it's positive, the relationship is positive.

Illustration

Table1: The correlation between the height of a person and self-confidence of 20 persons.

Person	Height (x)	Self Confidence (y)	xy	xx	yy
1	68	4.1	278.8	4624	16.81
2	71	4.6	326.6	5041	21.16
3	62	3.8	235.6	3844	14.44
4	75	4.4	330	5625	19.36
5	58	3.2	185.6	3364	10.24
6	60	3.1	186	3600	9.61
7	67	3.8	254.6	4489	14.44
8	68	4.1	278.8	4624	16.81
9	71	4.3	305.3	5041	18.49
10	69	3.7	255.3	4761	13.69
11	68	3.5	238	4624	12.25
12	67	3.2	214.4	4489	10.24
13	63	3.7	233.1	3969	13.69
14	62	3.3	204.6	3844	10.89
15	60	3.4	204	3600	11.56
16	63	4	252	3969	16
17	65	4.1	266.5	4225	16.81
18	67	3.8	254.6	4489	14.44
19	63	3.4	214.2	3969	11.56
20	61	3.6	219.6	3721	12.96
Sum	1308	75.1	4937.6	85912	285.45

N=20, $\sum xy$=4937.6, $\sum x$=1308, $\sum y$=75.1, $\sum x2$=85912, $\sum y2$=285.45 and R=0.73

So, the correlation for our twenty cases is 0.73, which is a fairly strong positive relationship. As in all hypothesis testing, we need to first determine the significance level. We may use the common significance level of alpha = 0.05. This means that we are conducting a test where the odds that the correlation is a chance occurrence is no more than 5 out of 100. We have to compute the degrees of freedom or df. The df is simply equal to N-2 or, in this example, is 20-2 = 18. Finally, we have to decide whether we are doing a one-tailed or two-tailed test. In this example, since we have no strong prior theory to suggest whether the relationship between height and self-confidence would be positive or negative, we'll opt for the two-tailed test. With these three pieces of information – the significance level (alpha = 0.05)), degrees of freedom (df = 18), and type of test (two-tailed) – we can now test the significance of the correlation we found. If we look up this value in the table at the back of any statistics book, we find that the critical value is .4438. This means that if the correlation is greater than .4438 or less than -.4438, we can conclude that the odds are less than 5 out of 100 that this is a chance occurrence. Since our correlation of .73 is actually quite a bit higher, we conclude that it is not a chance finding and that the correlation is "statistically significant". We can reject the null hypothesis and accept the alternative.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, r_s shows the correlation between two ordinal data. How one ordinal data changes as the other ordinal changes.

The formula for the Spearman rank correlation coefficient when there are no tied ranks is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

R Language provides two methods to calculate the correlation coefficient. By using the functions **cor()** or **cor.test()** it can be calculated. It can be noted that **cor()** computes the correlation coefficient whereas **cor.test()** computes test for association or correlation between paired samples. It returns both the correlation coefficient and the significance level (or p-value) of the correlation.

cor ( ) computes the correlation coefficient

cor.test () test for association/correlation between paired samples. It returns both the correlation coefficient and the significance level (or p-value) of the correlation.

cor(x, y, method = c("pearson", "spearman"))

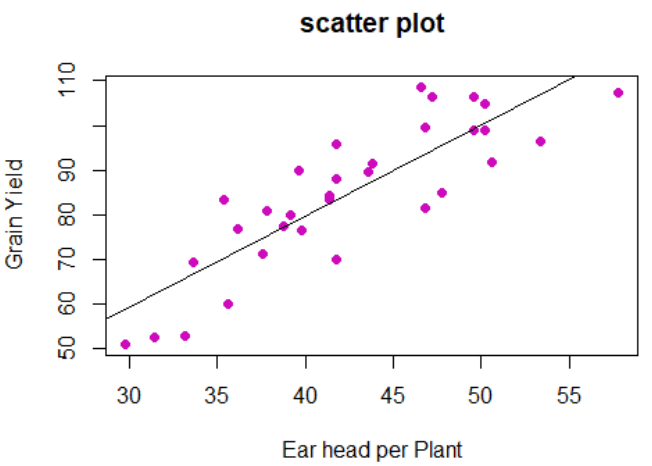
cor.test(x, y, method=c("pearson", "spearman"))

x, y: numeric vectors with the same length

method: correlation method.

Example:

Use singh.txt data
In R console **aa<- singh**
plot(grainyield~ear_plnt ,data=aa, main="scatter plot",xlab="Ear head per
Plant",ylab="Grain Yield",pch=19,col="62")



From the figure we can say that there is positive correlation between Ear head per Plant and Grain Yield (as number of Ear head per Plant increases the Grain Yield also increases).

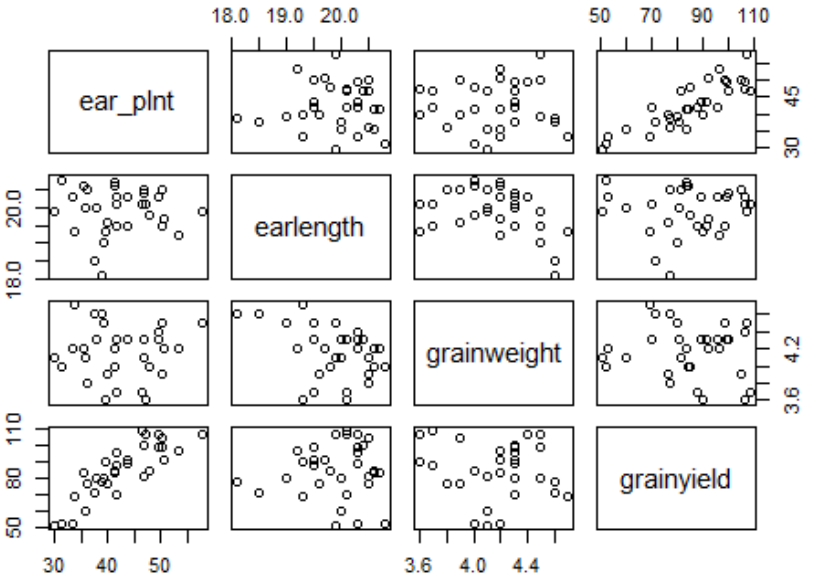
Pearson correlation coefficient calculated as

>cor(aa$grainyield,aa$ear_plnt, method ="pearson")
[1] 0.8514075

Here it is positive and high correlation between Ear head per Plant and Grain Yield i.e. r=0.85

If want to see relationship for entire data set one can use code

>pairs(aa[,-c(1:2)])



Pearson correlation for all variables together will be found by using code

>cor(aa[,-c(1:2)])

	ear_plnt	earlength	grainweight	grainyield
ear_plnt	1	0.021	0.022	0.851
earlength	0.021	1	-0.362	0.058
grainweight	0.022	-0.362	1	-0.111
grainyield	0.851	0.058	-0.111	1

 # Note: Correlation coefficients of types "pearson" (default), "kendall", or "spearman" can be obtained by using method.

>cor(aa[,-c(1:2)],method = "kendall")

Testing Correlation coefficient

(a) Null hypothesis: true correlation $\rho_{xy}= 0$
(b) Alternative hypothesis: true correlation$\rho_{xy} \neq 0$

If p-value is less than 0.05 then reject null hypothesis i.e. Correlation between X and Y variable is significant at 5 % level of significance.

```
>cor(aa$ear_plnt,aa$grain_yld, method ="pearson")
[1] 0.8514075
>cor.test(aa$ear_plnt,aa$grain_yld, method ="pearson")
            Pearson's product-moment correlation
data:  aa$ear_plnt and aa$grain_yld
t = 8.891, df = 30, p-value = 6.563e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7149760 0.9253754
sample estimates:
cor
0.8514075
```

Interpretation

(c) Karl Pearson's Correlation coefficient= 0.85
(d) Here p-value = 6.563e-10 = 6.563×10^{-6}
(e) The p-value is less than 0.01 so the correlation coefficient between grain yield and Ear head per plant is significantly different than 0

```
>cor(aa[,-c(1:2)], method ="pearson")
ear_plntear_lnthgrainwtgrain_yld
ear_plnt  1.00000000  0.02109627  0.02233878  0.8514075
ear_lnth  0.02109627  1.00000000 -0.36191303  0.0579030
grainwt   0.02233878 -0.36191303  1.00000000 -0.1108359
grain_yld 0.85140745  0.05790300 -0.11083586  1.0000000
```

A simple way of generating correlation with probability values is obtainedfor entire data set using the **Hmisc** package (Harrell, 2020).

```
> require("Hmisc")
>rcorr(as.matrix(aa[,-c(1:2)]), type="pearson")
ear_plntearlengthgrainweightgrainyield
ear_plnt       1.00      0.02       0.02       0.85
earlength      0.02      1.00      -0.36       0.06
grainweight    0.02     -0.36       1.00      -0.11
grainyield     0.85      0.06      -0.11       1.00
n= 32
```

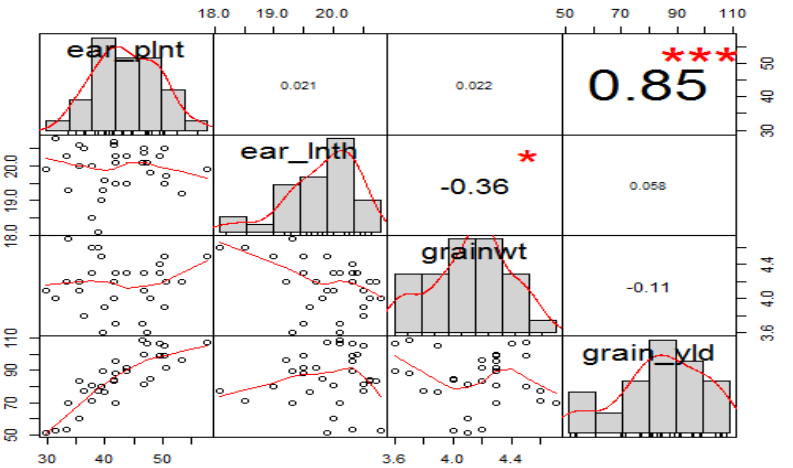
Probability values

ear_plntearlengthgrainweightgrainyield
ear_plnt 0.9088 0.9034 0.0000
earlength 0.9088 0.0418 0.7529
grainweight 0.9034 0.0418 0.5459
grainyield 0.0000 0.7529 0.5459

Draw scatter plots with correlation and its significance

> library("PerformanceAnalytics")

>chart.Correlation(aa[,-c(1:2)], histogram=TRUE, pch=19)



Regression Analysis

Regression Analysis is one of the most widely used statistical techniques in order to understand and model the relationship between two or more variables. It is used when you want to predict a continuous dependent variable from a number of independent variables. Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other. Regression analysis is often used to model or analyze data. It is used to understand the relationship between the variables, which can be further utilized to predict the precise outcome. In its simplest form regression analysis is very similar to correlation; in fact, the underlying mathematical models

are virtually identical. Regression analysis can, however, be used where there are many explanatory variables and where various data types are used together (Draper & Smith, 1998). The general regression model is:

$$Y = a + bx + e$$

Where *a* is a constant, X_1, X_2, *etc*. are the predictor variables, and the error term *e* is the difference between the observed and predicted value of Y.

The terminology associated with regression analysis is:

1. **Dependent variable or target variable:** Variable to predict.
2. **Independent variable or predictor variable:** Variables to estimate the dependent variable.
3. **Outlier:** Observation that differs significantly from other observations. It should be avoided since it may hamper the result.
4. **Multicollinearity:** Situation in which two or more independent variables are highly linearly related.
5. **Homoscedasticity or homogeneity of variance:** Situation in which the error term is the same across all values of the independent variables.

Regression analysis is primarily used for two distinct purposes. First, it is widely used for prediction and forecasting, which overlaps with the field of machine learning. Second, it is also used to infer causal relationships between independent and dependent variables.

1. The simplest case of linear regression is to find a relationship using a linear model (i.e line) between an input independent variable (input single feature) and an output dependent variable. This is called Bivariate Linear Regression.
2. On the other hand, when there is a linear model representing the relationship between a dependent output and multiple independent input variables is called Multivariate Linear Regression.
3. The dependent variable is continuous and independent variables may or may not be continuous. We find the relationship between

them with the help of the best fit line which is also known as the Regression line.

The regression equation of a line is given as;

$$y = m * x + b$$

Where,
x is Independent Variable
y is Dependent Variable
m is Slope of Line
b is y Intercept

To evaluate the best fit line, the most common method is the Least Square Method. In this method, the regression line is calculated by minimizing the least squared error between the regression line and the data points. Another method to find this line is also called the R Squared analysis. It is particularly useful when the relationship between the input variables and the output is not very complex. Also, it is very sensitive to outliers.

Multiple regression analysis is used to see if there is a statistically significant relationship between sets of variables. It's used to find trends in those sets of data.
Multiple regression analysis is *almost* the same as simple linear regression (Uyanık & Güler, 2013). The only difference between simple linear regression and multiple regression is in the number of predictors ("X" variables) used in the regression.

- Simple regression analysis uses a single x variable for each dependent "Y" variable. For example: (X_1, Y_1).
- Multiple regression uses multiple "X" variables for each independent variable: $(X_1)_1$, $(X_2)_1$, $(X_3)_1$, Y_1).

Multicollinearity

Linear Regression is a supervised learning algorithm used for continuous variables. When a Linear Regression model is built, there is a chance that some variables can be multicollinear in nature. Multicollinearity is a statistical terminology where more than one independent variable is correlated with each other. This multicollinearity results in reducing the reliability of statistical inferences. Multicollinearity in a regression model analysis occurs when two or more independent predictor variables are highly

correlated to each other, which results in the lack of unique information about the regression model. Hence, these variables must be removed when building a multiple regression model. Multicollinearity is a common problem in econometrics. The multicollinearity arises when we have too few observations to precisely estimate the effects of two or more highly correlated variables on the dependent variable. Graphically we can illustrate the problem of multicollinearity using Venn-diagrams. The Venn-diagrams below all represent the following regression model.

$$y = X_1 + X_2 + \in$$

Each circle depicts the variance of one variable of the regression model. That is, the circle y depicts the variance of the dependent variable y, the circle x_1 depicts the variance of variable x_1 and the circle x_2 shows the variance of the variable x_2. The overlapping areas show variation that variables have in common. For instance, the overlapping area of variable y and variable x_1 represents the variation of variable y that can be explained by variable x_1.

In first figure, the circles x_1 and x_2 both intersect with the circle y. However, there is no overlap between the circle x_1 and the circle x_2. In this case, variable x_1 and variable x_2 are both correlated with variable y, but the two explanatory variables themselves are uncorrelated. Thus, one can precisely identify the effect of each explanatory variable (X_1 and X_2) on the independent variable (Y).

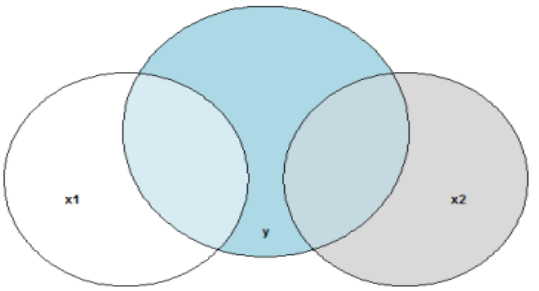


Figure1: Multicollinearity: Variable x_1 and x_2 are uncorrelated

Figure 2 shows a case in which there exists some correlation between the two explanatory variables. Note that, in Figure 2 there exists some overlap between the circle x_1 and the circle x_2 meaning that the two variables have some variation in common. You see that

it becomes less clear to determine what the effect of one explanatory variable on the dependent variable actually is, i.e. there is some area overlapping all three variables. Although there exists some correlation between variable x_1 and x_2, there is still enough variation left to determine the effect of x_1 and x_2 rather precisely.

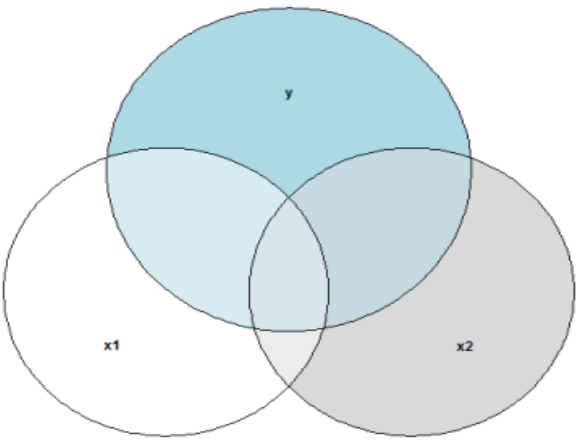


Figure 2: Moderate Multicollinearity: Variable x_1 and x_2 are somewhat correlated

Moderate multicollinearity is not much of a concern. However, if the correlation between two or more explanatory variables is very strong is get continuously harder to precisely estimate the pure effect of one explanatory variable on the dependent variable. Figure 3 depicts a case in which the variables x_1 and x_2 are strongly correlated. There is increasingly less variation left that can be associated to only one explanatory variable and y. In this case we need more data to precisely estimate the effect of one explanatory variable on the dependent variable. Generally, multicollinearity lets our estimates become less accurate.

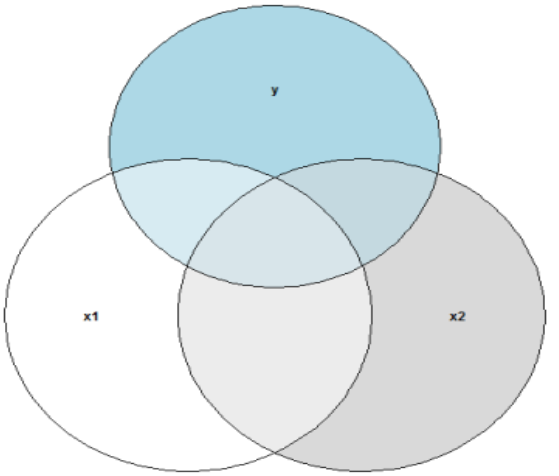


Figure 3: Strong Multicollinearity: Variable x_1 and x_2 are strongly correlated

Multicollinearity does not cause problems from a mathematical point of view as long as we do not have perfect multicollinearity. In the representation of a venn-diagram, perfect multicollinearity between variable X_1 and X_2 would mean that the circle of variable x_1 and the circle of variable X_2 are identical, i.e. there exists a perfectly overlap between the two circles. Hence, one variable is a linear combination of the other one. There is no variation left to be estimated. Variance inflation factor (VIF) is used for detecting the multicollinearity in a model, which measures the correlation and strength of correlation between the independent variables in a regression model. - If the value of VIF is less than 1: no correlation - If the value of VIF is between 1-5, there is moderate correlation - If the value of VIF is above 5: severe correlation.

The Variance Inflation Factor (VIF) measures the inflation in the coefficient of the independent variable due to the collinearities among the other independent variables. A VIF of 1 means that the regression coefficient is not inflated by the presence of the other predictors, and hence multicollinearity does not exist. **As a rule of thumb, a VIF exceeding 5 requires further investigation, whereas** VIFs above 10 indicate multicollinearity**.** Ideally, the Variance Inflation Factors are below 3.

R Codes:

```
rm(list=ls())
d=read.csv(file.choose(),header = T)
d
head(d)
tail(d)
m1=lm(d$bp~d$age)
library(lmtest)
coeftest(m1)
summary(m1)
```

First, let's perform the regression. Use lm (for linear model):

```
> fit<- lm(aa$grain_yld~aa$ear_plnt)

> summary(fit)
Call:
lm(formula = aa$grain_yld ~ aa$ear_plnt)

Residuals:
   Min     1Q  Median     3Q    Max
-13.265  -8.821   1.349   5.269  15.664

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.8212     9.8508  -0.185    0.855
aa$ear_plnt   2.0356     0.2289   8.891 6.56e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.636 on 30 degrees of freedom
Multiple R-squared:  0.7249,    Adjusted R-squared:  0.7157
F-statistic: 79.05 on 1 and 30 DF,  p-value: 6.563e-10
```

Interpretation

(f) Depended variable (Y) regress on independent variable ear head per plant (X_1)

(g) Regression equation can be written as $Y= -1.82 + 2.035\ X_1$

(h) R-squared is 0.72049 i.e. fitted regression model explain 72.049 % variation in Y, in other words we can say as 72.049 % variation in Y is explained by independent variable/s.

(i) The independent variable ear head per plant (X_1) is highly significant as p-value is less than 0.01

(j) The Fitted regression model is highly significant (F-value is 79.049 and p-value is 6.563e-10)

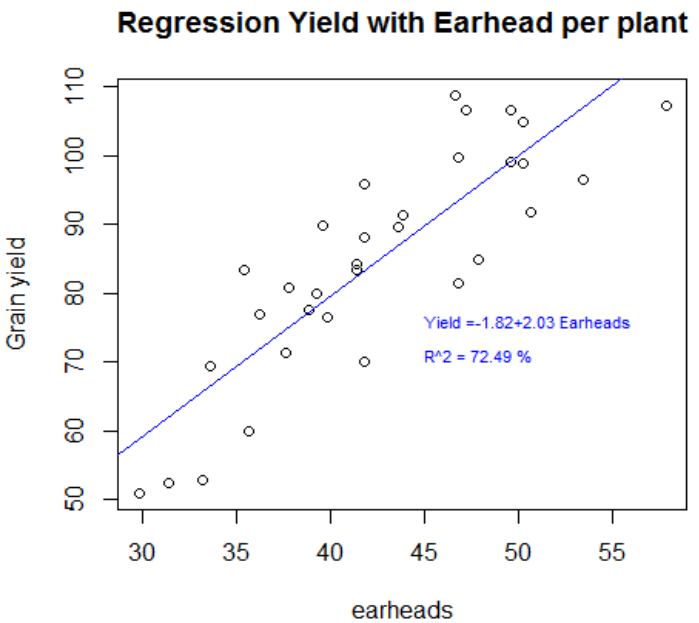
Scatterplot with regression line:

```
>plot(aa$grain_yld~aa$ear_plnt, main= "Regression Yield with Earhead per plant", xlab= "earheads",ylab="Grain yield")

>abline(fit1, col="blue")

>text(45,75, adj=c(0,0), labels="Yield =-1.82+2.03 Earheads",cex=0.7, col="blue")

>text(45,70, adj=c(0,0), labels= "R^2 = 72.49 %",cex=0.7, col="blue")
```



Multiple Regression Analysis

Use following codes for multiple regression using prd.txt data set. When regression model includes one dependent and two or more independent variables it is called as multiple regression model.

The independent variable X_1 and X_3 are shows significant effect at

```
> fit1 = lm(Y ~.,data=prd)
> summary(fit1)
Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) 325.43612   96.12721    3.385    0.00255 **
X1            0.06753   0.02899     2.329    0.02900 *
X2            2.55198   1.24824     2.044    0.05252 .
X3            3.80019   1.46114     2.601    0.01598 *
X4          -22.94947   2.70360    -8.488    1.53e-08 ***
X5            2.41748   1.80829     1.337    0.19433
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.039 on 23 degrees of freedom
Multiple R-squared:  0.8988,  Adjusted R-squared:  0.8768
F-statistic: 40.84 on 5 and 23 DF,  p-value: 1.077e-10
```

Model written as

$Y = 325.43 + 0.067\ X_1 + 2.552\ X_2 + 3.80\ X_3 - 22.949\ X_4 + 2.417\ X_5$

The 89.88 % variation in Y is explained by independent variables. The Fitted regression model is highly significant (F-value is 40.84 and p-value is 1.077e-10) and X_4 shows highly significant impact (p-value is < 0.01) at 1% level of significance
5% level of significance

Variable selection in regression

When one uses all independent variables against dependent variable some of the variables does not contribute significantly in model. The variables which contributes to regression model can be trained and others will be excluded based on different criteria such as p-value, AIC, Mallows Cp and RMSE.

The olsrr package in R does this job efficiently.

```
>library(olsrr)
> k=ols_step_backward_p(fit1,prem = 0.05)
```

Here backward elimination method is used in which 1st model will have all independent variables

Then based on p value <0.05 , the respective variable will be retained.

> k

Elimination Summary

Step	Variable Removed	Adj. R-Square	R-square	C(p)	AIC	RMSE
1	X5	0.8909	0.8727	5.7873	210.6363	8.1698
2	X1	0.8741	0.859	7.5963	212.7817	8.5978
3	X2	0.8587	0.8478	9.0975	214.1313	8.9321

(k) At 1st step X5 deleted as it has highest p-value in full model

(l) Removing X5 and model fitted variable with highest p-value will be deleted and so on

(m) At last step all non-significant (p>0.05) variable will be deleted and information of last model will be obtained as

```
>summary(k$model)
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept) 483.6703    39.5671  12.224 2.78e-12 ***
X3            4.7963     0.9511   5.043 3.00e-05 ***
X4          -24.2150     1.9405 -12.479 1.75e-12 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.932 on 26 degrees of freedom
Multiple R-squared:  0.8587,  Adjusted R-squared:  0.8478
F-statistic: 79.01 on 2 and 26 DF,  p-value: 8.938e-12
```

Here we can observe that model is significant and all the variables also significant. In built function such as step(fit1, direction = "backward", test = "F") can also be used.

References

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326): John Wiley & Sons.

Gupta, S., & Kapoor, V. (2020). *Fundamentals of mathematical statistics*: Sultan Chand & Sons.

Harrell, F. E. (2020). with contributions from Charles Dupont and many others (2019) Hmisc: Harrell Miscellaneous. R package version 4.2-0. In.

Loftus, S. C. (2021). *Basic statistics with R: reaching decisions with data*: Academic Press.

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences, 106*, 234-240.

How to Cite: Nirmala, B., Patil, S., and Rathod, S. (2023). Correlation and Regression Analysis. In S. Rathod, B. Sailaja, N. Bandumula, S. Arun Kumar, P. A. Lakshmi Prasanna, P. Jeyakumar, A. Waris, P. Muthuraman, and R. M. Sundaram (Eds.), *Statistical Procedures for Analysing Agricultural Data using R* (pp. 69-85). ICAR - Indian Institute of Rice Research, Hyderabad.