Lab Session 5

Chapter 4: Scatter plot and Correlation

Scatter plot

A Scatterplot displays the relationship between two quantitative variables. Each dot represents an observation. Their position on the X (horizontal) and Y (vertical) axis represents the values of the 2 variables. There are many ways to create a scatterplot in R. The basic function is plot(x, y), where x and y are numeric vectors denoting the (x,y) points to plot.

```
# Simple Scatterplot
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",  xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

The **scatterplot( )** function in the car package offers many enhanced features, including fit lines, marginal box plots, conditioning on a factor, and interactive point identification. Each of these features is optional.

```
# Enhanced Scatterplot of MPG vs. Weight  by Number of Car Cylinders
library(car)
scatterplot(mpg ~ wt | cyl, data=mtcars,xlab="Weight of Car", ylab="Miles Per Gallon",main="Enhanced Scatter Plot")
```

There are R functions available to create scatterplot matrices.

```
# Basic Scatterplot Matrix
pairs(~mpg+disp+drat+wt,data=mtcars,  main="Simple Scatterplot Matrix")
```

The lattice package provides options to condition the scatterplot matrix on a factor.

```
# Basic Scatterplot Matrix
pairs(~mpg+disp+drat+wt,data=mtcars,  main="Simple Scatterplot Matrix")
```

correlation

You can use the **cor( )** function to produce correlations between two quantitative variables.A simplified format is **cor(x, use=, method=)** where

```
# Correlations among numeric variables in  data frame mtcars. Use listwise deletion of missing data.
cor(mtcars, use="complete.obs", method="pearson")
```

Chapter 5:Regression

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables response and explanatory variables. In Linear Regression, these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph.

The general mathematical equation for linear regression is −y = ax + b

Following is the description of the parameters used −

- **y** is the response variable.
- **x** is the explanatory variable.
- **a** and **b** are constants which are called the coefficients (intercept and slope)

Steps to Establish a Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.

The steps to create the relationship is

- Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R.

Input Data

Below is the sample data representing the observations −

```
# Values of height
151, 174, 138, 186, 128, 136, 179, 163, 152, 131

# Values of weight.
63, 81, 56, 91, 47, 57, 76, 72, 62, 48
```

lm() Function

This function creates the relationship model between the predictor and the response variable.The basic syntax for **lm()** function in linear regression is lm(formula,data)

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between x and y.
- **data** is the vector on which the formula will be applied.\

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)# Plot the histogram for this sample.
# Apply the lm() function.
relation <- lm(y~x)
print(relation)

#to get more details of the model
summary(relation)
```

predict() Function

The basic syntax for predict() in linear regression is predict(object, newdata)

Following is the description of the parameters used

- **object** is the formula which is already created using the lm() function.
- **newdata** is the vector containing the new value for predictor variable.

```
# The predictor vector.
a <- data.frame(x = 170)
result <-  predict(relation,a)
print(result)
```

Visualize the Regression Graphically

```
plot(y,x,col = "blue",main = "Height & Weight Regression", abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "Weight in Kg",ylab = "Height in cm")
```

Residuals

The basic syntax to get residuals in linear regression is resid(object,..) or residuals(object, ...).

```
res <-resid(relation)
plot(x, res,ylab="Residuals", xlab="weight")
abline(0, 0)
```

Lab Exercise

1. Load the set of data “airquality” in the package “datasets” in R.
   a) Compute summary measurements for the quantitative variables and interpret your results. (Note: Avoid the effect of missing values if they exist)
   b) Identify the shape of the distributions of the quantitative variables using a suitable graphical method and comment on your findings.
   c) According to the results obtained in part b, name a suitable variable that can be approximated to normal distribution. Draw a normal curve for that variable.

2. Load the data set “iris” in the package “datasets”. This data set gives the measurements in centimeters of the variables sepal length and width and petal length and width and the species for 50 flowers from each of 3 species of iris.

   Construct a matrix scatter plot for the data and interpret the relationships among variables.

3. Load the data set “cats” in the package “MASS”. This data set contains Body weight (Bwt) and Heart weight (Hwt) of 144 domestic cats.
   a) Check whether there is any relationship between Bwt and Hwt, by using a suitable graphical method. Comment on your plot.
   b) Fit a simple linear regression to model the Heart weight of the cats using Body weight as the independent variable. Write the equation of the fitted model.
   c) Test the goodness of fit of the fitted model and justify your answer (R squared).
   d) Test the significance of the relationship between the variables Bwt and Hwt. (Mention all the steps clearly).
   e) Predict the heart weight of a cat if its body weight is 2.52 kg.