AssignmentLecture9

Zepeng Mu ID: 2014K8009915010

3/23/2017

Introduction

Simple linear regression is used to see whether the relationship between the independent value and dependent value satifies the formula $y = a * x + b$. During a simple linear regression analysis, the value of a and b are calculated, and the reliability of the result is also evaluated. Here I used a dataset generated by myself to perform the simple linear regression.

Assumptions of Simple Linear Regression

1. Normal distribution of Y value for X
2. Homogeneity of variance
3. The actual relationship is linear
4. Values of Y are independent to each other
5. X has no error

Dataset

The data set is generated in the following way: To begin with, X values is generated. Since I did not want the X values to be evenly distributed, I used the `sample()` function.

```
x = sample(seq(15, 30, by = 0.2), 60, rep = TRUE); x
```

```
##  [1] 20.6 19.4 20.8 23.0 28.0 28.0 23.2 29.8 15.0 18.0 28.0 23.0 28.4 21.2
## [15] 20.8 15.4 24.2 19.4 25.8 17.4 29.6 21.0 23.2 29.0 23.0 24.2 16.6 19.6
## [29] 19.2 20.6 24.2 26.6 22.2 19.0 18.4 18.8 19.0 25.0 28.8 24.4 17.0 29.2
## [43] 26.6 23.0 17.8 22.4 22.6 28.6 17.8 25.0 28.0 16.0 17.6 28.8 24.0 19.2
## [57] 16.0 29.8 19.4 24.4
```

However, identical X values may be generated, so I only chose the unique values:

```
xValue = unique(x); xValue
```

```
##  [1] 20.6 19.4 20.8 23.0 28.0 23.2 29.8 15.0 18.0 28.4 21.2 15.4 24.2 25.8
## [15] 17.4 29.6 21.0 29.0 16.6 19.6 19.2 26.6 22.2 19.0 18.4 18.8 25.0 28.8
## [29] 24.4 17.0 29.2 17.8 22.4 22.6 28.6 16.0 17.6 24.0
```

Here, I could assume that the value of X has no error. The next step is to get Y values. I first got exact Y values using a linear function:

```
y = 3 * xValue + 7; y
```

```
##  [1] 68.8 65.2 69.4 76.0 91.0 76.6 96.4 52.0 61.0 92.2 70.6 53.2 79.6 84.4
## [15] 59.2 95.8 70.0 94.0 56.8 65.8 64.6 86.8 73.6 64.0 62.2 63.4 82.0 93.4
## [29] 80.2 58.0 94.6 60.4 74.2 74.8 92.8 55.0 59.8 79.0
```

Note that by doing so I actually satisfied the third assumption mentioned above, that is, the actual relationship is linear. Then, I need to make some 'errors' and to get the values of Y that are not perfect linear to X:

```
yErr = sample(seq(-5, 5, by = 0.2), length(y), rep = TRUE); yErr
```

```
##  [1] -1.0 -1.4 -4.2 -1.4  0.0  0.6 -2.8  2.2  2.6  1.8  2.0 -3.0 -2.6 -1.4
## [15] -0.4  4.2  1.8  0.2 -1.0  4.2  3.6 -0.8 -0.4  5.0 -1.6  2.0 -1.6 -1.4
## [29] -4.8 -4.4  5.0  4.6  3.0 -4.0  2.4  4.2  3.2  2.2
```

```
yValue = y + yErr; yValue
```

```
##  [1]  67.8  63.8  65.2  74.6  91.0  77.2  93.6  54.2  63.6  94.0  72.6
## [12]  50.2  77.0  83.0  58.8 100.0  71.8  94.2  55.8  70.0  68.2  86.0
## [23]  73.2  69.0  60.6  65.4  80.4  92.0  75.4  53.6  99.6  65.0  77.2
## [34]  70.8  95.2  59.2  63.0  81.2
```

By doing so, I want make sure that the Y values are mutually independent. Next, a dataframe is generated.

```
regData = data.frame(yValue, xValue); regData
```

```
##    yValue xValue
## 1    67.8   20.6
## 2    63.8   19.4
## 3    65.2   20.8
## 4    74.6   23.0
## 5    91.0   28.0
## 6    77.2   23.2
## 7    93.6   29.8
## 8    54.2   15.0
## 9    63.6   18.0
## 10   94.0   28.4
## 11   72.6   21.2
## 12   50.2   15.4
## 13   77.0   24.2
## 14   83.0   25.8
## 15   58.8   17.4
## 16  100.0   29.6
## 17   71.8   21.0
## 18   94.2   29.0
## 19   55.8   16.6
## 20   70.0   19.6
## 21   68.2   19.2
## 22   86.0   26.6
## 23   73.2   22.2
## 24   69.0   19.0
## 25   60.6   18.4
## 26   65.4   18.8
## 27   80.4   25.0
## 28   92.0   28.8
## 29   75.4   24.4
## 30   53.6   17.0
## 31   99.6   29.2
## 32   65.0   17.8
## 33   77.2   22.4
## 34   70.8   22.6
## 35   95.2   28.6
## 36   59.2   16.0
## 37   63.0   17.6
## 38   81.2   24.0
```

Results

Up to now, I can perform the simple linear regression analysis.

```
reg = lm(yValue~xValue, data = regData); reg
```

```
##
## Call:
## lm(formula = yValue ~ xValue, data = regData)
##
## Coefficients:
## (Intercept)       xValue
##       8.343        2.959
```

```
s = summary(reg); s
```

```
##
## Call:
## lm(formula = yValue ~ xValue, data = regData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1470 -1.8930 -0.0797  2.1662  4.8489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3431     2.3709   3.519  0.00119 **
## xValue        2.9592     0.1047  28.266  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.889 on 36 degrees of freedom
## Multiple R-squared:  0.9569, Adjusted R-squared:  0.9557
## F-statistic:   799 on 1 and 36 DF,  p-value: < 2.2e-16
```

A plot with dots and regression is shown:

```
plot(xValue, yValue)
abline(reg)
```

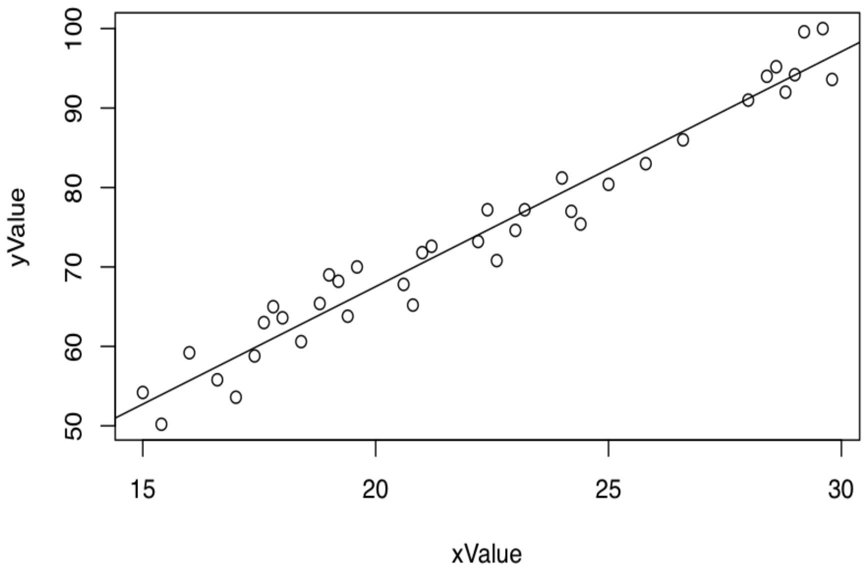

Figure 1 Simple linear regression beteen X and Y

After simple linear regression analysis, I still need to test the assumptions. The homogeneity of variance and the normal distribution of Y values are checked:

```
plot(reg$fitted, reg$resid)
```

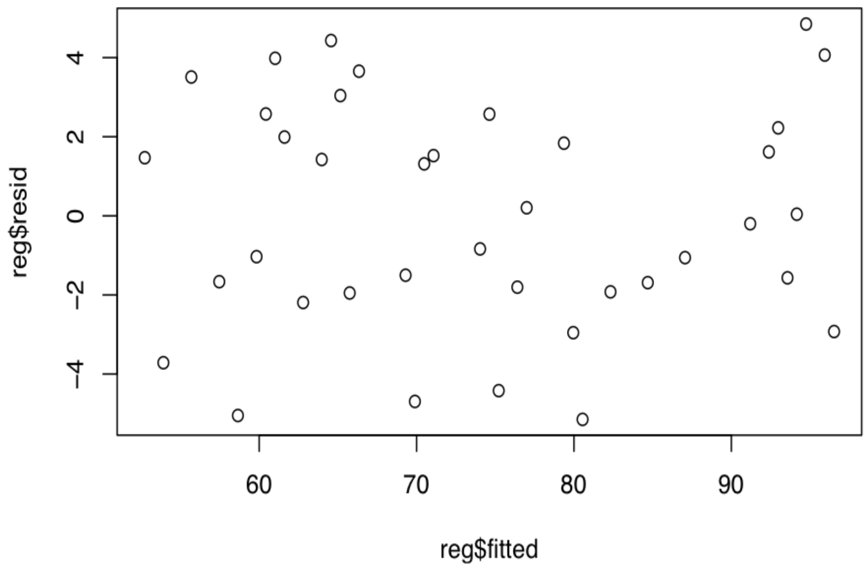

Figure 2 Checking the homogeneity of variance

```
n = shapiro.test(reg$resid); n
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  reg$resid
## W = 0.96024, p-value = 0.1933
```

Finally, I looked at R squred to evaluate the general realiability of the analysis.

```
rs = s$r.squared; rs
```

```
## [1] 0.9568851
```