Multiple Regression

Divit Vasu

21 September 2018

Multiple Linear Regression

The multiple linear regression model is used to predict response (independent) variable based on two or more predictor variable (dependent) variable.

The multiple linear regression model can be stated as follows

$Y_i = \beta_0 + \beta_1 * X_i 1 + \beta_2 X_i 2 + \ldots + \beta_p x_i p + \epsilon_i$ where

$Y_i$is the i^{th} value of the response variable, $X_{ij}$is the i^{th} observtaion of the j^{th} predictor variable, $\beta_0, \beta_1, \ldots, \beta_p$are the parameters (regression coefficients), ϵ_i random error term with E(ϵ_i) = 0 and V(ϵ_i) = σ^2,ϵ_i~IN(0,σ^2)

Definition

- Load the data trees from datasets package.
- Display the structure of trees data set.
- Check the dimension of trees data set.
- Obtain the summary statistics for trees data set.
- Fit a simple linear regression model with Volume as dependent and Girth as independent variable.
- Fit a simple linear regression model with Volume as dependent and Height as independent variable.
- Fit a multiple linear regression model with Volume as dependent and Height and Girth as independent variable.

- Load the data set

```
data("trees")
```

- Display structure of data set

```
str(trees)
```

```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

- Display dimension of trees data set

```
dim(trees)
```

```
## [1] 31  3
```

- Display Summary of data set

```
summary(trees)
```

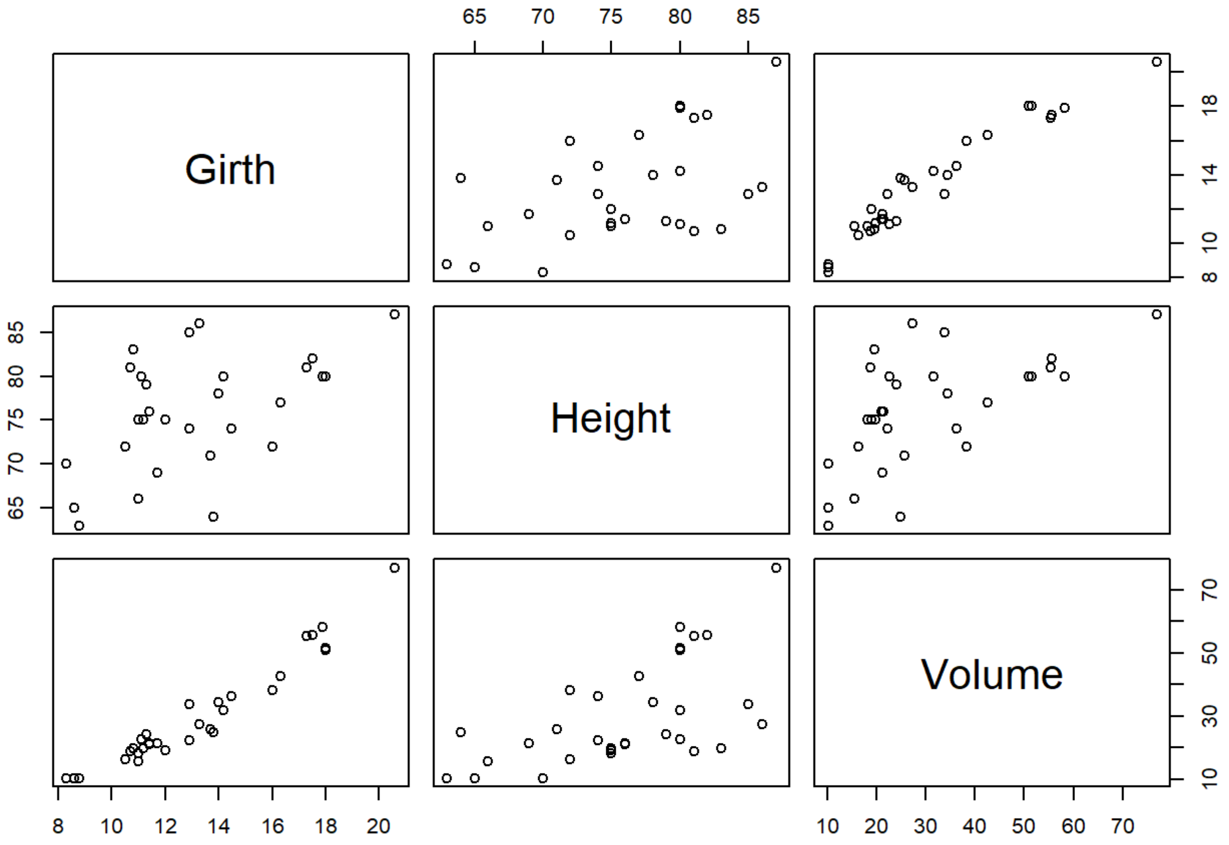
```
##      Girth           Height       Volume    
##  Min.   : 8.30   Min.   :63   Min.   :10.20  
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40  
##  Median :12.90   Median :76   Median :24.20  
##  Mean   :13.25   Mean   :76   Mean   :30.17  
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30  
##  Max.   :20.60   Max.   :87   Max.   :77.00  
```

- Correlation coefficient matrix

```
cor(trees)
```

```
##            Girth    Height    Volume
## Girth  1.0000000 0.5192801 0.9671194
## Height 0.5192801 1.0000000 0.5982497
## Volume 0.9671194 0.5982497 1.0000000
```

```
plot(trees) # Scatter plot
```



- Fit linear model Volume on Girth

```
model1<-lm(Volume~Girth,data = trees)
model1
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Coefficients:
## (Intercept)        Girth
##     -36.943        5.066
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

For both the coefficient (intercept) and (slope) the p-values are less than 0.05, we concldue that the regression coefficients are significant at 0.05 level of significance.

The multiple R^2 is 0.9353. That is 93.53 percent of the variation in Volume of tree is explained by the Girth.

- Fit linear model Volume on Height

```
model2<-lm(Volume~Height,data = trees)
model2
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Coefficients:
## (Intercept)       Height
##     -87.124        1.543
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

For both the coefficient (intercept) and (slope) the p-values are less than 0.05, we concldue that the regression coefficients are significant at 0.05 level of significance. The multiple R^2 is 0.3579. That is 35.79 percent of the variation in Volume of tree is explained by the Height. The coefficient of determination for model 1 is R^2 is 0.9353 and for model 2 is R^2 is 0.3579. Girth is preferable to predict the volume of timber using simple linear regression model, as the amount of variation explained by girth about the volume is more compared to the amount of variation explained by height.

- Fit linear model Volume on Height and Girth

```
model3<-lm(Volume~Height+Girth,data = trees)
model3
```

```
##
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
##
## Coefficients:
## (Intercept)       Height        Girth
##    -57.9877       0.3393       4.7082
```

```
summary(model3)
```

```
## 
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.4065 -2.6493 -0.2876  2.2003  8.4847 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Height        0.3393     0.1302   2.607   0.0145 *  
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442 
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Conclusion

For all the coefficient the p-values are less than 0.05, we concldue that the regression coefficients are significant at 0.05 level of significance.

The multiple R^2 is 0.948. That is 94.8 percent of the variation in Volume of tree is explained by the Girth and Height.