Simple Linear Regression

Kingsuk_Jana

2023-04-07

Create data frame

```
Height_Wight_data<- data.frame(
  Height = c(72, 68, 69,68,64,72,72,66,68,70,74,74,63,66.25,62,67.5,70,68.5,66,73,64,63,66,71,68,69,65,6
  Weight = c(200, 165,160,135,120,162,190,139,155,155,185,170,137,125,114,130,196,160,135,155,125,125,2(
)
dim(Height_Wight_data)
```

```
## [1] 28  2
```

1. Response variable(Y) is weight.

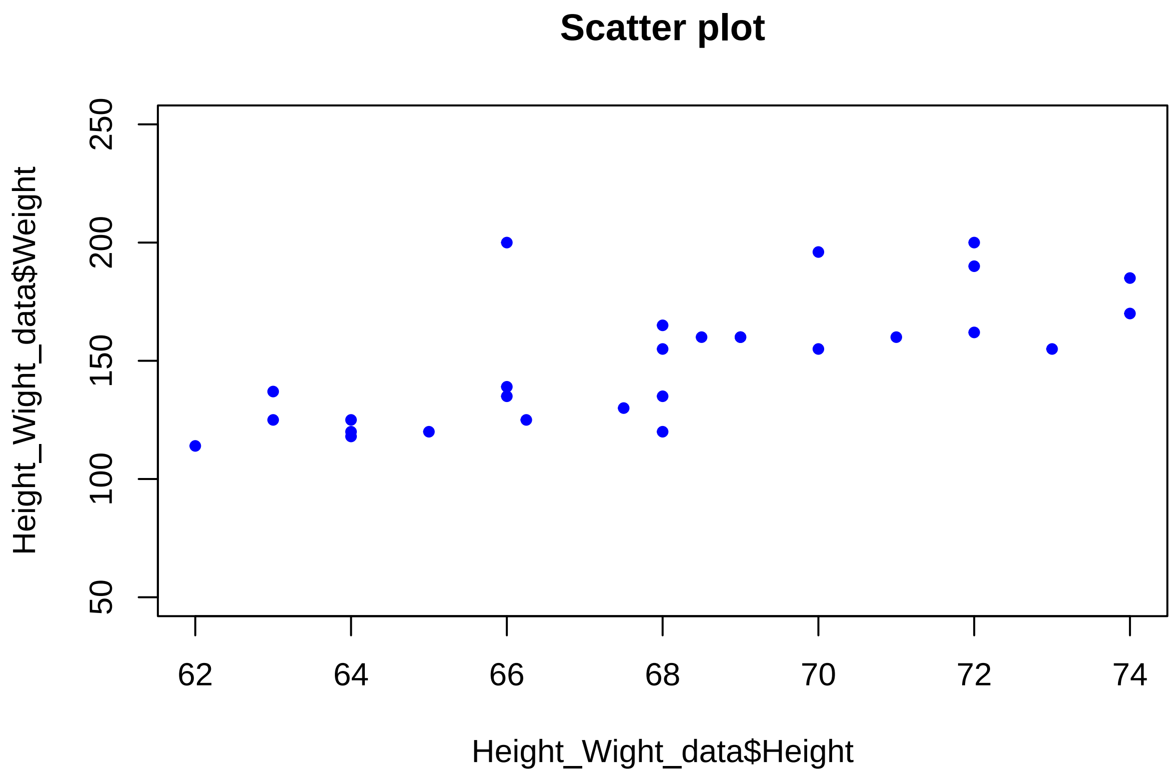
2. Predictor variable(X) is height.

3. In simple linear regression, we model the relationship between one predictor variable and response variable through the linear mathematical equation y = Beta_0 + Beta_1*x.

4. Here 'Beta_0' is called 'y-intercept' Where the linear regression line intercept with y axis. Whereas 'Beta_1' is called 'slope' and it quantify how this two variables are related.

5. Scatter plot: That describe the relationship between response and predictor variable.

```
plot(Height_Wight_data$Height, Height_Wight_data$Weight, main = "Scatter plot",
     pch=20, col="blue", ylim = c(50,250))
```



6. Decide from the above scatter plot whether the response and predictor variable are positively, negatively, or no correlation.

```
cor(Height_Wight_data$Height, Height_Wight_data$Weight)
```

```
## [1] 0.7111321
```

7. Write down the simple linear regression model(SLR)

Y = Beta_0 + Beta_1*x + error. Beta_0 = Population y-intercept. Beta_1 = Population slope and. Error is the deviation of Y from Beta_0+Beta_1*x .

8. Assumptions of Simple Linear Regression(SLR)

Linearity: The relationship between response(Y) and predictor(X) and it must be linear. Check this assumption by examining a scatterplot of X and Y.

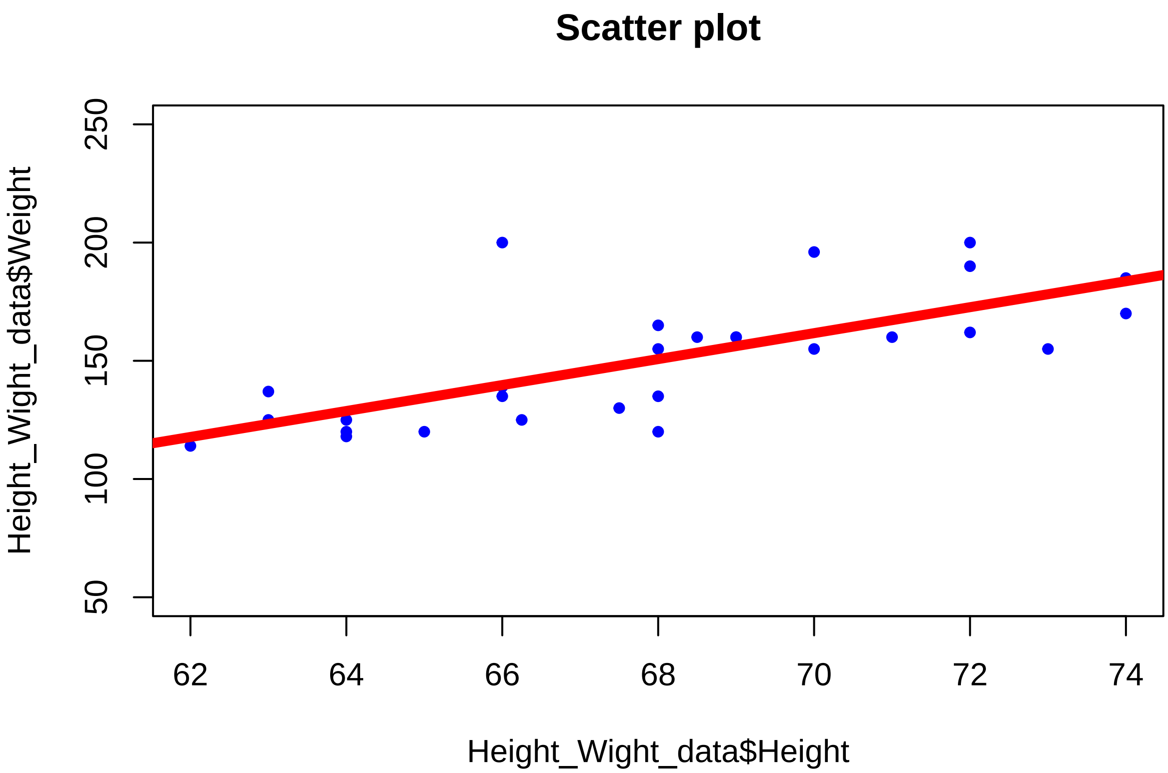
Independence of errors: There is not a relationship between the Y variable and the residuals.

9.Use least square method to estimate the coefficients of SLR model.

```r
linear_reg = lm(Weight ~ Height, data = Height_Wight_data)
summary(linear_reg)
```

```
##
## Call:
## lm(formula = Weight ~ Height, data = Height_Wight_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.718 -11.486  -3.777   4.846  60.258
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -222.479     72.421  -3.072  0.00494 **
## Height         5.488      1.064   5.158 2.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.11 on 26 degrees of freedom
## Multiple R-squared:  0.5057, Adjusted R-squared:  0.4867
## F-statistic:  26.6 on 1 and 26 DF,  p-value: 2.218e-05
```

```r
plot(Height_Wight_data$Height,Height_Wight_data$Weight,main = "Scatter plot",
     pch=20, col="blue", ylim = c(50,250))
#here the estimates are
#Beta_0_hat= -222.479
#Beta_1_hat= 5.488.
abline(lm(Weight ~ Height, data = Height_Wight_data), col="red", lwd=5)
```



The fitted regression is wight = -222.5 + 5.49*height

Interpretation of slope: As we can see from the above result, the estimated value of slope(Beta_1) is 5.488. It represent as one inch increase in height the estimated increment in weight is 5.488(pound).

Interpretation of coefficient: The intercept(Beta_o) is -222.5. That means estimated weight of a person -222 pounds when someone have height 0. It is not possible to get a person whose height 0.

#10.confidence Interval for population slope:

The confidence interval for population slope is (Estimated_Beta_1 - SE(Estimated_Beta_1)*t(alpha/2), Estimated_Beta_1 + SE(Estimated_Beta_1)*t(alpha/2))

Here the estimated_Beta_1 is 5.49 and the standard error of the estimated_Beta_1 is 1.064. The t(alpha/2) value represent the tabulated value of t distribution with 28-2=26 degrees of freedom. We find the t value to be 2.056.

Putting all values together we can get the confidence interval as: (5.49 - 2.056*1.064, 5.49 + 2.056*1064) =(3.31, 7.67). We are 95% sure that population slope is in between 3.31 and 7.67. In other words, we are 95% sure that, as height increase by one inch, that the weight is increase by between 3.32 and 7.67 pounds, on average.

#11. Coefficient of Ditermination(R-square):

Coefficient of determination measures the percentage of variability with-in the response variable can be explained by the regression model. Therefore, the value of r-square close to 100% means the model is useful and the value close to zero indicate the model is not useful to explain the variability in the response variable.

From the output we can get the value of R-square as 0.5057 or 50.57%. This value means, 50.57% variability in weight can be explained by the height.

#12 Correlation test

```r
cor.test(Height_Wight_data$Height, Height_Wight_data$Weight, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Height_Wight_data$Height and Height_Wight_data$Weight
## t = 5.1576, df = 26, p-value = 2.218e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4601312 0.8568743
## sample estimates:
##       cor
## 0.7111321
```

#t- is the t-test statistic value (t = 5.16), #df- is the degrees of freedom (df= 28-2), #p-value is the significance level of the t-test (p-value = 0.000002). #conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [0.46, 0.85]); #sample estimates is the correlation coefficient (Cor.coeff = -0.87).'