Regression Analysis in R

Yegon Shadrack

2024-09-04

## 1.Simple linear regression

### Load the data

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)

### Load the data

The data comes from car dataset showing the car speed and the stopping distance(dist)

cars %>% head(5)

## speed dist  
## 1 4 2  
## 2 4 10  
## 3 7 4  
## 4 7 22  
## 5 8 16

### Create the Model

cars %>% lm(dist~speed,data=.) %>% summary()

##   
## Call:  
## lm(formula = dist ~ speed, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.069 -9.525 -2.272 9.215 43.201   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.5791 6.7584 -2.601 0.0123 \*   
## speed 3.9324 0.4155 9.464 1.49e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.38 on 48 degrees of freedom  
## Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438   
## F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

### Findings

1. P-value associated with F-statistics is very small showing that at least one of the predictor variable is significant predictor of Outcome variable.
2. R-squared=0.65,Indicates that 65% of the change in stopping distance can be explained by explained by the change in speed.

### conclusion

model: dist=-17.58+3.93(speed) Increasing the speed of the car by 1 unit increases the car stopping distance by 3.93 ft

## 2.Multiple linear regression

### Load the data

The data is from Trees dataset and it shows the Diameter(girth), Height and Volume for Black Cherry Trees. I wanted to investigate whether there is any relationship between Volume,diameter(girth) and height.

trees %>% head(5)

## Girth Height Volume  
## 1 8.3 70 10.3  
## 2 8.6 65 10.3  
## 3 8.8 63 10.2  
## 4 10.5 72 16.4  
## 5 10.7 81 18.8

### Create the Model

trees %>% lm(Volume~Girth+Height,data=.) %>% summary()

##   
## Call:  
## lm(formula = Volume ~ Girth + Height, data = .)  
##   
## 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 \*\*\*  
## Girth 4.7082 0.2643 17.816 < 2e-16 \*\*\*  
## Height 0.3393 0.1302 2.607 0.0145 \*   
## ---  
## 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

### Findings

1. P-value associated with F-statistics is less than 0.5,showing that at least one of the predictor variable is significant.
2. Individual p-values of the coefficients are all significant indicating that height and girth is significant predictor of volume.
3. Adjusted R-squared=0.94,showing that 94% of the change in volume can be explained by both the change in girth and height.

### Conclusions

model: Volume=-57.99+4.71(Girth)+0.34(Height)

* keeping other factors constant,1 inch increase in Girth results in 4.71 cubic ft increase in volume
* Also ,keeping girth constant,a one unit increase in height ,increases the volume by 0.33 cubic ft