Simple Linear Regression

2025-05-14

Table of Contents

# 1 Load the Dataset

pacman::p\_load("readr")  
  
clv\_data <- read\_csv("./data/clv\_data.csv")  
head(clv\_data)

## # A tibble: 6 × 2  
## purchase\_frequency customer\_lifetime\_value  
## <dbl> <dbl>  
## 1 3 110.   
## 2 7 190.   
## 3 6 160.   
## 4 2 94.4  
## 5 4 133.   
## 6 8 223.

# 2 Initial EDA

dim(clv\_data)

## [1] 500 2

sapply(clv\_data, class)

## purchase\_frequency customer\_lifetime\_value   
## "numeric" "numeric"

str(clv\_data)

## spc\_tbl\_ [500 × 2] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ purchase\_frequency : num [1:500] 3 7 6 2 4 8 0 4 8 3 ...  
## $ customer\_lifetime\_value: num [1:500] 110.3 190.2 160 94.4 133.2 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. purchase\_frequency = col\_double(),  
## .. customer\_lifetime\_value = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

**Descriptive Statistics**

## 2.1 Measures of Frequency

## 2.2 Measures of Central Tendency

summary(clv\_data)

## purchase\_frequency customer\_lifetime\_value  
## Min. :-1.000 Min. : 26.13   
## 1st Qu.: 4.000 1st Qu.:122.04   
## Median : 5.000 Median :148.21   
## Mean : 4.914 Mean :148.25   
## 3rd Qu.: 6.000 3rd Qu.:175.88   
## Max. :11.000 Max. :262.04

## 2.3 Measures of Distribution

sapply(clv\_data[,], var)

## purchase\_frequency customer\_lifetime\_value   
## 4.146898 1642.315996

sapply(clv\_data[,], sd)

## purchase\_frequency customer\_lifetime\_value   
## 2.036393 40.525498

pacman::p\_load("e1071")  
sapply(clv\_data[,], kurtosis, type = 2)

## purchase\_frequency customer\_lifetime\_value   
## -0.1220038 -0.1484811

sapply(clv\_data[,], skewness, type = 2)

## purchase\_frequency customer\_lifetime\_value   
## -0.04021915 -0.01608242

cov(clv\_data, method = "spearman")

## purchase\_frequency customer\_lifetime\_value  
## purchase\_frequency 20409.91 20235.73  
## customer\_lifetime\_value 20235.73 20874.99

cor.test(clv\_data$customer\_lifetime\_value, clv\_data$purchase\_frequency, method = "spearman")

##   
## Spearman's rank correlation rho  
##   
## data: clv\_data$customer\_lifetime\_value and clv\_data$purchase\_frequency  
## S = 409190, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.9803588

cor(clv\_data, method = "spearman")

## purchase\_frequency customer\_lifetime\_value  
## purchase\_frequency 1.0000000 0.9803588  
## customer\_lifetime\_value 0.9803588 1.0000000