STAT202 Assignment 1

David Ewing

2024-07-31

# 0. conditionally install (synchronise between home and uni laptops)  
  
if (!requireNamespace("tidyverse", quietly = TRUE)) {  
 install.packages("tidyverse")  
}  
if (!requireNamespace("conflicted", quietly = TRUE)) {  
 install.packages("conflicted")  
}

# 0.1 Load lib (and address conflicts)

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(conflicted)   
conflicted::conflict\_prefer("filter", "dplyr")

[conflicted] Will prefer dplyr::filter over any other package.

conflicted::conflict\_prefer("lag", "dplyr")

[conflicted] Will prefer dplyr::lag over any other package.

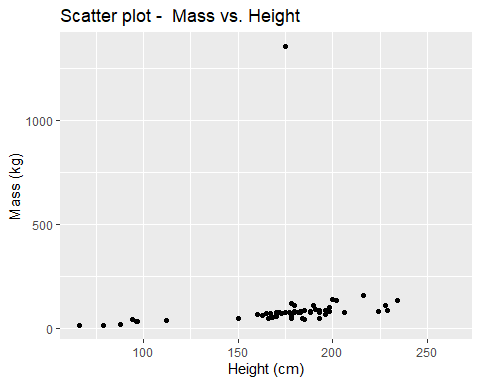
## 1. Load dataset

data(starwars)  
df <- starwars  
View(starwars)

## 2. Create scatter plot

ggplot(df, aes(x = height, y = mass)) +  
 geom\_point() +  
 labs(title = "Scatter plot - Mass vs. Height",   
 x = "Height (cm)", y = "Mass (kg)")

Warning: Removed 28 rows containing missing values or values outside the scale range  
(`geom\_point()`).



# Assessment: WEAK and positive . if the outlyer was removed then I would suggest stronger and positive

## 3. missing data

na\_rows <- starwars %>% filter(if\_any(everything(), is.na))  
na\_graph <- starwars %>% filter(is.na(mass) | is.na(height))  
starwars |> select (mass,height) |> summary()

mass height   
 Min. : 15.00 Min. : 66.0   
 1st Qu.: 55.60 1st Qu.:167.0   
 Median : 79.00 Median :180.0   
 Mean : 97.31 Mean :174.6   
 3rd Qu.: 84.50 3rd Qu.:191.0   
 Max. :1358.00 Max. :264.0   
 NA's :28 NA's :6

## 4. drop missing data rows

star\_no <- starwars |> drop\_na(mass, height) #drop na  
star\_no <- star\_no |> filter (mass <= 250)  
star\_no |> select (mass,height) |> summary()

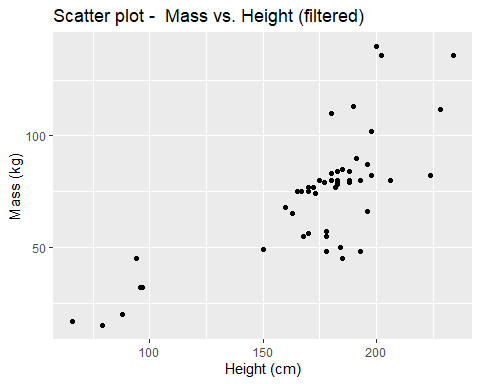
mass height   
 Min. : 15.00 Min. : 66.0   
 1st Qu.: 55.30 1st Qu.:170.0   
 Median : 79.00 Median :181.0   
 Mean : 75.58 Mean :174.3   
 3rd Qu.: 84.00 3rd Qu.:192.5   
 Max. :159.00 Max. :234.0

## 5. Student\_ID as value

Student\_ID = 82171165  
set.seed(Student\_ID)  
my\_starno <-star\_no |> sample\_n(50)   
#my\_starno |> summary()

## 6. scatter plot of filtered data

ggplot(my\_starno, aes(x = height, y = mass)) +  
 geom\_point() +  
 labs(title = "Scatter plot - Mass vs. Height (filtered)",   
 x = "Height (cm)", y = "Mass (kg)")



# Assessment: MODERATE and positive

## 7. linear regression model

model\_1 <- lm(mass ~ height, data = my\_starno)  
print(summary(model\_1))

Call:  
lm(formula = mass ~ height, data = my\_starno)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-37.201 -5.908 1.214 5.379 50.679   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -28.39482 12.41314 -2.287 0.0266 \*   
height 0.58858 0.07023 8.381 5.89e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 17.76 on 48 degrees of freedom  
Multiple R-squared: 0.594, Adjusted R-squared: 0.5856   
F-statistic: 70.24 on 1 and 48 DF, p-value: 5.885e-11

coefficients <- summary(model\_1)$coefficients

# The intercept is where the height is zero and is useless on its own but is required to maintain the slope of the line to properly fit a curve relative to the observed data points.

# The slope shows how the mass of a character is expected to increase for every centimetre of highght increase as per the group of observed data points.

````

# 8. centre the height

mystar\_no <- my\_starno %>% mutate(cent\_height = height - mean(height, na.rm = TRUE))  
model\_2 <- lm(mass ~ cent\_height, data = mystar\_no)  
summary(model\_2)

Call:  
lm(formula = mass ~ cent\_height, data = mystar\_no)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-37.201 -5.908 1.214 5.379 50.679   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 73.48800 2.51138 29.262 < 2e-16 \*\*\*  
cent\_height 0.58858 0.07023 8.381 5.89e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 17.76 on 48 degrees of freedom  
Multiple R-squared: 0.594, Adjusted R-squared: 0.5856   
F-statistic: 70.24 on 1 and 48 DF, p-value: 5.885e-11

coefficients\_2 <- summary(model\_2)$coefficients

# Differences:

# Primary difference is the intercept.

# r, r-squared and adjusted r-square # remains the same. Indicating to me that the slope is the same and that the data shifted and this shift is primarily indicated in the intercept.

# 9. scatter plot of mystar\_no data

ggplot(mystar\_no, aes(x = height, y = mass)) +  
 geom\_point() +  
 labs(title = "Scatter plot - Mass vs. Height (mystar\_no)",   
 x = "Height (cm)", y = "Mass (kg)")

