STAT202 Lab 2: Multiple Linear Regression START

David Ewing

Due on 7 August 2024

# Introduction

This document contains the analysis for Lab 2. It explores biometric data using multiple linear regression to understand the relationships between various predictors and the response variable weight\_kg. The analysis will follow these steps:”

# Step 0: setup

loading libraries:

set.seed(82171165) #set seed   
  
knitr::opts\_chunk$set(  
 echo = TRUE, # Show all code by default  
 message = TRUE, # Include package messages  
 warning = TRUE # Include warnings if they occur  
)  
  
  
  
library(conflicted)  
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

library(readxl)

## Warning: package 'readxl' was built under R version 4.4.2

library(performance)

## Warning: package 'performance' was built under R version 4.4.2

library(GGally)

## Warning: package 'GGally' was built under R version 4.4.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(flextable)

## Warning: package 'flextable' was built under R version 4.4.2

library(skimr)

## Warning: package 'skimr' was built under R version 4.4.2

library(data.table)  
conflict\_prefer("filter", "dplyr"); conflict\_prefer("select", "dplyr")

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

# Step 1: Read biometrics.xlsx

biomet <- read\_excel("../data/biometrics.xlsx")  
skim\_summary <- skim(biomet)  
skim\_summary |>  
 select(skim\_variable, n\_missing)

## # A tibble: 6 × 2  
## skim\_variable n\_missing  
## <chr> <int>  
## 1 sex 0  
## 2 weight\_kg 0  
## 3 height\_cm 0  
## 4 handspan\_cm 0  
## 5 finger2\_cm 1  
## 6 finger4\_cm 1

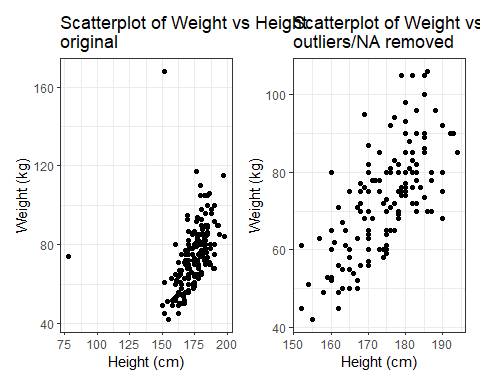
# Step 2: Scatterplot of Weight vs Height

befor <- biomet |>  
 ggplot(aes(x = height\_cm, y = weight\_kg)) + # Scatterplot  
 geom\_point() +  
 labs(  
 title = "Scatterplot of Weight vs Height \noriginal",  
 x = "Height (cm)",  
 y = "Weight (kg)"  
 ) +  
 theme\_bw()

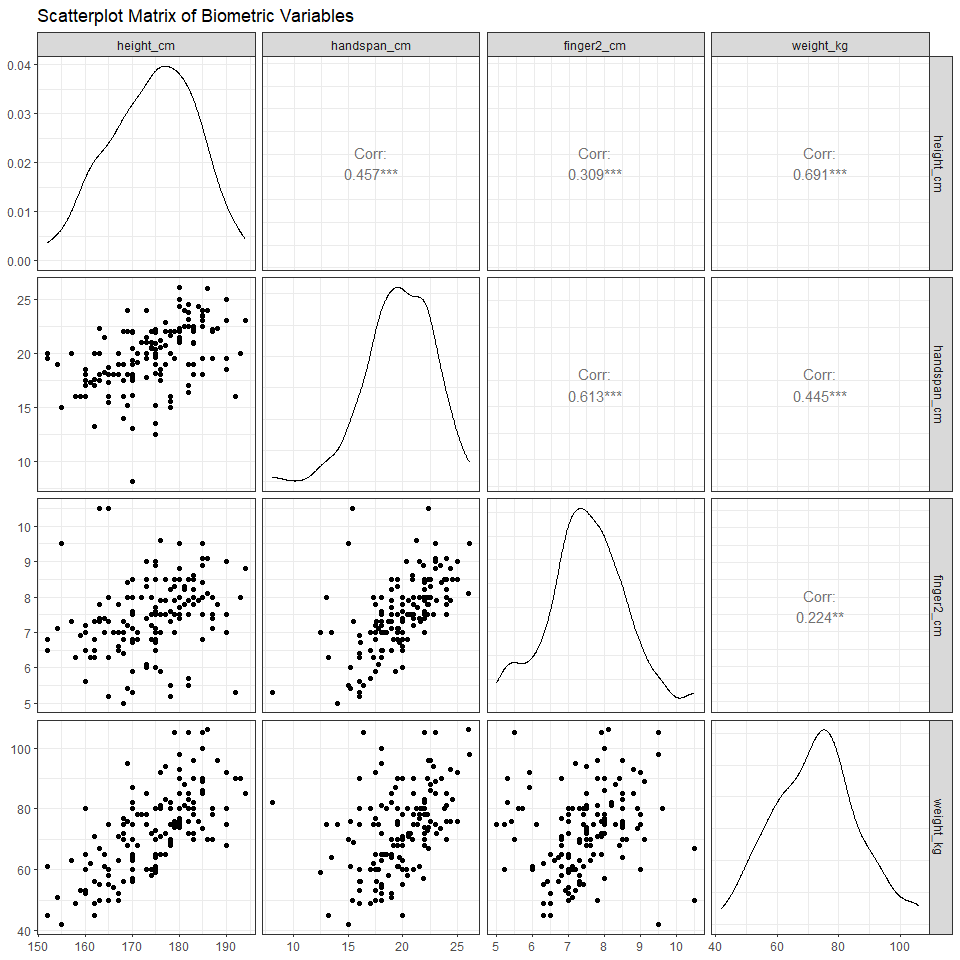
biomet\_0 <- biomet |>  
 drop\_na( finger2\_cm, finger4\_cm) # drop missing values identified in skim()   
  
remove\_outliers <- function(data, column) { # Function to remove outliers   
 outliers <- boxplot.stats(data[[column]])$out # identify outliers  
 data |> filter(!(data[[column]] %in% outliers)) # filter outliers  
}  
  
library(patchwork)

## Warning: package 'patchwork' was built under R version 4.4.2

biomet\_my <- biomet |>  
 drop\_na( finger2\_cm, finger4\_cm) |> # drop missing values identified in skim()   
 remove\_outliers("height\_cm") |> # remove outliers   
 remove\_outliers("weight\_kg") |> # remove outliers   
 slice\_sample(n = 150) # take a random sample  
   
after <- biomet\_my |>  
 ggplot(aes(x = height\_cm, y = weight\_kg)) + # Scatterplot  
 geom\_point() +  
 labs(  
 title = "Scatterplot of Weight vs Height \noutliers/NA removed",  
 x = "Height (cm)",  
 y = "Weight (kg)"  
 ) +  
 theme\_bw()  
   
 befor + after



# Scatterplot matrix for selected variables  
biomet\_my |>  
 select(height\_cm, handspan\_cm, finger2\_cm, weight\_kg) |>  
 ggpairs(title = "Scatterplot Matrix of Biometric Variables") +  
 theme\_bw()



biomet\_my |>  
 select(height\_cm, handspan\_cm, finger2\_cm, weight\_kg) |>  
 ggpairs(  
 lower = list(continuous = wrap("smooth", method = "lm", se = TRUE)),  
 title = "Scatterplot Matrix of Biometric Variables\nwith Regression Lines"  
 ) +  
 theme\_bw()

