Model answer for assignment 2

Luis A. Apiolaza

9 August 2024

In this lab you will work with a data set on biometric features of FORE224-STAT202 students. Students recorded their weight\_kg (response variable in kg), and the follow- ing predictors: height\_cm (in cm), sex, handspan\_cm (distance in cm between the tips of their thumb and little finger), finger2\_cm (length of index in cm) and finger4\_cm (length of annular in cm). We will ignore sex in this lab.

Remember to create a folder for this assignment and to create a new project in that folder. This way R will know where to read files from and where to save output files. *In my setup, all files are located in a folder called data, which explains while all the file names start with data/.*

Your script will need to start with the following code:

library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── 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(readxl)  
library(performance)  
library(GGally) # This is here to later used GGally

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

library(patchwork) # I'm using this for an optional comparison of plots at the end

## Warning: package 'patchwork' was built under R version 4.3.1

## Question 1

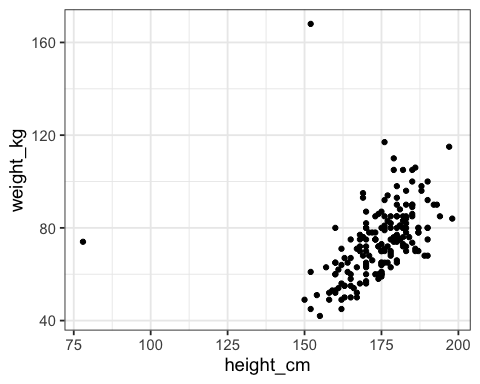
**Read the biometrics.xlsx file and assign it to the name biomet.**

biomet <- read\_excel("data/biometrics.xlsx")

## Question 2 (0.25 points)

**Create a scatterplot where weight\_kg is the response and height\_cm is the predictor. Note: all the scatterplots in this assignment must use theme theme\_bw().**

theme\_set(theme\_bw(base\_size = 14)) # default theme for all plots  
  
biomet |>   
 ggplot(aes(x = height\_cm, y = weight\_kg)) + geom\_point()

 There are two clear outliers: individual shorter than 90 cm tall (and almost 80 kg of weight), and another one who weights over 160 kg (but only 150 cm tall).

## Question 3 (0.25 points)

**Remove the two outliers using code, also remove any observations that have missing values, and take a sample of size 150 observations. Assign the sample to a data frame called my\_biom. You will use my\_biom for the following questions. Remember to use set.seed() with your student number before taking the sample.**

# Replace 1999 with your Student ID number  
set.seed(1999)  
  
my\_biomet <- biomet |>  
 drop\_na(height\_cm, weight\_kg, handspan\_cm, finger2\_cm) |>  
 filter(height\_cm > 90 & weight\_kg < 160) |>  
 sample\_n(150)

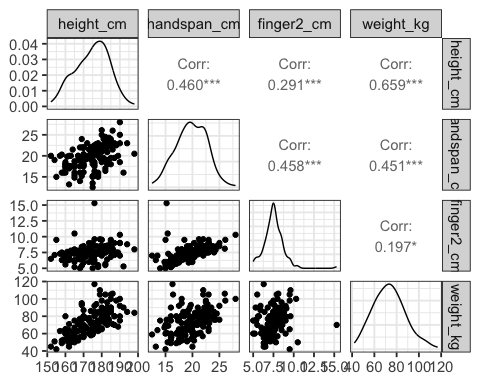
## Question 4 (1 point)

**Create a series of scatterplots between weight\_kg, height\_cm, handspan\_cm and finger2\_cm and write a 50-word comment on the relationships you observe between the variables, including direction (positive, negative) and strength. This requires either creating plots one at the time or just one scatterplot matrix. For the latter you’ll need the ggpairs() function of the GGally package.**

*It is a good idea to list the response as the last variable, so the bottom row of the scatterplot matrix contains the response on y all the time.*

my\_biomet |>   
 ggpairs(columns = c("height\_cm", "handspan\_cm", "finger2\_cm", "weight\_kg"))

## Warning in geom\_point(): All aesthetics have length 1, but the data has 16 rows.  
## ℹ Did you mean to use `annotate()`?

 There is a moderate to strong positive relationship between weight and height, while there is a weak-moderate relationship between weight and hand span, and weight and finger2. There are moderate positive correlations among all predictors.

## Question 5 (1 point)

**Fit five different linear regression models (call them m1, m2, …, m5) using weight\_kg as the response variable and using height\_cm, handspan\_cm, finger2\_cm, height\_cm + handspan\_cm and height\_cm + handspan\_cm + finger2\_cm as predictors. Notice the changes of goodness of fit when moving from models with 1, 2 and 3 predictors. Have a look at R-squared, Adjusted R-squared and residual standard errors for the models. Write a 50-word comment on the improvement of fit when moving from m1 through m5.**

m1 <- lm(weight\_kg ~ height\_cm, data = my\_biomet)  
  
m2 <- lm(weight\_kg ~ handspan\_cm, data = my\_biomet)  
  
m3 <- lm(weight\_kg ~ finger2\_cm, data = my\_biomet)  
  
m4 <- lm(weight\_kg ~ height\_cm + handspan\_cm, data = my\_biomet)  
  
m5 <- lm(weight\_kg ~ height\_cm + handspan\_cm + finger2\_cm, data = my\_biomet)  
  
summary(m1)

##   
## Call:  
## lm(formula = weight\_kg ~ height\_cm, data = my\_biomet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.267 -7.995 -1.927 6.141 42.175   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -107.14949 16.94434 -6.324 2.86e-09 \*\*\*  
## height\_cm 1.03395 0.09688 10.673 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.9 on 148 degrees of freedom  
## Multiple R-squared: 0.4349, Adjusted R-squared: 0.4311   
## F-statistic: 113.9 on 1 and 148 DF, p-value: < 2.2e-16

summary(m2)

##   
## Call:  
## lm(formula = weight\_kg ~ handspan\_cm, data = my\_biomet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.931 -8.107 -2.291 7.604 51.502   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28.3482 7.4145 3.823 0.000193 \*\*\*  
## handspan\_cm 2.2791 0.3709 6.144 7.07e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.94 on 148 degrees of freedom  
## Multiple R-squared: 0.2032, Adjusted R-squared: 0.1979   
## F-statistic: 37.75 on 1 and 148 DF, p-value: 7.073e-09

summary(m3)

##   
## Call:  
## lm(formula = weight\_kg ~ finger2\_cm, data = my\_biomet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.003 -8.859 -0.705 8.192 46.432   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.9315 7.2643 7.700 1.8e-12 \*\*\*  
## finger2\_cm 2.3233 0.9516 2.442 0.0158 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.22 on 148 degrees of freedom  
## Multiple R-squared: 0.03872, Adjusted R-squared: 0.03222   
## F-statistic: 5.961 on 1 and 148 DF, p-value: 0.0158

summary(m4)

##   
## Call:  
## lm(formula = weight\_kg ~ height\_cm + handspan\_cm, data = my\_biomet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.536 -7.344 -2.285 5.439 45.649   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -102.2922 16.6763 -6.134 7.54e-09 \*\*\*  
## height\_cm 0.8991 0.1068 8.419 3.11e-14 \*\*\*  
## handspan\_cm 0.9452 0.3443 2.745 0.00681 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.67 on 147 degrees of freedom  
## Multiple R-squared: 0.4625, Adjusted R-squared: 0.4551   
## F-statistic: 63.23 on 2 and 147 DF, p-value: < 2.2e-16

summary(m5)

##   
## Call:  
## lm(formula = weight\_kg ~ height\_cm + handspan\_cm + finger2\_cm,   
## data = my\_biomet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.764 -7.136 -2.590 5.792 45.107   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -100.8700 16.7259 -6.031 1.28e-08 \*\*\*  
## height\_cm 0.9106 0.1073 8.485 2.21e-14 \*\*\*  
## handspan\_cm 1.0941 0.3724 2.938 0.00384 \*\*   
## finger2\_cm -0.8459 0.8072 -1.048 0.29640   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.66 on 146 degrees of freedom  
## Multiple R-squared: 0.4665, Adjusted R-squared: 0.4555   
## F-statistic: 42.55 on 3 and 146 DF, p-value: < 2.2e-16

The best fit model with one predictor uses height\_cm with an adjusted-R2 of 0.43 and a residual standard error of 10.9.

The model fits slightly improves moving to m4 and m5, with adjusted-R2 increasing to 0.46 and 0.46, while the residuak standard error reduces to 10.67 and 10.66.

## Question 6 (0.25 point)

**Compare models m4 and m5 using the anova(m4, m5) function. Which model fits best?**

anova(m4, m5)

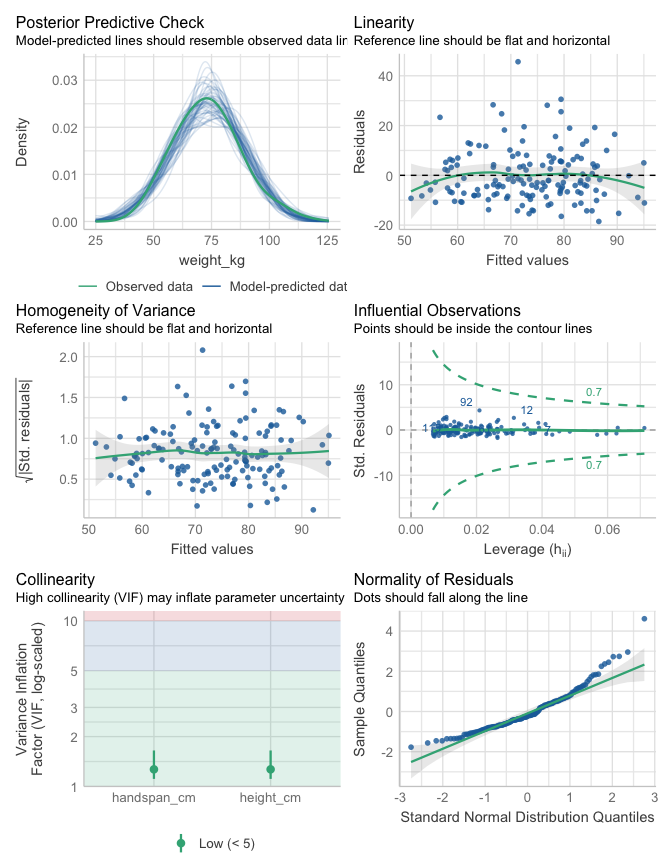
## Analysis of Variance Table  
##   
## Model 1: weight\_kg ~ height\_cm + handspan\_cm  
## Model 2: weight\_kg ~ height\_cm + handspan\_cm + finger2\_cm  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 147 16728   
## 2 146 16603 1 124.88 1.0982 0.2964

Model 5 is not statistically significantly different from model 4. Therefore, we will continue with model 4 as it is simpler.

## Question 7 (0.5 point)

**Considering the best model, have a look at any potential outliers. Use the check\_model() function to visualize the distribution of residuals. Write 50 words explaining what you learned about the residuals of your model.**

check\_model(m4)



1. The Linearity plot shows no important deviation from linearity (residuals vs fitted values plot).
2. The Normality Q-Q plot shows some deviations from normality (away from the dashed line) on both sides of the distribution.
3. The Homogeneity of variance plot shows a slightly increasing residual variance for larger fitted values.
4. There are no strongly influential observations (with a Cook’s Distance larger than 0.5).

## Question 8

**Now we will use the tricarpa.csv data set. Read it in R and name the data frame tricarpa.**\*

tricarpa <- read\_csv("data/tricarpa.csv")

## Rows: 962 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (3): MOE, dry\_density, acoustic\_velocity  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Question 9

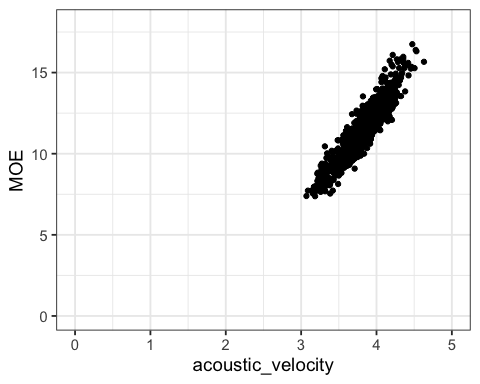
**Reset the random seed to your student ID number and take a sample of 900 observations, called this sample my\_tri.**

set.seed(1999)  
my\_tri <- tricarpa |> sample\_n(900)

## Question 10 (0.25 point)

**Produce a scatterplot of MOE vs acoustic\_velocity; that is, MOE in the y-axis and acoustic\_velocity in the x-axis. This time we want both axes to go all the way to 0. You can add + lims(y = c(0, 18)) to the plot. Use something similar to change the scale of x, taking into account the ranges of the variables.**

my\_tri |>  
 ggplot(aes(x = acoustic\_velocity, y = MOE)) +  
 geom\_point() + lims(x = c(0, 5), y = c(0, 18))



## Question 11 (0.5 point)

**Fit a linear regression model of MOE vs acoustic\_velocity and explain the meaning of the regression coefficients (using the values you estimated) in the context of the problem.**

moe\_vs\_acoustic <- lm(MOE ~ acoustic\_velocity, data = my\_tri)  
summary(moe\_vs\_acoustic)

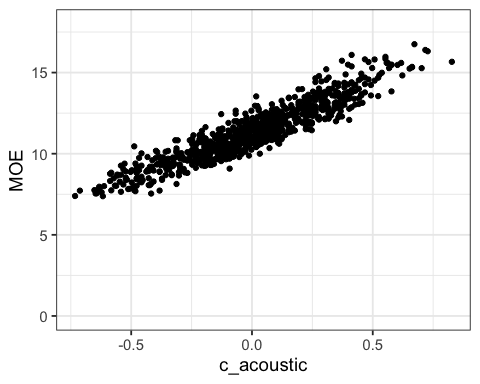
##   
## Call:  
## lm(formula = MOE ~ acoustic\_velocity, data = my\_tri)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71389 -0.46310 -0.02579 0.41095 2.30177   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.19771 0.30691 -36.49 <2e-16 \*\*\*  
## acoustic\_velocity 5.92770 0.08051 73.63 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6514 on 898 degrees of freedom  
## Multiple R-squared: 0.8579, Adjusted R-squared: 0.8577   
## F-statistic: 5421 on 1 and 898 DF, p-value: < 2.2e-16

The intercept means that MOE is -11.2 for an individual with acoustic\_velocity of 0. The slope means that for every additional unit of acoustic\_velocity, MOE increases by 5.93.

## Question 12 (0.5 point)

**Create a new variable in my\_tri, called c\_acoustic, which is the centered version of acoustic\_velocity. Produce a scatterplot of MOE vs c\_acoustic, with the y axis going all the way to 0.**

my\_tri <- my\_tri |>  
 mutate(c\_acoustic = acoustic\_velocity - mean(acoustic\_velocity))  
  
my\_tri |>  
 ggplot(aes(x = c\_acoustic, y = MOE)) + geom\_point() +  
 lims(y = c(0, 18))

 ## Question 13 (0.5 point)

**Fit a linear regression model of MOE vs c\_acoustic and explain the meaning of the regression coefficients (using the values you estimated) in the context of the problem. Refer to the plots produced in items 10- and 12- to help in your answer.**

moe\_vs\_c\_acoustic <- lm(MOE ~ c\_acoustic, data = my\_tri)  
summary(moe\_vs\_c\_acoustic)

##   
## Call:  
## lm(formula = MOE ~ c\_acoustic, data = my\_tri)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71389 -0.46310 -0.02579 0.41095 2.30177   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.34219 0.02171 522.38 <2e-16 \*\*\*  
## c\_acoustic 5.92770 0.08051 73.63 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6514 on 898 degrees of freedom  
## Multiple R-squared: 0.8579, Adjusted R-squared: 0.8577   
## F-statistic: 5421 on 1 and 898 DF, p-value: < 2.2e-16

The slope in this model still means that for every additional unit of acoustic\_velocity, MOE increases by 5.93. In contrast, now the intercept means that MOE is 11.34 for an individual with **average** acoustic\_velocity.

**Bonus comparison**

We are now plotting the non-centered and centered regressions side-by-side. One way of doing this is to create each plot separately (I am highlighting the intercept, and adding some formatting) to later join them using patchwork:

plot1 <- my\_tri |>  
 ggplot(aes(x = acoustic\_velocity, y = MOE)) + geom\_point() +  
 labs(x = "Acoustic velocity", y = "Modulus of Elasticity") +  
 lims(x = c(0, 6), y = c(-12, 18)) +   
 geom\_smooth(method = "lm", se = FALSE, fullrange = TRUE) +  
 geom\_line(data = tibble(acoustic\_velocity = c(0, 0),   
 MOE = c(0, coef(moe\_vs\_acoustic)[1])),   
 colour = "red")  
  
plot2 <- my\_tri |>  
 ggplot(aes(x = c\_acoustic, y = MOE)) + geom\_point() +  
 labs(x = "Acoustic velocity - mean(Acoustic velocity)", y = "Modulus of Elasticity") +  
 lims(y = c(0, 18)) +   
 geom\_smooth(method = "lm", se = FALSE, fullrange = TRUE) +  
 geom\_line(data = tibble(c\_acoustic = c(0, 0),   
 MOE = c(0, coef(moe\_vs\_c\_acoustic)[1])),   
 colour = "red")  
  
# Patchwork permits putting plots side-by-side  
plot1 + plot2

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 15 rows containing missing values or values outside the scale range  
## (`geom\_smooth()`).

## `geom\_smooth()` using formula = 'y ~ x'

