Annand DSE5001 Assignment 7

Joseph Annand

2023-06-24

This assignment includes OpenIntro Chapter exercises 8.6, 8.22, and 8.32, followed by the Chapter 8 Lab.

# Load Packages

library(tidyverse)

## ── 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 3.4.2 ✔ 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(openintro)

## Loading required package: airports  
## Loading required package: cherryblossom  
## Loading required package: usdata

library('statsr')

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

## Loading required package: BayesFactor  
## Loading required package: coda

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

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).  
##   
## Type BFManual() to open the manual.  
## \*\*\*\*\*\*\*\*\*\*\*\*  
##   
## Attaching package: 'statsr'  
##   
## The following objects are masked from 'package:openintro':  
##   
## calc\_streak, evals, nycflights, present

library('broom')

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

# Question 8.6

## Part A

There is a positive relationship between husband’s age and wife’s age, meaning that generally as husband’s age increases, wife’s age increases, too.

## Part B

There is a positive relationship between husband’s height and wife’s height, meaning that generally taller husbands will have taller wives and vice versa.

## Part C

The age plot shows a stronger correlation. The data points are closer together and form a more linear shape.

## Part D

No, the conversion will not affect the correlation. Since both variables are adjusted by the same conversion factor, the shape of the plot is unchanged.

# Question 8.22

## Part A

The relationship between the number of calories and amount of carbohydrates appears to be that, generally, the amount of carbohydrates increases with the number of calories; food menu items with higher calories will generally have more carbohydrates.

## Part B

Calories is the explanatory variable, and carbs is the response variable.

## Part C

We may want to fit a regression to this data to determine how strong the relationship is between calorie content and carbs in a food menu item. If the regression line shows that there is a strong relationship between the two variables, then we may use the regression to predict amount of carbohydrates based on calorie content. If not, then we may need to look for another variable to predict carbohydrates.

## Part D

The data does show a linear trend between number of calories and amount of carbohydrates. The residuals appear to be nearly normal based on their distribution plot. The observations of different food menu items are independent. The residual plot, however, suggests that the variability of the errors are related to the value of the x variable, number of calories. Therefore, all conditions required for fitting a least squares line are not met.

# Question 8.32

## Part A

There is positive, linear relationship between cans of beer and blood alcohol. As the number of cans of beer drank increases, BAC increases.

## Part B

y = -0.0127 + 0.0180\*x

The intercept tells us that a student who drank zero cans of beer would have a BAC of -0.0127 grams per deciliter. The slope tells us that for each can of beer a student drinks, their BAC increases by 0.0180 grams per deciliter.

## Part C

Null hypothesis: Drinking more cans of beer is not associated with an increase in blood alcohol.

Alternate hypothesis: Drinking more cans of beer is associated with an increase in blood alcohol.

The p-value for the beer variable is so low that we may reject the null hypothesis and conclude that drinking more cans of beer is associated with an increase in blood alcohol.

## Part D

bac <- 0.89  
  
r\_squared <- 0.89 ^ 2  
r\_squared

## [1] 0.7921

The R-squared value tells us the strength of the linear relationship between the number of cans of beers drank and blood alcohol. It describes how much of the variation of blood alcohol is explained by the least squares line. In this case, 79.21% of the variation in blood alcohol is explained by the number of cans of beer drank.

## Part E

No, the relationship will not be as strong. The surveyed population will likely be more diverse than the group of 16 students who volunteered for the initial study. With a more diverse, other variables; like gender, weight, and drinking habits; will have a greater effect of the variation in blood alcohol and the r-squared value will be lower.

# Chapter 8 Lab

## Excercise 1

# Display data set  
hfi

## # A tibble: 1,458 × 123  
## year ISO\_code countries region pf\_rol\_procedural pf\_rol\_civil  
## <dbl> <chr> <chr> <chr> <dbl> <dbl>  
## 1 2016 ALB Albania Eastern Europe 6.66 4.55  
## 2 2016 DZA Algeria Middle East & North… NA NA   
## 3 2016 AGO Angola Sub-Saharan Africa NA NA   
## 4 2016 ARG Argentina Latin America & the… 7.10 5.79  
## 5 2016 ARM Armenia Caucasus & Central … NA NA   
## 6 2016 AUS Australia Oceania 8.44 7.53  
## 7 2016 AUT Austria Western Europe 8.97 7.87  
## 8 2016 AZE Azerbaijan Caucasus & Central … NA NA   
## 9 2016 BHS Bahamas Latin America & the… 6.93 6.01  
## 10 2016 BHR Bahrain Middle East & North… NA NA   
## # ℹ 1,448 more rows  
## # ℹ 117 more variables: pf\_rol\_criminal <dbl>, pf\_rol <dbl>,  
## # pf\_ss\_homicide <dbl>, pf\_ss\_disappearances\_disap <dbl>,  
## # pf\_ss\_disappearances\_violent <dbl>, pf\_ss\_disappearances\_organized <dbl>,  
## # pf\_ss\_disappearances\_fatalities <dbl>, pf\_ss\_disappearances\_injuries <dbl>,  
## # pf\_ss\_disappearances <dbl>, pf\_ss\_women\_fgm <dbl>,  
## # pf\_ss\_women\_missing <dbl>, pf\_ss\_women\_inheritance\_widows <dbl>, …

The data set is a 1,458 rows x 123 columns tibble. Each row represents a different country.

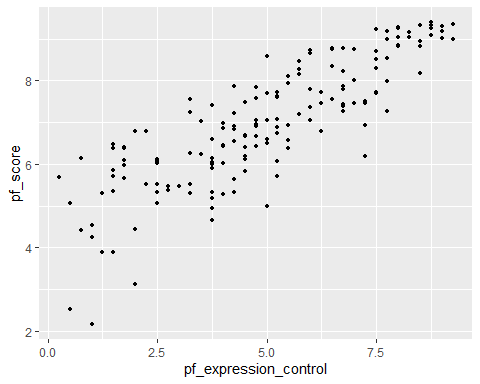
## Exercise 2

# Filter data to have only information from the year 2016  
hfi\_2016 <- filter(hfi, year == 2016)  
hfi\_2016

## # A tibble: 162 × 123  
## year ISO\_code countries region pf\_rol\_procedural pf\_rol\_civil  
## <dbl> <chr> <chr> <chr> <dbl> <dbl>  
## 1 2016 ALB Albania Eastern Europe 6.66 4.55  
## 2 2016 DZA Algeria Middle East & North… NA NA   
## 3 2016 AGO Angola Sub-Saharan Africa NA NA   
## 4 2016 ARG Argentina Latin America & the… 7.10 5.79  
## 5 2016 ARM Armenia Caucasus & Central … NA NA   
## 6 2016 AUS Australia Oceania 8.44 7.53  
## 7 2016 AUT Austria Western Europe 8.97 7.87  
## 8 2016 AZE Azerbaijan Caucasus & Central … NA NA   
## 9 2016 BHS Bahamas Latin America & the… 6.93 6.01  
## 10 2016 BHR Bahrain Middle East & North… NA NA   
## # ℹ 152 more rows  
## # ℹ 117 more variables: pf\_rol\_criminal <dbl>, pf\_rol <dbl>,  
## # pf\_ss\_homicide <dbl>, pf\_ss\_disappearances\_disap <dbl>,  
## # pf\_ss\_disappearances\_violent <dbl>, pf\_ss\_disappearances\_organized <dbl>,  
## # pf\_ss\_disappearances\_fatalities <dbl>, pf\_ss\_disappearances\_injuries <dbl>,  
## # pf\_ss\_disappearances <dbl>, pf\_ss\_women\_fgm <dbl>,  
## # pf\_ss\_women\_missing <dbl>, pf\_ss\_women\_inheritance\_widows <dbl>, …

## Exercise 3

# Create a scatter plot with expression control on x-axis and personal freedom score on the y axis using 2016 data from all countries  
ggplot(hfi\_2016,   
 mapping = aes(x = pf\_expression\_control, y = pf\_score)  
) + geom\_point(size = 1.0)



# Determine correlation coefficient of the two variables from scatter plot  
hfi\_2016 %>%  
 summarise(cor(pf\_expression\_control, pf\_score))

## # A tibble: 1 × 1  
## `cor(pf\_expression\_control, pf\_score)`  
## <dbl>  
## 1 0.845

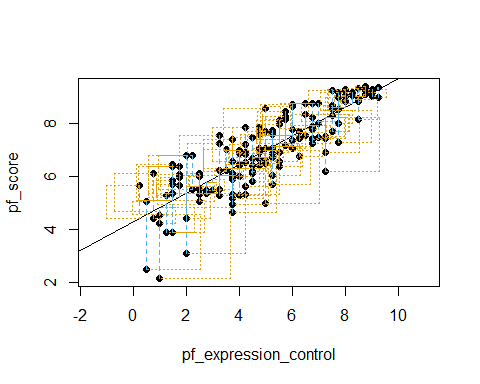
We use a scatter plot to display the relationship between the personal freedom score and expression control since we are comparing two continuous variables. The relationship does appear to be linear. The correlation coefficient shows a strong, positive, and linear relationship between the two variables. This suggests that a linear model may be applied to try to predict personal freedom score. To be comfortable with actually using the model, the model should be evaluated using inference and/or a hypothesis test.

## Exercise 4

The relationship appears to be linear. The relationship is positive, meaning as pf\_expression\_control increases, personal freedom score also increases. The correlation coefficient is 0.85 and suggest that the relationship is strong. The variability of the data seems to be greater in the lower end of the plot than in the higher end of the plot, which may inhibit applying a linear model.

## Exercise 5

# Create interactive plot to find minimum sum of squares through trial and error  
plot\_ss(x = pf\_expression\_control, y = pf\_score, data = hfi\_2016, showSquares = TRUE)



## Click two points to make a line.   
## Call:  
## lm(formula = y ~ x, data = pts)  
##   
## Coefficients:  
## (Intercept) x   
## 4.2838 0.5418   
##   
## Sum of Squares: 102.213

The smallest sum of squares I got during the interactive exercise was 6505.454 with an intercept of 11.2388 and x of 0.4062.

## Exercise 6

# Create linear model that uses least squares regression for total human freedom using expression control as predictor variable  
m1 <- lm(hf\_score ~ pf\_expression\_control, data = hfi\_2016)  
  
# Get information on the model  
tidy(m1)

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 5.05 0.123 41.1 5.97e-87  
## 2 pf\_expression\_control 0.368 0.0224 16.5 2.73e-36

# Get r-squared value  
glance(m1)

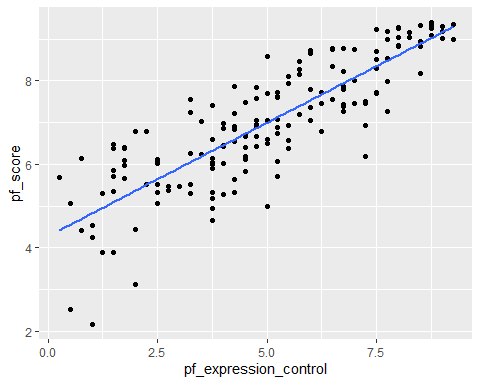
## # A tibble: 1 × 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.629 0.627 0.660 271. 2.73e-36 1 -161. 329. 338.  
## # ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

The equation of the regression line is y = 5.05 + 0.368 \* x. The slope tells us that for every one unit that pf\_expression\_control increases, the total human freedom score increases by 0.368 units.

## Exercise 7

# Create linear regression for personal freedom using expression control as predictor variable  
m2 <- lm(pf\_score ~ pf\_expression\_control, data = hfi\_2016)  
  
# Plot the linear regression  
ggplot(data = hfi\_2016, aes(x = pf\_expression\_control, y = pf\_score)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE)

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



They could predict a country’s personal freedom score by using the equation of the line and set x equal to 3.

# Find predicted personal freedom score given that expression control is 3  
y\_hat <- 4.28 + 0.542 \* 3  
y\_hat

## [1] 5.906

# Find observation in data set where expression control is equal to 3  
select(hfi\_2016, countries, pf\_score, pf\_expression\_control) %>%  
 filter(pf\_expression\_control == 3)

## # A tibble: 1 × 3  
## countries pf\_score pf\_expression\_control  
## <chr> <dbl> <dbl>  
## 1 Central Afr. Rep. 5.47 3

The model predicts that a country with a pf\_expression\_control of 3 will have a pf\_score of 5.906. We can see from our actual data set that the Central African Republic, which has a pf\_expression\_control of 3, has a pf\_score of 5.47.

# Determine residual  
res\_3 <- 5.47 - 5.906  
res\_3

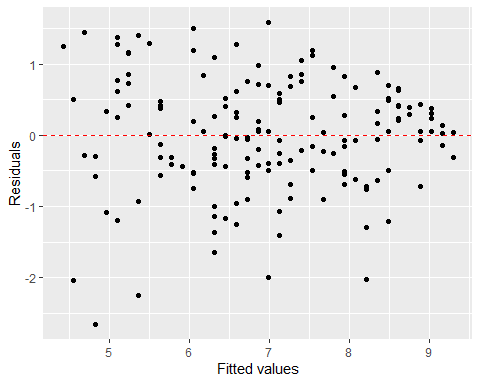
## [1] -0.436

The residual for this prediction is -0.436, so the model overestimated.

## Exercise 8

# augment calculated predicted values and residuals  
m2\_aug <- augment(m2)

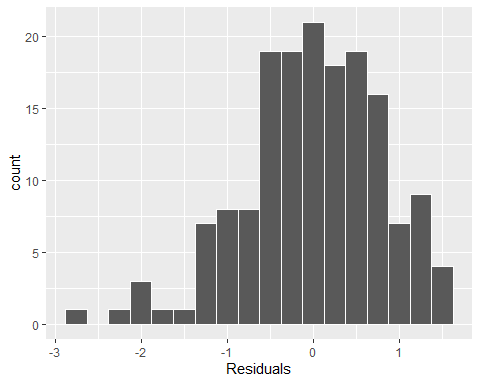
# Create a residual plot using m2 predicted values and residuals  
ggplot(data = m2\_aug, aes(x = .fitted, y = .resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, linetype = "dashed", color = "red") +  
 xlab("Fitted values") +  
 ylab("Residuals")



The residual plot shows that the points are distributed around zero and there is no apparent curvature, so the linearity condition is satisfied.

## Exercise 9

# Create a histogram showing the normal distribution for residuals  
ggplot(data = m2\_aug, aes(x = .resid)) +  
 geom\_histogram(binwidth = 0.25, color = "white") +  
 xlab("Residuals")



The residuals do not appear to be nearly normally distributed. Based on the histogram, the distribution is left-skewed.

## Exercise 10

The constant variability condition does appear to be violated. The variability of the pf\_score is greater at lower pf\_expression\_control values.