Assignment 1

Holly Finertie (HF2379)

1/17/2020

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 1.0.0 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ───────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)

## Problem 1:

Please find tables below that provide summaries (min, mean, median, IQR, max) of the quantitative features of the dataset.

data = read\_csv("./data/data.csv") %>%   
 janitor::clean\_names()

## Parsed with column specification:  
## cols(  
## Age = col\_double(),  
## BMI = col\_double(),  
## Glucose = col\_double(),  
## Insulin = col\_double(),  
## HOMA = col\_double(),  
## Leptin = col\_double(),  
## Adiponectin = col\_double(),  
## Resistin = col\_double(),  
## MCP.1 = col\_double(),  
## Classification = col\_double()  
## )

summary(data)

## age bmi glucose insulin   
## Min. :24.0 Min. :18.37 Min. : 60.00 Min. : 2.432   
## 1st Qu.:45.0 1st Qu.:22.97 1st Qu.: 85.75 1st Qu.: 4.359   
## Median :56.0 Median :27.66 Median : 92.00 Median : 5.925   
## Mean :57.3 Mean :27.58 Mean : 97.79 Mean :10.012   
## 3rd Qu.:71.0 3rd Qu.:31.24 3rd Qu.:102.00 3rd Qu.:11.189   
## Max. :89.0 Max. :38.58 Max. :201.00 Max. :58.460   
## homa leptin adiponectin resistin   
## Min. : 0.4674 Min. : 4.311 Min. : 1.656 Min. : 3.210   
## 1st Qu.: 0.9180 1st Qu.:12.314 1st Qu.: 5.474 1st Qu.: 6.882   
## Median : 1.3809 Median :20.271 Median : 8.353 Median :10.828   
## Mean : 2.6950 Mean :26.615 Mean :10.181 Mean :14.726   
## 3rd Qu.: 2.8578 3rd Qu.:37.378 3rd Qu.:11.816 3rd Qu.:17.755   
## Max. :25.0503 Max. :90.280 Max. :38.040 Max. :82.100   
## mcp\_1 classification   
## Min. : 45.84 Min. :1.000   
## 1st Qu.: 269.98 1st Qu.:1.000   
## Median : 471.32 Median :2.000   
## Mean : 534.65 Mean :1.552   
## 3rd Qu.: 700.09 3rd Qu.:2.000   
## Max. :1698.44 Max. :2.000

## Problem 2:

Using code below, the continuous BMI variable was transformed into a categorical variable (bmi\_cat) with the following BMI categories:

* Severely underweight: BMI < 16.5kg/m^2
* Underweight: 16.5 <= BMI <= 18.5 kg/m^2
* Normal weight: 18.5 <= BMI <=24.9 kg/m^2
* Overweight: 25 <= BMI <= 29.9 kg/m^2
* Obesity class I: 30 <= BMI <= 34.9 kg/m^2
* Obesity class II: 35 <= BMI <= 39.9 kg/m^2
* Obesity class III: BMI >= 40 kg/m^2

data\_bmi = data %>%   
 mutate(bmi\_cat = as.factor(case\_when(  
 bmi <= 16.4 ~ "Severely underweight",   
 bmi >= 16.5 & bmi <= 18.4 ~ "Underweight",  
 bmi >=18.5 & bmi <= 24.9 ~ "Normal weight",  
 bmi >= 25 & bmi <= 29.9 ~ "Overweight",   
 bmi >= 30 & bmi <= 34.9 ~ "Obesity 1",   
 bmi >= 35 & bmi <= 39.9 ~ "Obesity 2",   
 bmi >= 40 ~ "Obesity 3"  
 )),   
 bmi\_cat = fct\_relevel(  
 bmi\_cat,   
 str\_c(c("Underweight", "Normal weight",   
 "Overweight", "Obesity 1", "Obesity 2")))  
 )

## Problem 3:

data\_final = data\_bmi %>%   
 mutate(  
 Arm = recode(classification,   
 `1` = "control",   
 `2` = "case"),   
 outcome = recode(Arm,   
 "control" = 0,   
 "case" = 1))  
   
  
plot = data\_final %>%   
 ggplot(aes(x = bmi\_cat, fill = Arm)) +  
 geom\_bar(stat = "count") +   
 xlab("BMI Category") +  
 ylab("Count of Breast CAncer Cases and Controls") +  
 labs(  
 title = "Proportion of Breast Cancer Cases and Controls by BMI Category"  
 )  
  
plot

A screenshot of a social media post

Description automatically generated

## Problem 4:

logit\_reg =   
 glm(outcome ~ glucose + homa + leptin + bmi + age,   
 family = binomial(link = "logit"), data = data\_final) %>%   
 broom::tidy() %>%   
 mutate(  
 "Lower Limit" = estimate - (std.error\*1.96),   
 "Upper Limit" = estimate + (std.error\*1.96)  
 ) %>% filter(term == "homa") %>%   
 select(term, estimate, "Lower Limit", "Upper Limit") %>%   
 knitr::kable()  
  
  
logit\_reg

|  |  |  |  |
| --- | --- | --- | --- |
| term | estimate | Lower Limit | Upper Limit |
| homa | 0.2738822 | -0.0631907 | 0.6109551 |

As seen in the table above, the beta estimate associated with a 1-unit change in HOMA is 0.274 with 95% CI(-0.063, 0.611).

## Problem 5:

linear\_reg =   
 lm(insulin ~ bmi + age + glucose, data = data\_final) %>%   
 broom::tidy() %>%   
 mutate(  
 "Lower Limit" = estimate - (std.error\*1.96),   
 "Upper Limit" = estimate + (std.error\*1.96)  
 ) %>% filter(term == "age") %>%   
 select(term, estimate, "Lower Limit", "Upper Limit") %>%   
 knitr::kable()  
  
linear\_reg

|  |  |  |  |
| --- | --- | --- | --- |
| term | estimate | Lower Limit | Upper Limit |
| age | -0.0540217 | -0.1558221 | 0.0477787 |

As seen in the table above, the beta estimate associated with a 1-unit change in age is -0.054 with 95% CI(-0.156, 0.048).