Assignment\_1

1/23/2022

First, load in the required libraries and the data.

library(tidyverse)  
bc\_data = read.csv("bcdata.csv")

### Question #1

Construct a table providing summaries of the quantitative features of the dataset.Summaries should include the mean, median, minimum value, and maximum value.

summary\_data = summary(bc\_data)  
as.data.frame.matrix(summary\_data) %>%   
 knitr::kable()

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | BMI | Glucose | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP.1 | Classification |
| X | Min. :24.0 | Min. :18.37 | Min. : 60.00 | Min. : 2.432 | Min. : 0.4674 | Min. : 4.311 | Min. : 1.656 | Min. : 3.210 | Min. : 45.84 | Min. :1.000 |
| X.1 | 1st Qu.:45.0 | 1st Qu.:22.97 | 1st Qu.: 85.75 | 1st Qu.: 4.359 | 1st Qu.: 0.9180 | 1st Qu.:12.314 | 1st Qu.: 5.474 | 1st Qu.: 6.882 | 1st Qu.: 269.98 | 1st Qu.:1.000 |
| X.2 | Median :56.0 | Median :27.66 | Median : 92.00 | Median : 5.925 | Median : 1.3809 | Median :20.271 | Median : 8.353 | Median :10.828 | Median : 471.32 | Median :2.000 |
| X.3 | Mean :57.3 | Mean :27.58 | Mean : 97.79 | Mean :10.012 | Mean : 2.6950 | Mean :26.615 | Mean :10.181 | Mean :14.726 | Mean : 534.65 | Mean :1.552 |
| X.4 | 3rd Qu.:71.0 | 3rd Qu.:31.24 | 3rd Qu.:102.00 | 3rd Qu.:11.189 | 3rd Qu.: 2.8578 | 3rd Qu.:37.378 | 3rd Qu.:11.816 | 3rd Qu.:17.755 | 3rd Qu.: 700.09 | 3rd Qu.:2.000 |
| X.5 | Max. :89.0 | Max. :38.58 | Max. :201.00 | Max. :58.460 | Max. :25.0503 | Max. :90.280 | Max. :38.040 | Max. :82.100 | Max. :1698.44 | Max. :2.000 |

### Question 2

Recode BMI into the WHO-defined categories

new\_data = bc\_data %>%  
 mutate(  
 BMI = case\_when(  
 BMI < 16.5 ~ "Severely underweight",  
 BMI > 16.5 & BMI < 18.5 ~ "Underweight",  
 BMI >= 18.5 & BMI <= 24.9 ~ "Normal weight",   
 BMI >= 25 & BMI <= 29.9 ~ "Overweight",  
 BMI >= 30 & BMI <= 34.9 ~ "Obsesity class I",  
 BMI >= 35 & BMI <=39.9 ~ "Obesity class II",  
 TRUE ~ "Obesity class III")  
 ) %>%   
 mutate(  
 BMI = factor(BMI, levels = c("Severely underweight", "Underweight", "Normal weight", "Overweight", "Obsesity class I", "Obesity class II", "Obesity class III"))  
 )  
  
str(new\_data$BMI)

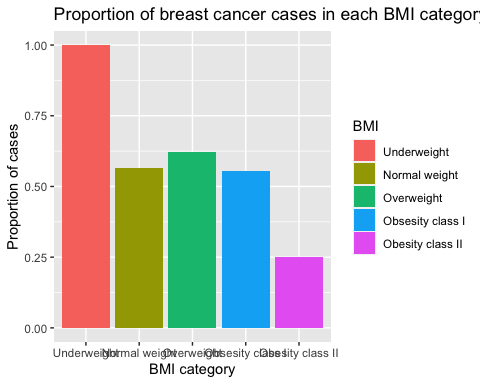
## Factor w/ 7 levels "Severely underweight",..: 3 3 3 3 3 3 3 3 3 3 ...

### Question 3

Create a bar chart showing the proportion of breast cancer cases and controls within each BMI category.

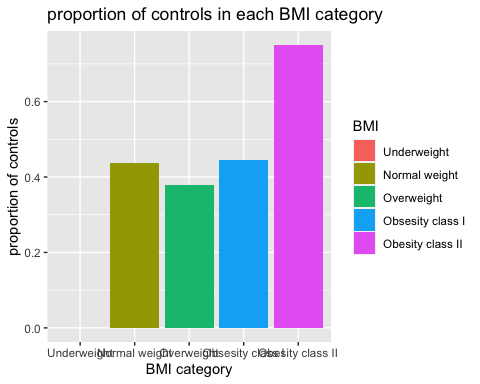
#### Proportion of Breast Cancer

new\_data2 =  
new\_data %>%   
 group\_by(BMI) %>%   
 summarize(  
 control = sum(Classification == 1) / n(),  
 bc = sum(Classification == 2)/ n()  
 ) %>%   
 ggplot(aes(x = BMI, y = bc, fill = BMI)) +  
 geom\_bar(stat = "identity") +  
 ggtitle("Proportion of breast cancer cases in each BMI category") +  
 xlab("BMI category") + ylab("Proportion of cases")  
  
new\_data2



#### Proportion of Control

new\_data3 =  
new\_data %>%   
 group\_by(BMI) %>%   
 summarize(  
 control = sum(Classification == 1) / n(),  
 bc = sum(Classification == 2)/ n()  
 ) %>%   
 ggplot(aes(x = BMI, y = control, fill = BMI)) +  
 geom\_bar(stat = "identity") +   
 ggtitle("proportion of controls in each BMI category") +  
 xlab("BMI category") + ylab("proportion of controls")  
  
new\_data3



### Question 4

Construct a **logistic** regression model:

* outcome: breast cancer classification
* independent variables: glucose, HOMA-IR, leptin, BMI (continuous), age

Fill in the beta estimate and 95% confidence interval associated with a 1-unit change in HOMA-IR.

model\_data = bc\_data %>%   
 janitor::clean\_names() %>%   
 mutate(  
 classification = ifelse(classification == 1, 0, 1)  
 )  
  
logistic\_model = glm(classification ~ glucose + homa + leptin + bmi + age, data = model\_data, family = "binomial")  
  
summary(logistic\_model)  
##   
## Call:  
## glm(formula = classification ~ glucose + homa + leptin + bmi +   
## age, family = "binomial", data = model\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2944 -0.8901 0.1308 0.8084 2.1371   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.626065 2.355177 -1.540 0.123654   
## glucose 0.081699 0.023526 3.473 0.000515 \*\*\*  
## homa 0.273882 0.171976 1.593 0.111259   
## leptin -0.008574 0.015783 -0.543 0.586979   
## bmi -0.104261 0.056642 -1.841 0.065668 .   
## age -0.022881 0.014377 -1.592 0.111496   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 159.57 on 115 degrees of freedom  
## Residual deviance: 120.81 on 110 degrees of freedom  
## AIC: 132.81  
##   
## Number of Fisher Scoring iterations: 6  
confint(logistic\_model)   
## 2.5 % 97.5 %  
## (Intercept) -8.54138756 0.754487774  
## glucose 0.03956613 0.132397841  
## homa -0.02555240 0.653222623  
## leptin -0.04019445 0.022416142  
## bmi -0.21944692 0.004398024  
## age -0.05192184 0.004856327

For every one unit change in HOMA-IR, the log odds of having breast cancer increases by **0.2738822** (95% CI: -0.025, 0.653).

<https://stats.oarc.ucla.edu/r/dae/logit-regression/>

### Question 5

Construct a **linear** regression model:

* outcome: insulin
* independent variables: BMI (continuous), age, glucose

Fill in the beta estimate and 95% confidence interval associated with a 1-unit change in age.

linear\_model = lm(insulin ~ bmi + age + glucose, data = model\_data)  
summary(linear\_model)  
##   
## Call:  
## lm(formula = insulin ~ bmi + age + glucose, data = model\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.161 -4.359 -2.118 2.124 46.269   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13.49576 5.85941 -2.303 0.0231 \*   
## bmi 0.14969 0.16382 0.914 0.3628   
## age -0.05402 0.05194 -1.040 0.3005   
## glucose 0.22982 0.03752 6.126 1.37e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.731 on 112 degrees of freedom  
## Multiple R-squared: 0.2675, Adjusted R-squared: 0.2479   
## F-statistic: 13.64 on 3 and 112 DF, p-value: 1.207e-07  
confint(linear\_model)  
## 2.5 % 97.5 %  
## (Intercept) -25.1054353 -1.88608318  
## bmi -0.1748942 0.47427491  
## age -0.1569321 0.04888876  
## glucose 0.1554864 0.30414939

For every one year increase in age, the insulin decreases by **-0.0540217** microU/mL (95% CI: -0.157, 0.048).