Machine Learning Precourse Assignment

Dionna Attinson, dna2118

1/20/2020

#### Step 1: First, we will read in the dataset and name it Data

Data = read\_csv("./dataR2.csv")

## 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()  
## )

#### Step 2: Next, we will create a summary output of the dataset

Data %>%   
 select(-Classification) %>%   
 summary(Data) %>%   
 knitr::kable(digits = 1)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | BMI | Glucose | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP.1 |
|  | 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 |
|  | 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 |
|  | 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 |
|  | 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 |
|  | 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 |
|  | 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 |

#### Step 3: Now, we will recode BMI into the WHO-defined categories below

* +Severely underweight - BMI less than 16.5kg/m^2
* +Underweight - BMI under 18.5 kg/m^2
* +Normal weight - BMI greater than or equal to 18.5 to 24.9 kg/m^2
* +Overweight – BMI greater than or equal to 25 to 29.9 kg/m^2
* +Obesity class I – BMI 30 to 34.9 kg/m^2
* +Obesity class II – BMI 35 to 39.9 kg/m^2
* +Obesity class III – BMI greater than or equal to 40 kg/m^2

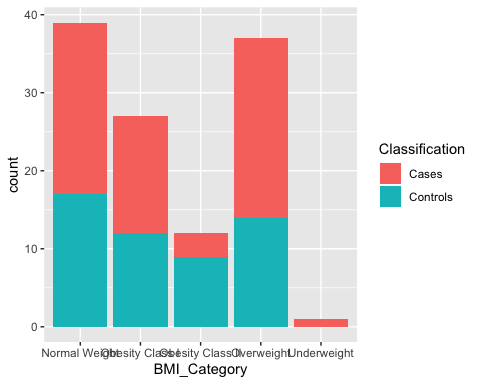
Data =   
 Data %>%   
 mutate(  
 BMI\_Category = BMI,  
 BMI\_Category = as.numeric(BMI\_Category),  
 Classification = as.character(Classification))

Data =   
 Data %>%   
 mutate(BMI\_Category =   
 case\_when(BMI\_Category >= 40 ~ 'Obesity Class III',  
 BMI\_Category >= 35 & BMI\_Category <= 39.9 ~ 'Obesity Class II',  
 BMI\_Category >= 30 & BMI\_Category <= 34.9 ~ 'Obesity Class I',  
 BMI\_Category >= 25 & BMI\_Category <= 29.9 ~ 'Overweight',  
 BMI\_Category >= 18.5 & BMI\_Category <= 24.9 ~ 'Normal Weight',  
 BMI\_Category >= 16.5 & BMI\_Category <= 18.4 ~ 'Underweight',  
 BMI\_Category <= 16.4 ~ "Severely Underweight"))

#### Step 4: Now we will create a bar chart showing the proportion of breast cancer cases and controls within each BMI category

PlotData =   
Data %>%   
 mutate(Classification =   
 recode(Classification,  
 "1" = "Controls",  
 "2" = "Cases"))

PlotData %>%   
ggplot(aes(x=BMI\_Category, fill=Classification)) +  
geom\_bar()



#### Step 5: Now we will create a logistic and linear regression model using the variables provided

* Construct a logistic regression model using breast cancer classification as the outcome and glucose, HOMA, leptin, BMI (continuous) and age as the independent variables.
* Fill in the beta estimate and 95% confidence interval associated with a 1-unit change in HOMA5.
* Construct a linear regression model using insulin as the outcome and BMI (continuous), age, and glucose as the independent variables.
* Fill in the beta estimate and 95% confidence interval associated with a 1-unit change in age.

Data\_Regression =   
 Data %>%   
 mutate(  
 Classification = recode(Classification,   
 "1" = "0",   
 "1" = "0",  
 "2" = "1",  
 "2" = "1"),  
 Classification = as.numeric(Classification))

model1 =   
 Data\_Regression %>%   
 lm(Classification ~ Glucose + HOMA + Leptin + BMI + Age, data = ., family = binomial())

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

model1 %>%   
 broom::tidy() %>%   
 filter(term == "HOMA") %>%   
 mutate(OR = exp(estimate),  
 High\_CI = exp(estimate + 1.96\*std.error),  
 Low\_CI = exp(estimate - 1.96\*std.error)) %>%   
 select(term, log\_OR = estimate, OR, p.value, Low\_CI, High\_CI) %>%   
 knitr::kable(digits = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| term | log\_OR | OR | p.value | Low\_CI | High\_CI |
| HOMA | 0.005 | 1.005 | 0.775 | 0.973 | 1.038 |

#### In the logistic regression above, we see that at a 5% level of significance, cases have 1.315 times the odds for HOMA in comparison to the control group. We are 95% confident that the true odds lies between 0.973 and 1.038.

model2 =   
 Data\_Regression %>%   
 lm(Insulin ~ BMI + Age + Glucose, data = .,)  
  
model2 %>%   
 broom::tidy() %>%   
 filter(term == "Age") %>%   
 mutate(High\_CI = exp(estimate + 1.96\*std.error),  
 Low\_CI = exp(estimate - 1.96\*std.error)) %>%   
 select(term, estimate, p.value, Low\_CI, High\_CI) %>%   
 knitr::kable(digits = 3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | p.value | Low\_CI | High\_CI |
| Age | -0.054 | 0.301 | 0.856 | 1.049 |

#### In the linear regression above, we see that at a 5% level of signficance, for every one unit increase in age, insulin decreases by 0.054.