Data 2010 Project

Lucas Berzuk, Ethan Robson, Hugo Ng, Rodell Salonga

April 5, 2024

Introduction

Abstract

Obesity is a chronic disease that is defined by having excessive amount of body fat. It increases the risk of other diseases and health problems including heart disease, high blood pressure, diabetes and certain cancers. First world countries are experiencing an epidemic of obesity in recent years. This includes Canada, which reported a 30% obesity rate out of all Canadian adults, an increase from the 21% reported in 2003 [2]. There must be a change and this includes diagnosing obesity early on for people and providing the preventative therapies to decrease body fat percentage. This involves measuring levels of obesity in people.

There are many different tools for measuring obesity levels, these include the body mass index, waist-to-hip ratio, skin fold thickness, and other more costly and intense body fat measuring procedures. The most common measure is the body mass index, BMI for short, which is calculated by using an individuals weight and height, with the following formula: $BMI = \frac{weight}{height^2}$. However in many cases the height and weight of individuals are not available in the diagnosing of obesity level.

In this paper we investigate different models that can be to predict obesity levels based on other variables other than height and weight. The models will be based on the following variables.

We build various regression models with the target variable being BMI and using the rest of the variables in the dataset to see if we are able to predict obesity with the lifestyle of a person. We end up using Linear, Lasso, and Ridge regression to build our models and calculate the root mean squared error (RMSE).

//**TODO** We then build classification models including...

Data Analysis

This data comes from a study performed in Mexico, Peru and Colombia. It has 17 attributes and 2111 data points. The 17 attributes comprise 4 different variable types: 6 continuous, 2 ordinal, 5 categorical, and 4 binary. This data consists of an individual's behavioural patterns such as eating habits and physical condition, as well as their level of obesity. It is important to note that up to 77% of this data has been synthetically generated because of a greatly unbalanced number of samples between the different obesity categories in the sample data. The balancing was performed using the SMOTE filter with the Weka tool. The purpose was to create data points within the obesity categories with lower counts. This lead to an overall data set with uniform counts for the different obesity levels. The data was collected through a voluntary online survey that had 16 questions with a variety of responses. It was accessible for 30 days for users to complete. The researchers of this study calculated Body Mass Index using the equation $BMI = \frac{weight}{height^2}$ to determine the BMI of every respondent. The BMI values were then placed into different obesity level categories based on the World Health Organization and Mexican Normativity. The categories can be observed in Table 1.

Table 1: Categorizing Obesity Levels

Obesity Level	Insufficient Weight	Normal Weight	Overweight Level I	Overweight Level II	Obesity Type I	Obesity Type II	Obesity Type III
BMI	<18.5	18.5-22.9	23.0-25.9	26.0-29.9	30.0-34.9	35.0-39.9	>40

//TODO - Talk about our preprocessing stuff (do we need this????)

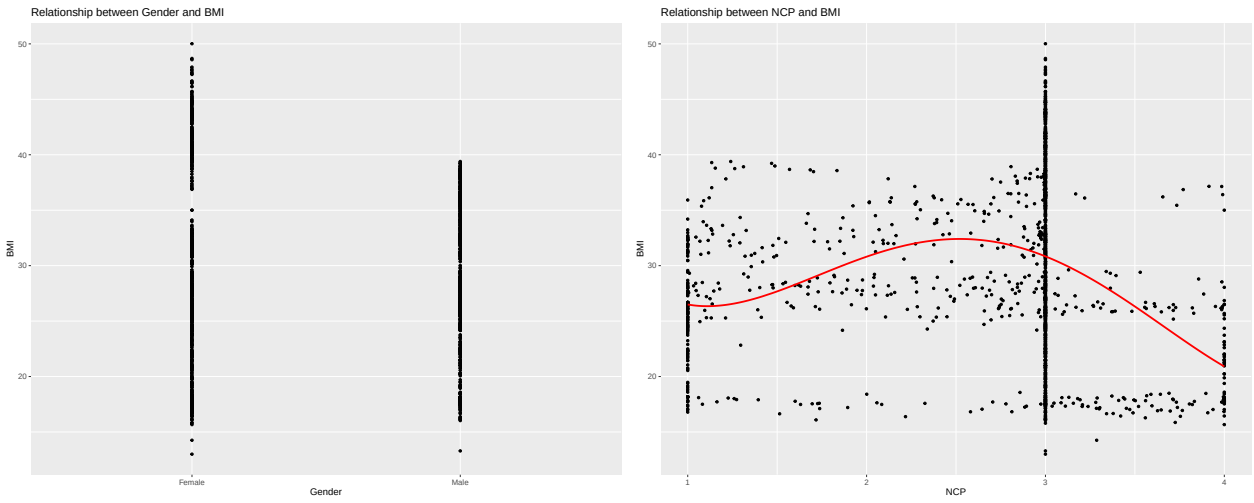
Methods

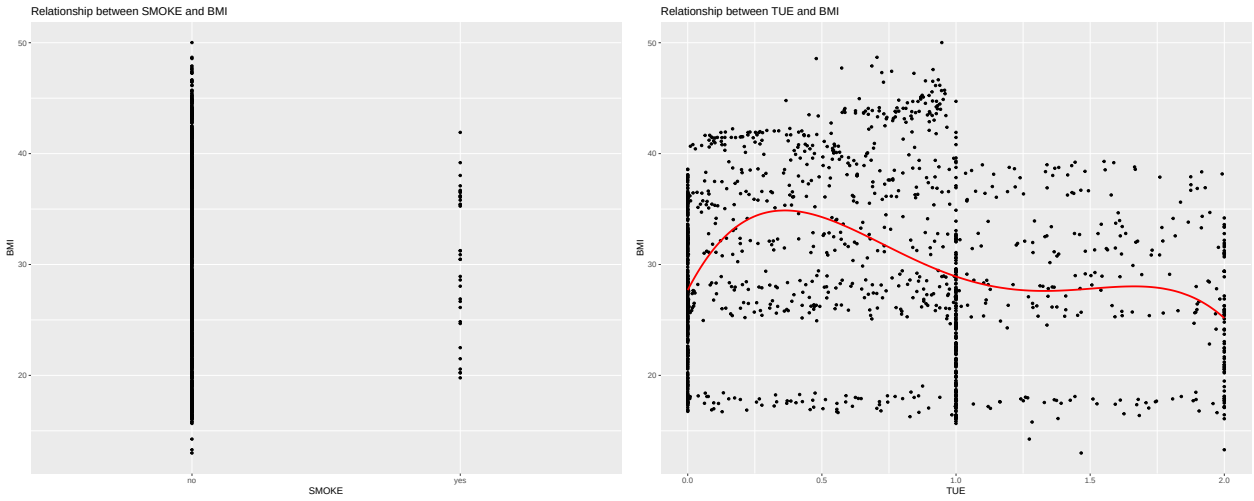
Regression

We decide to model many forms of regression on our data to calculate RMSE and see how far off predicted values would be from observed values for this obesity dataset. We first started off by performing standard Linear regression with BMI where $BMI = \frac{weight}{height^2}$, on every variable that was in the dataset (aside from height and weight because that is what is used in BMI). For the categorical variables we had to factor them before performing regression because otherwise we would not be able to test them. These are the RMSE values that we found for every variable in the dataset.

```
## Gender = 7.939414
## Age = 8.178174
## History of overweight = 9.08678
## Freq consumption of high caloric food = 8.134233
## Freq consumption of vegetables = 8.167971
## Number of main meals = 7.932018
## Consumption of food between meals = 8.807099
## Smoke = 7.929959
## Consumption of water = 8.065398
## Calorie consumption monitoring = 8.106003
## Freq of physical activity = 8.109581
## Time using electronics = 7.973657
## Consumption of alcohol = 8.130043
## Transportation used = 8.07459
```

Noticing that *Gender*, *Number of main meals*, *Smoke* and *Time using electronics* were the variables that lead to the lowest RMSE values (under 8), we decided to graph these variables with cubic regression for visual input. Here are those graphs.





After some consideration, we decided to try spline regression on the 2 numerical variables out of these 4 to see if that would improve the RMSE values any further. These were the results:

```
## Number of main meals = 8.540398
## Time using electronics = 8.488051
```

It turns out that spline regression did not improve the error in our case so we had to try something else.

We moved to attempt Lasso regression to yet again try and improve our values. When performing Lasso regression on the 4 original variables that all had RMSE under 8, this is what we got with a lambda of 1.2.

```
## RMSE = 7.968656
```

This is a good RMSE value but it is not lower than any 1 of the 4 original RMSE values gained from using Linear regression. Because we still can not seem to get it lower that the original 4, we decided to do Lasso regression with all of the variables and this was the result with a lambda of 4:

```
## RMSE = 7.968656
```

This turned out to give us the exact same RMSE. Maybe we could try and use less variables? This is with just the Gender and SMOKE variable at the lambda of 0.27.

```
## RMSE = 7.96817
```

The result is just slightly lower than the previous 2 Lasso regression attempts so it seems like we may be at a dead end here. But we tried one more just to see what happened. Ridge regression gave us similar RMSE values to Lasso but just slightly higher no matter what variables we used in the model. The range of BMIs that we acquired from the dataset are from:

```
## Minimum = 12.99868
```

```
## Maximum = 50.01402
```

So the RMSE floating around 8 is not the end of the world but we definitely wish that we could improve it more.

Next we wanted to try to see if there was any evidence of multicollinearity. So we created a correlation matrix and scanned the values to see if there was anything close enough to 1 or -1 to be considered evidence of multicollinearity, but there was nothing close enough to prove this so we state that there is no multicollinearity in this dataset.

```
##                Age     Height     Weight       FCVC        NCP        CH2O
## Age    1.00000000 0.02080862 0.21793935 0.02225751 0.04393122 0.034840229
## Height 0.02080862 1.00000000 0.47117782 0.04608912 0.25248988 0.204148356
## Weight 0.21793935 0.47117782 1.00000000 0.19534515 0.11347087 0.190483184
## FCVC   0.02225751 0.04608912 0.19534515 1.00000000 0.05008597 0.052498525
## NCP    0.04393122 0.25248988 0.11347087 0.05008597 1.00000000 0.042227619
## CH2O   0.03484023 0.20414836 0.19048318 0.05249853 0.04222762 1.000000000
## FAF    0.16415271 0.30640582 0.05650405 0.01100181 0.11328858 0.161417486
## TUE    0.30123390 0.05600413 0.06577166 0.10525079 0.04894118 0.009913777
##                FAF         TUE
## Age    0.16415271 0.301233897
## Height 0.30640582 0.056004127
## Weight 0.05650405 0.065771664
## FCVC   0.01100181 0.105250787
## NCP    0.11328858 0.048941181
## CH2O   0.16141749 0.009913777
## FAF    1.00000000 0.068454890
## TUE    0.06845489 1.000000000
```

Classification

Results

// **TODO** - Restate results from the methods section, talk about what they mean and why they happened

After performing all of our regression testing, we saw that the lowest RMSE values we were able to obtain through various methods of regression was always around 8. There was no method that we could find that would allow us to narrow the model even further in order to get a more accurate prediction of BMI.

Linear regression put our RMSE values in the range of *7.929959* to *9.08678*. Spline regression was worse and it left us with two values, those being: *8.540398* and *8.488051*. Lasso regression seemed to be the most consistent in getting us RMSE values under 8, no matter what variables were used to build the model. We resulted with RMSE of *7.968656* and *7.96817* accross 3 Lasso regression tests. And finally with the ridge regression (whose results are not included in this paper for brevity) left us with RMSE values of *8.034446* and *8.793966*.

This could be due to various reasons, one that came to mind is because of the synthetically generated data in this dataset. That generated data could have lead to inaccurate inputs which would hold us back from being able to accurately predict BMI of a person. Although the data generated was useful for filling in noticeable gaps, it has most likely lead to some sort of skew and/or bias in the dataset which prevents us from being able to get a more accurate prediction.

Conclusions

//*__TODO__ - Essay type write up for the conclusion

To conclude, blah blah blah i dont think i can conclude yet because were not done.

References

1. Estimation of Obesity Levels Based On Eating Habits and Physical Condition
2. An overview of weight and height measurements on World Obesity Day

Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE, tidy.opts = list(width.cutoff = 9
    tidy = TRUE)
library(kableExtra)
# df = data.frame('Obesity Level' = c('Insufficient Weight', 'Normal Weight', 'Overweight
# Level I', 'Overweight Level II', 'Obesity Type I', 'Obesity Type II', 'Obesity Type
# III'), 'BMI' = c('<18.5', '18.5-22.9', '23.0-25.9', '26.0-29.9', '30.0-34.9',
# '35.0-39.9', '>40')) kable(df, caption = '\\label{tab1}Categorizing Obesity Levels')
# %>% column_spec(2:2,border_left = T)

mat = matrix(c("BMI", "<18.5", "18.5-22.9", "23.0-25.9", "26.0-29.9", "30.0-34.9", "35.0-39.9",
    ">40"), nrow = 1, ncol = 8, byrow = TRUE)

colnames(mat) = c("Obesity Level", "Insufficient Weight", "Normal Weight", "Overweight Level I",
    "Overweight Level II", "Obesity Type I", "Obesity Type II", "Obesity Type III")
kable(mat, align = "c", caption = "\\label{tab1}Categorizing Obesity Levels")

# Setup Code Chunk Add any libraries etc used here.
library(ggplot2)
library(tidyverse)
library(splines)
library(glmnet)

# Data importing
dataset = read.csv("ObesityDataSet_raw_and_data_sinthetic.csv")

# factor all categorical variables
dataset[sapply(dataset, is.character)] = lapply(dataset[sapply(dataset, is.character)], as.factor)

set.seed(1)
row.number = sample(1:nrow(dataset), 0.7 * nrow(dataset))
trainData = dataset[row.number, ]
testData = dataset[-row.number, ]

# BMI to use
BMI = trainData$Weight/(trainData$Height^2)
# Performing regression on every variable with BMI

# Gender
m = lm(BMI ~ Gender, data = trainData)
p = predict(m, newdata = testData)
rmse1 = sqrt(mean((BMI - p)^2))

# Age
m = lm(BMI ~ Age, data = trainData)
p = predict(m, newdata = testData)
```

```r
rmse2 = sqrt(mean((BMI - p)^2))

# family_history_with_overweight
m = lm(BMI ~ family_history_with_overweight, data = trainData)
p = predict(m, newdata = testData)
rmse3 = sqrt(mean((BMI - p)^2))

# FAVC - Frequent consumption of high caloric food
m = lm(BMI ~ FAVC, data = trainData)
p = predict(m, newdata = testData)
rmse4 = sqrt(mean((BMI - p)^2))

# FCVC - Frequent consumption of vegetables
m = lm(BMI ~ FCVC, data = trainData)
p = predict(m, newdata = testData)
rmse5 = sqrt(mean((BMI - p)^2))

# NCP - Number of main meals
m = lm(BMI ~ NCP, data = trainData)
p = predict(m, newdata = testData)
rmse6 = sqrt(mean((BMI - p)^2))

# CAEC - Consumption of food between meals
m = lm(BMI ~ CAEC, data = trainData)
p = predict(m, newdata = testData)
rmse7 = sqrt(mean((BMI - p)^2))

# SMOKE
m = lm(BMI ~ SMOKE, data = trainData)
p = predict(m, newdata = testData)
rmse8 = sqrt(mean((BMI - p)^2))

# CH2O - Consumption of water
m = lm(BMI ~ CH2O, data = trainData)
p = predict(m, newdata = testData)
rmse9 = sqrt(mean((BMI - p)^2))

# SCC - Calories consumption monitoring
m = lm(BMI ~ SCC, data = trainData)
p = predict(m, newdata = testData)
rmse10 = sqrt(mean((BMI - p)^2))

# FAF - Frequency of physical activity
m = lm(BMI ~ FAF, data = trainData)
p = predict(m, newdata = testData)
rmse11 = sqrt(mean((BMI - p)^2))

# TUE - Time using electronics
m = lm(BMI ~ TUE, data = trainData)
p = predict(m, newdata = testData)
rmse12 = sqrt(mean((BMI - p)^2))

# CALC - Consumption of alcohol
```

```r
m = lm(BMI ~ CALC, data = trainData)
p = predict(m, newdata = testData)
rmse13 = sqrt(mean((BMI - p)^2))

# MTRANS - Transportation used
m = lm(BMI ~ MTRANS, data = trainData)
p = predict(m, newdata = testData)
rmse14 = sqrt(mean((BMI - p)^2))

cat("Gender =", rmse1, "\nAge =", rmse2, "\nHistory of overweight =", rmse3, "\nFreq consumption of high
    rmse4, "\nFreq consumption of vegetables =", rmse5, "\nNumber of main meals =", rmse6,
    "\nConsumption of food between meals =", rmse7, "\nSmoke =", rmse8, "\nConsumption of water =",
    rmse9, "\nCalorie consumption monitoring =", rmse10, "\nFreq of physical activity =", rmse11,
    "\nTime using electronics =", rmse12, "\nConsumption of alcohol =", rmse13, "\nTransportation used =
    rmse14)
# Lets graph the regression models

# Gender
m = lm(BMI ~ poly(Gender, 1), data = trainData)
trainData |>
    mutate(fitted = fitted(m)) |>
    ggplot(aes(x = Gender)) + geom_point(aes(y = BMI), size = 1) + ggtitle("Relationship between Gender

# NCP
m = lm(BMI ~ poly(NCP, 4), data = trainData)
trainData |>
    mutate(fitted = fitted(m)) |>
    ggplot(aes(x = NCP)) + geom_point(aes(y = BMI), size = 1) + geom_line(aes(y = fitted),
    colour = "red", linewidth = 1) + ggtitle("Relationship between NCP and BMI")

# SMOKE
m = lm(BMI ~ poly(SMOKE, 1), data = trainData)
trainData |>
    mutate(fitted = fitted(m)) |>
    ggplot(aes(x = SMOKE)) + geom_point(aes(y = BMI), size = 1) + ggtitle("Relationship between SMOKE an

# TUE
m = lm(BMI ~ poly(TUE, 4), data = trainData)
trainData |>
    mutate(fitted = fitted(m)) |>
    ggplot(aes(x = TUE)) + geom_point(aes(y = BMI), size = 1) + geom_line(aes(y = fitted),
    colour = "red", linewidth = 1) + ggtitle("Relationship between TUE and BMI")
# Spline regression on the numerical variables to see if its a better approach

# NCP
m = lm(BMI ~ ns(NCP, df = 5), data = trainData)
p = predict(m, newdata = testData)
rmse1 = sqrt(mean((BMI - p)^2))

# TUE
m = lm(BMI ~ ns(TUE, df = 5), data = trainData)
p = predict(m, newdata = testData)
rmse2 = sqrt(mean((BMI - p)^2))
```

```r
cat("Number of main meals =", rmse1, "\nTime using electronics =", rmse2)
# Lasso regression with all the lower error variables

# Create model
fit = lm(BMI ~ Gender + NCP + SMOKE + TUE, data = trainData)
X = model.matrix(fit)
y = BMI

# Remove intercept
X = X[, -1]
fit_lasso = glmnet(X, y)
X_test = model.matrix(~Gender + NCP + SMOKE + TUE, data = testData)
X_test = as.matrix(X_test)
# Remove intercept
X_test = X_test[, -1]
y_test = BMI[1:634]
predLasso = predict(fit_lasso, newx = X_test, s = 1.2)
rmseLasso = sqrt(mean((y_test - predLasso)^2))
cat("RMSE =", rmseLasso)
# Lasso regression all variables (no weight and height)

# Create model
fit = lm(BMI ~ Gender + Age + family_history_with_overweight + FAVC + FCVC + NCP + CAEC + SMOKE +
    CH2O + SCC + FAF + TUE + CALC + MTRANS, data = trainData)
X = model.matrix(fit)
y = BMI

# Remove intercept
X = X[, -1]
fit_lasso = glmnet(X, y)
X_test = model.matrix(~Gender + Age + family_history_with_overweight + FAVC + FCVC + NCP +
    CAEC + SMOKE + CH2O + SCC + FAF + TUE + CALC + MTRANS, data = testData)
X_test = as.matrix(X_test)
# Remove intercept
X_test = X_test[, -1]
y_test = BMI[1:634]
predLasso = predict(fit_lasso, newx = X_test, s = 4)
rmseLasso = sqrt(mean((y_test - predLasso)^2))
cat("RMSE =", rmseLasso)
# Lasso regression all variables (no weight and height)

# Create model
fit = lm(BMI ~ Gender + SMOKE, data = trainData)
X = model.matrix(fit)
y = BMI

# Remove intercept
X = X[, -1]
fit_lasso = glmnet(X, y)
X_test = model.matrix(~Gender + SMOKE, data = testData)
X_test = as.matrix(X_test)
# Remove intercept
X_test = X_test[, -1]
```

```r
y_test = BMI[1:634]
predLasso = predict(fit_lasso, newx = X_test, s = 0.27)
rmseLasso = sqrt(mean((y_test - predLasso)^2))
cat("RMSE =", rmseLasso)
cat("Minimum =", min(BMI))
cat("Maximum =", max(BMI))
# Checking for multicollinearity
double_vars = sapply(dataset, function(x) is.numeric(x) && !all(x%%1 == 0))
double_data = trainData[, double_vars]
corMatrix = cor(double_data)
abs(corMatrix)
```