Data 2010 Project

Lucas Berzuk, Ethan Robson, Hugo Ng, Rodell Salonga

April 5, 2024

Introduction

- *Talk about the data analysis and probably a few other things, maybe like where we were gojng to go with it but then what changed*

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 model with the target variable being BMI...

We then build classification models including...

- Talk about our data and preprocessing stuff

Methods

Regression

For regression 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.

1. *Gender* = 7.939414
2. *Age* = 8.178174
3. *History of overweight* = 9.08678
4. *Freq consumption of high caloric food* = 8.134233
5. *Freq consumption of vegetables* = 8.167971
6. *Number of main meals* = 7.932018
7. *Consumption of food between meals* = 8.807099
8. *Smoke* = 7.929959
9. *Consumption of water* = 8.065398
10. *Calorie consumption monitoring* = 8.106003
11. *Freq of physical activity* = 8.109581
12. *Time using electronics* = 7.973657

13. *Consumption of alcohol* = 8.130043
14. *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 do spline regression on the 2 numerical variables out of these 4 to see if that would improve the model any further. These were the results:

1. *Number of main meals* = 8.540398
2. *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 models. When performing lasso regression on the 4 original variables that all had RMSE under 8, this is what we got with a lambda of 2.

Lasso regression = 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:

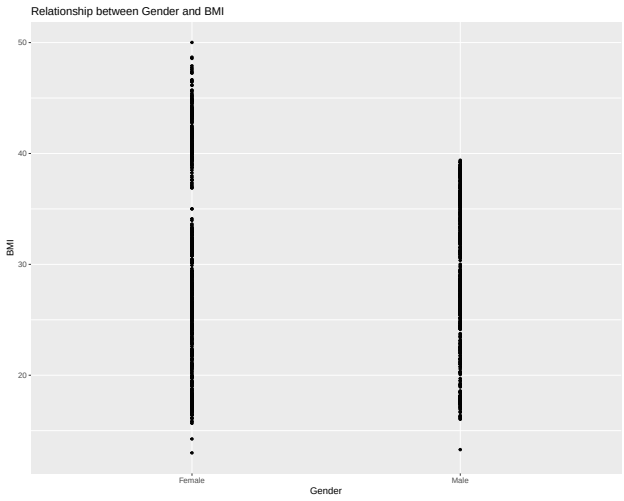
Lasso regression = 7.968656

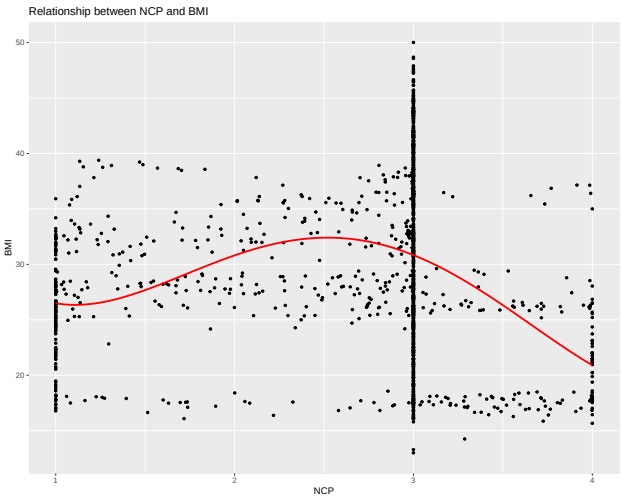
Which infact, is worse than before but this is something we kind of expected because more variables just means more noise. Maybe we could try and use less variables? This is with just the Gender and SMOKE variable at the lambda of 0.27.

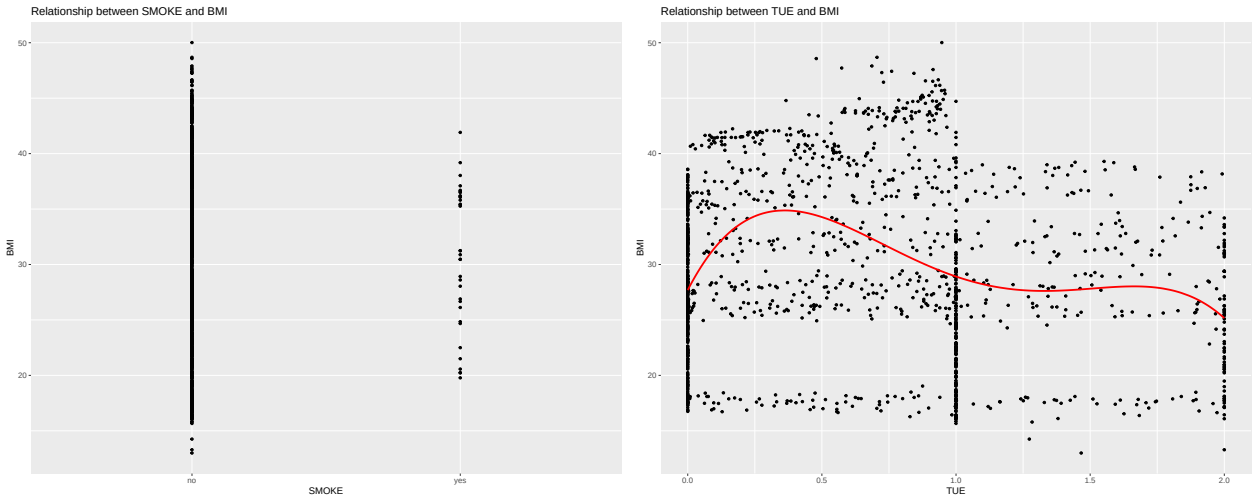
Lasso regression = 7.96817

The result is just slightly lower than the previous 2 lasso regression attempts so it seems like we may be at a deal end here.

Here are the graphs of the cubic regression models for the 4 original variables that we were trying to improve.







Next we wanted to try to see if there was any evidence of multicollinearity. Upon taking a look at the correlation matrix we did not see any values that were even close to being considered multicollinearity.

Classification

Results

- Just the results from all of the shit we do to the data

Conclusions

- Essay type write up for the conclusion

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