STAT 51200: Final Project

Application of Modeling Techniques for Predicting Body Fat Content

Ireland Straub, Prantar Borah, Sanaz Matinmehr

Abstract

The study aimed to build a **predictive model** for **visceral adipose tissue (VAT)** mass using data from the **2017-2018 NHANES [2]** questionnaire. Variables like age, gender, dietary habits, sleep patterns, alcohol consumption, and anthropometric measures were explored to understand their impact on visceral fat. Model selection techniques identified a **7-predictor model** as optimal. Surprisingly, **dietary** variables **did not significantly affect** the response, leading to their exclusion despite initial expectations. This research underscores the importance of **diverse predictor variables** in **understanding visceral fat** and highlights the necessity of addressing model assumptions for robust predictions.

Introduction

Prediction models have been considered as essential over the years, especially within the field of medical data research. They can be created to predict or find trends amongst diseases, patient diagnoses, or new discoveries in the medical field. The Center for Disease Control considers prediction models to be incredibly useful for precise patient diagnostics and if they require **further testing [3]**. Physicians may use these models to predict physiological aspects of the patient, such as weight to see correlations between it and if they are susceptible to a possible illness.

Body Mass Index (BMI) is the most common measurement used to diagnose obesity. However, it has been well documented that excess visceral body fat is a stronger predictor of obesity and related diseases**,** such as **heart disease and diabetes** [1]. Our goal was to create a model to predict visceral fat mass based on different types of components, like BMI. The following study was based on the dataset retrieved from NHANES from 2017 to 2018. We explored different types of variables that would influence the response and yield the best predictive model. These included aspects such as age and sex, various body measurements (i.e. BMI), sleeping habits, and dietary habits. Initially, we believe that the variables related to diet would be the biggest factors that influence VAT mass. We provide further evidence that tests our claim and we reported the findings in the following sections.

Model Description

In order to create a good model for **VAT**, we explored variables from several different datasets and assessed their relationship with the response variable. These datasets are provided from the 2017-2018 NHANES questionnaire. We hypothesized the following variables would have an effect on the response: **age** (only participants aged **0 to 79** were used), **gender**, **number of frozen meals eaten in the last 30 days, number of fast food meals eaten in the last 30 days, number of meals eaten in the last 30 days that were not home prepared, average number of alcoholic drinks per day for the last 12 months, whether or not they were self reported to snore frequently, average number of hours slept per night on workdays, weight** (in kg)**, BMI, and waist circumference** (in centimeters). After selecting only variables of interest, removing rows which contained unusable observations, and combining each set by participant, we were left with a final dataset which contained **1711** observations of **12** variables. In order to test the model, the last **200** observations were removed from the model building data, which left us with **1511** observations.

Different model selection procedures were employed in order to compare several potential models. A model which contained 7 predictors was determined to be the best by several criteria: namely

SSE, adjusted R^2, and Mallow's Cp. The BIC criteria was minimized for a model containing 6 predictors. Thus, we considered **two possible models** chosen by a best subsets search procedure; one which contained **age, gender, sleep hours, snoring frequently, weight, waist circumference, and BMI**, and one which contained **all of the previous except for BMI**. Surprisingly, **none of the variables** related to **diet** were significant enough in reducing the variation in the response by any criteria to be chosen through any selection procedure. Although it was slightly more complex, **we decided to continue the analysis with the model containing 7 predictors.**

The next step was assessing whether the model assumptions were met, i.e. normally distributed residuals, constant variance amongst the residuals, and that the residuals were independent from the response. The **normal probability plot** of the residuals **showed significant departure**s of normality at the tails. The plot of **residuals versus fitted** values **exhibited a megaphone shape**, becoming more spread out for larger fitted values. The plot of **absolute residuals versus fitted** values **showed a slight linear relationship** between the errors and fitted values. Therefore, we determined that a **transformation on the response variable would be necessary** in order to stabilize the variance.

Fig.1_Fitted VAT vs Residual VAT

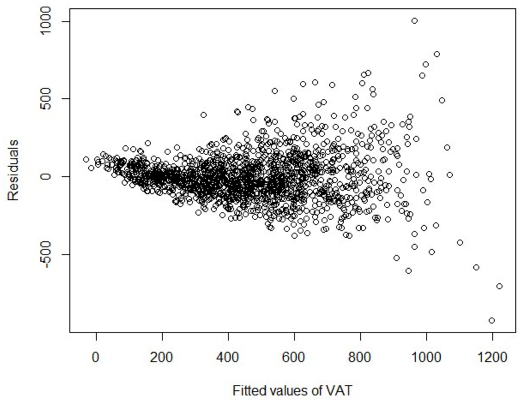


Fig.2_Absolute VAT vs Fitted VAT

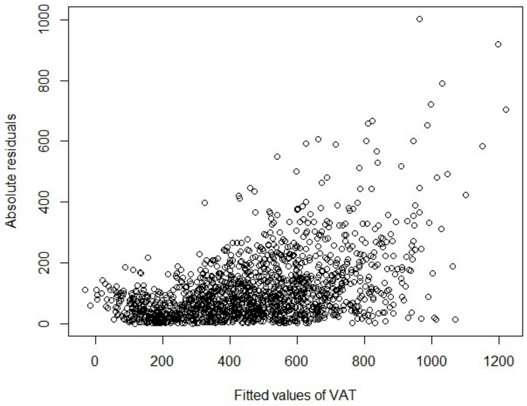
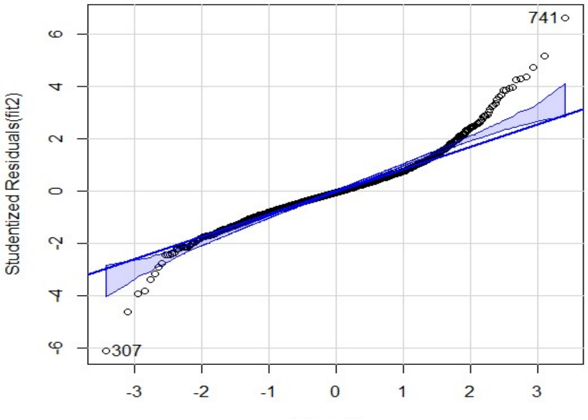


Fig.3_QQ plot of Studentized Residuals for Fit2 Model



The **box-cox procedure** was used to automatically identify **λ=0.26** as the optimal transformation on Y, and a new model was fitted to the transformed Y. This transformation both lowered the values of the estimated regression coefficients and their standard errors. It was also effective in stabilizing the variance, as supported by the new residual plots. There are still slight departures from normality at the bottom tail of the QQ plot, but it has **improved substantially**.

Fig.4_Residuals vs Fitted Y

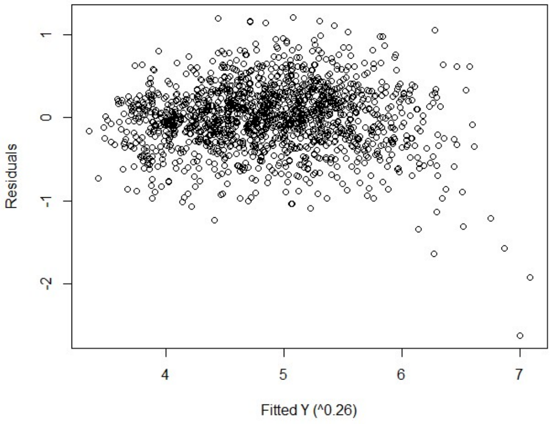


Fig_5_Absolute Residuals vs. Fitted Y

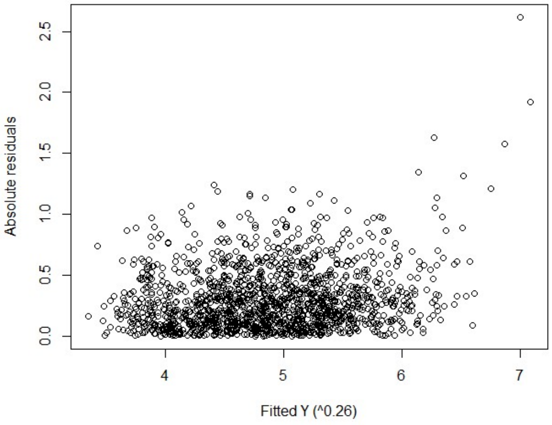
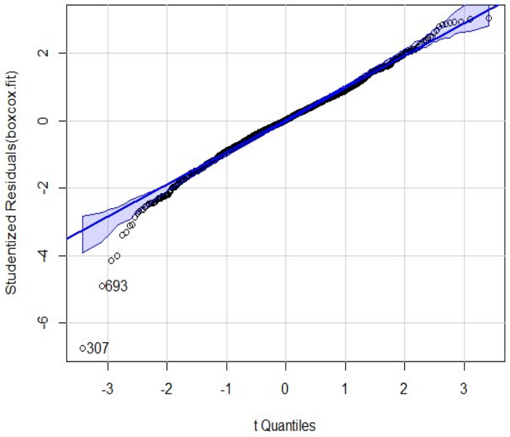


Fig.6_QQ plot of Box-Cox Transformation



From the nature of the chosen predictors, we **expected** there to be a **moderate problem with collinearity**. We looked at the variance inflation factors for the different variables, and found it was concerningly high for the weight, BMI, and waist circumference variables. We opted to **utilize ridge regression**, as it could also help reduce the influence of some extreme observations. The new coefficients determined from this method were very slightly smaller than the coefficients from the box cox fit. Since the two models were extremely similar, we decided to take **both models into consideration** for the final choice. The **adjusted** $\mathbf{R^2}$ for the **box-cox** only model was **0.726**, and the **adjusted** $\mathbf{R^2}$ for the **ridge regression** model was **0.725**.

Both models were tested with the 200 observations in the validation dataset. The **box-cox** fit resulted in a **MSPR** of **0.1396**, and the box-cox and **ridge** model resulted in a **MSPR** of **0.1397**. Clearly, both models are equally sufficient for prediction. We decided to **settle on the more simplistic box-cox** only model, since the **multicollinearity did not seem to greatly impact** the regression coefficients or predictive ability of the model.

Statistical Analysis

The **final model** is as follows:

$$\boldsymbol{\widehat{Y}'_i = 0.967 + 0.018X_{i1} - 0.228X_{i2} + 0.020X_{i3} + 0.071X_{i4} - 0.012X_{i5} + 0.021X_{i6} + 0.036X_{i7}}$$

Where Y' is visceral adipose tissue mass (to the power of 0.26), X_1 is age, X_2 is sex with 1 representing females and 0 for males, X_3 is average number of hours slept per workday, X_4 is heavy snoring, X_5 is weight (in kg), X_6 is BMI, and X_7 is waist circumference (in cm). The model can be interpreted as the mean change in VAT mass (to the power of 0.26) per one unit change in any predictor variable, given that the remaining variables are held constant.

Summary output for the model:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.61522 -0.23216  0.00652  0.25097  1.20023


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9669892  0.0903454  10.703  < 2e-16 ***
AGE          0.0181114  0.0009418  19.232  < 2e-16 ***
FEMALE      -0.2280191  0.0269369  -8.465  < 2e-16 ***
SLEEP        0.0198645  0.0065730   3.022  0.00255 **
SNORE        0.0705331  0.0249811   2.823  0.00481 **
WEIGHT      -0.0116352  0.0015321  -7.594 5.40e-14 ***
BMI          0.0207533  0.0048119   4.313 1.72e-05 ***
WAIST        0.0363455  0.0020224  17.971  < 2e-16 ***
```

All regression coefficients have a **small standard error**. All predictor variables are statistically significant, with the **t-values** ranging from **3.02** to **19.23**. This is convincing evidence that all of the **chosen parameters should be in the model**.

Output continued:

```
Residual standard error: 0.3981 on 1503 degrees of freedom
Multiple R-squared:  0.7272,   Adjusted R-squared:  0.726
F-statistic: 572.4 on 7 and 1503 DF,  p-value: < 2.2e-16
```

The **F-test** for significance produced a value of **572.4** with a p-value of zero. We can conclude that the regression model is **statistically significant**.

Conclusion

We conducted a thorough analysis of a set of variables that we predicted would affect visceral adipose tissue mass. These included **demographic information, sleeping patterns, alcohol use, eating habits, and various body measurements**. We **initially predicted** that **all** of these variables **may be useful** in predicting visceral adipose tissue mass. Through several variable selection procedures, we found that snoring, average number of hours sleeping on weekdays, gender, age, weight, BMI, and waist circumference were statistically significant in predicting VAT mass. To our **surprise**, **the amount of fast food, frozen meals, and non-home cooked meals eaten did not appear** to be **correlated** to the response. This could possibly be due to the nature of the data, which only asked participants about their eating habits of the previous 30 days. **We can conclude** that there is a relationship between **the amount of visceral fat** and **whether or not a person snores**, as well as **the amount of sleep they get on average**. **Gender, age, weight, BMI, and waist circumference** are also related to **visceral fat**.

If given more time, we would like to explore other predictor variables that could improve the model. The regression coefficient for weight had the opposite algebraic sign of what we would expect, so it is obvious that some important predictors have been omitted from the model. It would be interesting to examine factors such as mental health, prescription medications, and dietary supplements. Additionally, other variables such as environmental and socioeconomic components could be tested as well.

Worked Cited

[1] Després, J. (2011). Excess visceral adipose Tissue/Ectopic fat. *Journal of the American College of Cardiology*, *57*(19), 1887–1889. https://doi.org/10.1016/j.jacc.2010.10.063

[2] "NHANES Questionnaires, Datasets, and Related Documentation." *Wwwn.cdc.gov*, wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017.

[3] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T. A., Gönen, M., Obuchowski, N. A., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/ede.0b013e3181c30fb2