Kyndalanne Pike, Shengjia Kang, Tam Nguyen

Dr. Frazier

Applied Statistical Methods

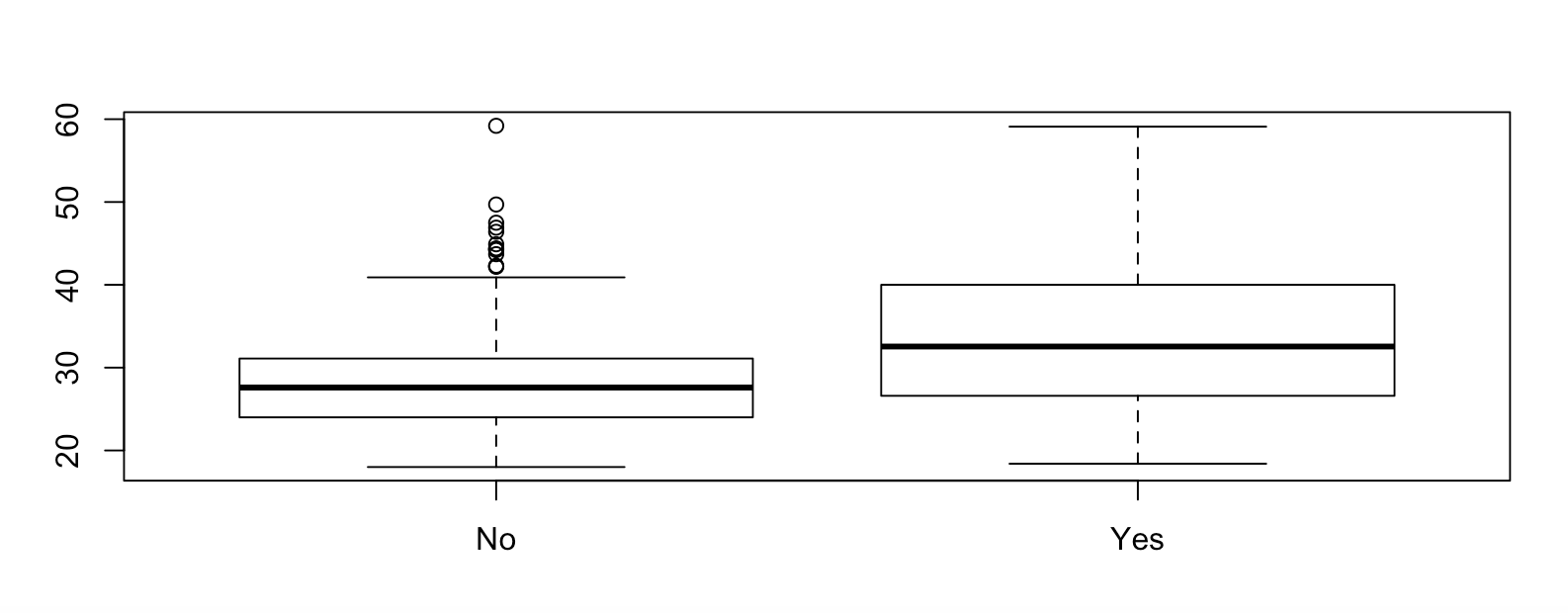
29 September 2018

**NHANES Modeling BMI**

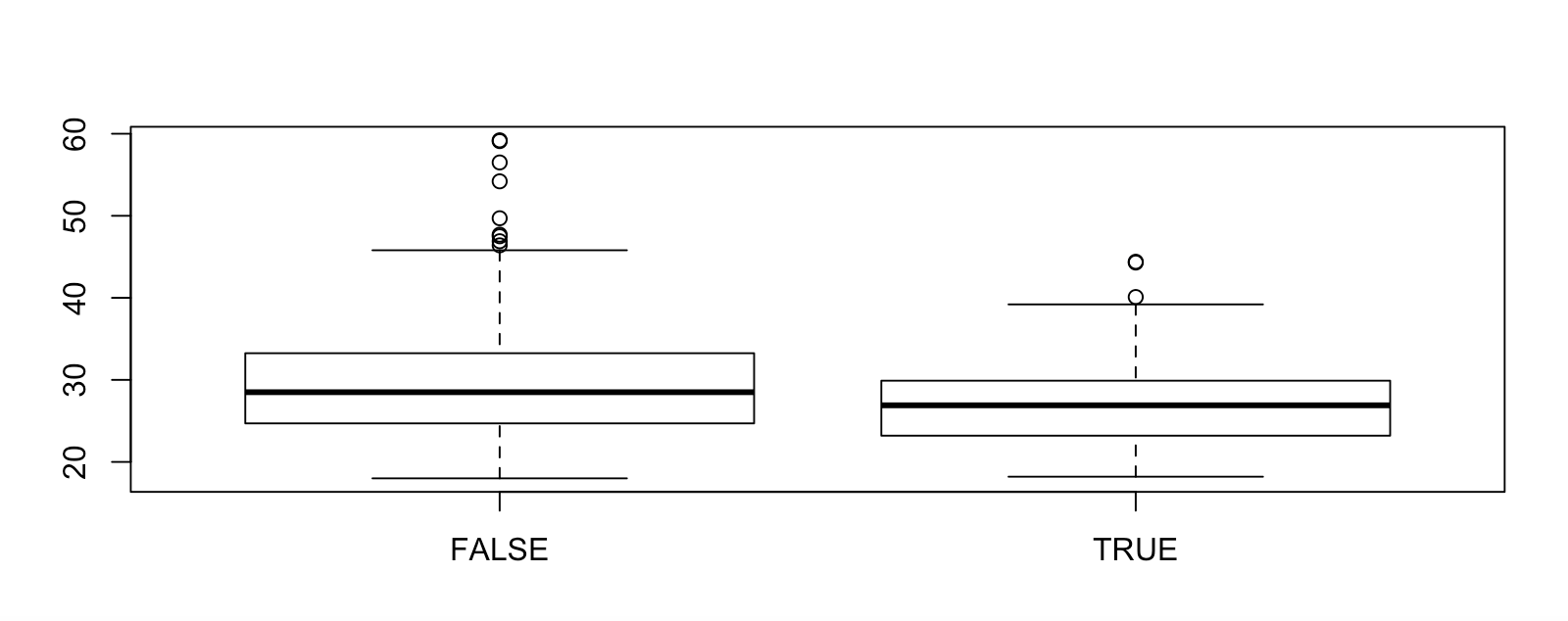
For our model, we used the NHANES dataset, which is considered to be fairly representative of US civilians. In this dataset, 48 terms were documented for 500 individuals. In these 48 terms, 20 of them are categorical/ordinal and 28 are numerical. Independence was upheld as one person’s BMI and other indicators should not affect others. The age interval was from 30 to 80, with a mean of 53. This data set is not representative of children or young adults. In this study, 51% of the subjects are male and 49% are female. Among 500 subjects, 63% of them are white, 12% are black, and 25% are other races. Also, there are 76 subjects (15%) who have diabetes (32 females and 44 males). 43% of the subjects in this study are smokers, and 15% of the subjects are drug users. According to the distribution of age, race, and gender of US citizen on the report of the United States Census Bureau, the contribution of factors in NHANES dataset has a slightly different at approximately 3%, which is not really significant. Consequently, we believe that this dataset can be a representative dataset for US civilians and useful in further investigation.

BMI is a function of weight and height. In this study, we will dig deeper to see if there are other factors that are related. In our model, we try to predict log(BMI) instead of BMI to make highly skewed distributions of BMI less skewed. Firstly, we looked at some categorical/ordinal terms and see if the distribution of BMI is significantly different. Some intriguing factors are: Gender, Race, TV hours, and Diabetes.

In this part, we now present some major relationship between BMI and various parameters such as gender, race, TV hours and diabetes. Although the distribution for female is more spread out, we do not think there are any major difference between the two genders. We believe that gender is an indicator for testosterone, since the amount of testosterone will determine the gender of a person. The distribution of BMI does look different for different races. Especially between White/Black, Mexican/Hispanic, and Others, the differences seem significant. However, 63% of the sample are white, it is questionable if we should include race as an indicator due to complexity. Then we looked at how TV hours may be related to BMI, Figure 1. We use Yes value to represent people who watches TV less than 2 hours per day, and No value for people who watches TV more than or equal to 2 hours per day.



***Figure 1:*** *Distribution of indicator watching TV less than 2 hours per day*

From the Figure 1, we can say that the distribution of BMI is different for people who watch TV less than 2 hours a day and people who watch more than 2 hours a day. Diabetes is another reasonable factor for predicting BMI. Using binary value to shows the distribution of diabetes, Figure 2 represents the different between the density of people who has diabetes and people who don’t. 

***Figure 2:****BMI distributions separated by whether or not individuals had diabetes.*

After examining indicators, we then moved to numerical terms. Table 1 is the correlation matrix between variables that we considered could be predictors for natural log of BMI.

***Table 1:*** *Correlation between BMI and numerical variables*

|  |  |
| --- | --- |
| **Numerical Variables** | **Correlation** |
| **Log(BMI)** | 1.0000 |
| **Age** | 0.0097 |
| **Pulse\*** | 0.1601 |
| **BPSysAve\*** | 0.1415 |
| **BPDiaAve\*** | 0.1246 |
| **Testosterone\*** | -0.1037 |
| **DirectChol\*\*\*** | -0.3480 |
| **Total Cholesterol** | 0.0269 |
| **UrineVol1** | 0.0247 |
| **UrineVol1** | -0.0730 |

Direct cholesterol was most highly correlated with BMI, but the relationship does not appear to be very strong. None of the variables were highly correlated with BMI.

Two models were considered. The more complex model predicted the natural log of BMI from Pulse, average systolic blood pressure and natural log of average systolic blood pressure, total testosterone, and direct HDL cholesterol. As indicators we used gender, race, whether or not the person had diabetes, and whether the person reported watching less than two hours a day of TV.

The conditions for inference were met, linearity was upheld, variance of the residuals was fairly constant, and the model was very normal according to residuals plots. This model explained approximately 30% of the variability in the data, with an adjusted R2 of 0.2986, and the ρ-value (<2.2e-16) indicated that the model was significant.

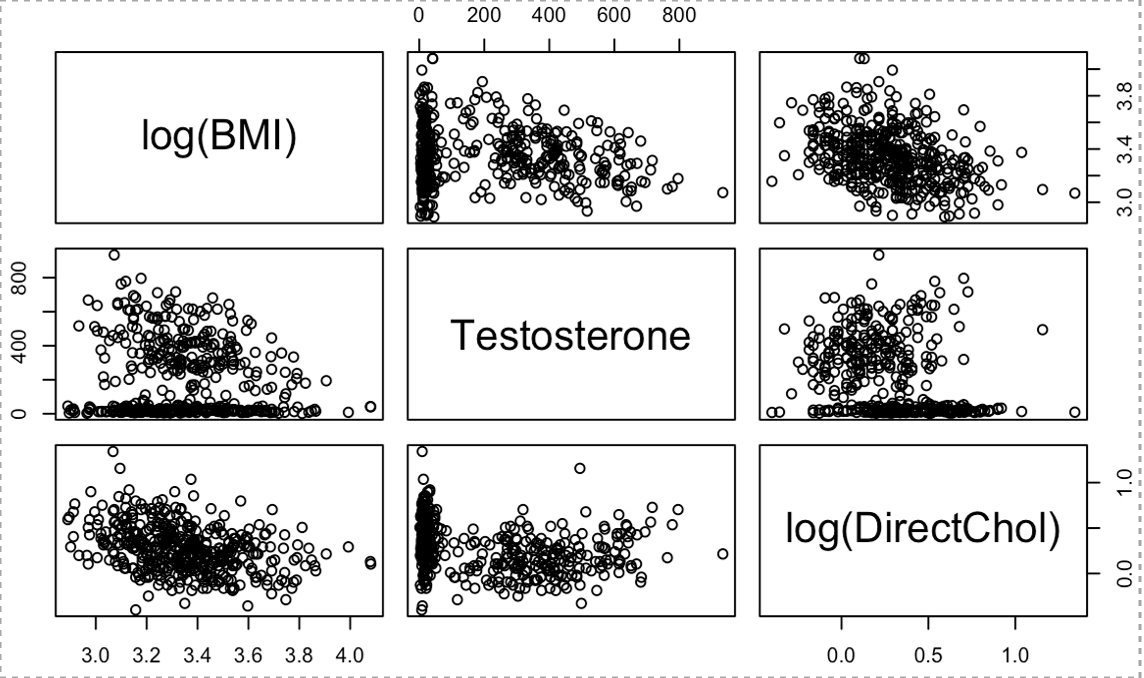
However, while looking at the VIF value of all parameters in this model in the table below, average systolic blood pressure should be excluded from our model since it has a significant high VIF value (104 > 1). Overall, even when our model has significant p-value and high R^2, this model is too complex and need to cut-off some parameters.

**Table 2:** VIF values of the complex model

|  |  |
| --- | --- |
| **Variables** | **GVIF** |
| **Pulse\*\*\*** | 1.0579 |
| **Avg systolic blood pressure**  **(BPSysAve)** | 104.2231 |
| **log(BPSysAve)** | 104.5278 |
| **Testosterone** | 3.9371 |
| **Direct Cholesterol\*\*\*** | 1.2300 |
| **Gender** | 4.2976 |
| **Race\*\*** | 1.1118 |
| **TV hours\*\*\*** | 1.0677 |
| **Diabetes\*\*** | 1.1238 |

The figure below shows the relationship between log(BMI) vs log(DirectChol) and log(BMI) vs Testosterone. Those are two significant numerical variables which doesn’t have collinearly relationships with the natural log of BMI. We didn’t examine Pulse in the simpler model, since pulse is believed to have an association with the amount of direct cholesterol in the body. We then can see the trend of linear relationship among these variables in the figure below.

**Table 3:** Matrix of scatter plots for numerical variables used in final model: testosterone and natural log of direct cholesterol against natural log BMI



In the second model, we exclude all factors that do not show the significant relationship to the natural log of BMI to have a simpler model. This new model predicted the natural log of BMI from Testosterone, and natural log of Direct HDL Cholesterol in mmol/L. As indicators, we included whether or not the person reported watching TV for less than 2 hours per day, and whether or not the person has diabetes.

**FINAL MODEL:**

***ln(BMI)*** = 3.467 - 0.0001***Testosteron***e – 0.3049***ln(DirectChol)*** – 0.0672***TVHrsDay*** + *0.177****Diabete****s*

***Table 4:*** *Coefficients, p-values, and confidence intervals for simpler model*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Coefficient** | **p-value** | **Confidence Interval** |
| **Intercept** | 3.467 | < 2e-16 | [3.4323, 3.5019] |
| **Testosterone** | -1.912e-04 | 3.43e-16 | [-0.0002, -0.0001] |
| **log(DirectChol)** | -3.049e-01 | < 2e-16 | [-0.3718, -0.2380] |
| **TV hours** | -6.724e-02 | 0.000542 | [-0.1051, -0.0293] |
| **Diabetes** | 1.177e-01 | 1.21e-06 | [0.0706, 0.1646] |

In the second model, we exclude all factors that do not show the significant relationship to the natural log of BMI to have a simpler model. This new model predicted the natural log of BMI from Testosterone, and natural log of Direct HDL Cholesterol in mmol/L. As indicators, we included whether or not the person reported watching TV for less than 2 hours per day, and whether or not the person has diabetes.

According to residuals plots, the conditions for inference were met, linearity was upheld, the variance of the residuals was constant, and the model was very normal. This model explained approximately 23% of the variability in the data, with an adjusted R2 of 0.23, F-stat equals 35.58 on 4 and 459 DP, and the ρ-value (< 2.2e-16) ~ 0, indicated that the model was significant.

In this model, all factors have p-value ~ 0, which shows the significant relationship between the natural log of BMI and the all the variables. The relationship between the natural log of BMI and the indicator factor whether that person watched TV for less than 2 hours/day has highest p-value = 0.000542. The natural log of Direct HDL Cholesterol has the smallest p-value (< 2e-16), which means this is the most significant factor in this model.

As already mentioned above, the direct HDL Cholesterol, Testosterone, and the indicator variables such as whether that person had diabetes or not, and whether that person watched TV less than 2 hours a day have a linearly relationship with the natural log of BMI, which leads to the significant p-value in the model.

Overall, the second model predicts 95% confidence that for every unit change in total testosterone, the natural log of BMI will change decreases from 0.027% to 0.011%. An additional mmol/L of the natural log of direct HDL Cholesterol will affect the natural log of BMI will vary in range from 37.19% to 23.80%. Significantly, the indicator variable, whether that person had diabetes or not, will affect the natural log of BMI from 7.06% to 16.46%. Otherwise, with the indicator that whether that person watched TV for less than 2 hours a day or not, will slightly decrease the natural log of BMI from 10.52% to 2.93%. All these interpretations are holding all other variables constants.

This model explains 23% of the variability in the dataset, however, it shows significance in all variables. Additionally, with only two numerical variables, testosterone and the natural log of HDL Cholesterol, and the two indicators, whether that person had diabetes or not, and whether that person watched TV less than 2 hours a day, this model is simpler than the previous one and also has a significant result. In conclusion, this is our final model which meets all the requirement of a linearity model and all the variables in the model are strongly related to the BMI value.