Data analysis report

Prajwol Poudel

STA302

Introduction

This paper will explore the NHANES dataset and perform a complete data analysis on it based on the techniques explored in class. The information in the NHANES dataset is collected from the National Health and Nutrition Examination Survey in the United States and is used to determine the prevalence of major diseases and risk factors of diseases, provide national standards on measurements such as height, weight, blood pressure etc. and to assess effects of nutritional statuses on health (2017). Smoking is a habit that carries a lot of health risks and can have multiple negative effects on the health of a person. Therefore, the main purpose of this study is to observe the effect of smoking on the main outcome of interest: combined systolic blood pressure reading while delivering a model with the best variables for prediction.

Methodology

This first step is to conduct exploratory data analysis on the relationship between smoking status and combined systolic blood pressure to identify anomalies, patterns, and interesting relationships between the two groups. Using a t-test which conducts hypothesis testing we compare the means of the two groups and check if the effect of smoking status is significant or not. The t-test result showcases that there exists a relationship between smoking status and blood pressure levels as the p-value is lower than the significance level. The boxplot (Appendix 1) of smoking status showcases that the mean and median blood pressure levels of non-smoking participants is higher than their smoking counterparts. The histogram of combined systolic blood pressure levels produces a right-skewed histogram. Then a multiple linear regression model was constructed with all the variables included and stepwise selection methods as well as lasso was used for model selection. The stepwise selection process: variables are included or excluded in the model where the AIC or BIC values are calculated for each step and the model with the lowest AIC or BIC values is the final selected model. The selected model accuracy was tested with a 10-fold cross validation, calibration plots were generated for each selected model and similarly the test errors were also recorded. The final model with a parsimonious final model, the lowest test error and the most accurate calibration plot was selected. The variable of interest, smoking status of an individual is included in the final model if it was not selected as one of the significant variables during the variable selection step. The resulting final model with the lowest BIC value from the stepwise selection was chosen as it performed better in all of the important criteria's. Model diagnostics are conducted for the final model including checking for any correlation between the variables using vif, checking for leverage points or influential observations using cook's distance, hat values, dfbetas, dffits and checking if all the model assumptions are met with the use of qq plots, anova and residual plots.

Results

After the variable selection process, BIC stepwise selection produced the more parsimonious model selecting: Age, Poverty, Weight, and Sleeping trouble as the significant variables. AIC stepwise selection selected eight variables in the model while LASSO only selected one. LASSO selection produced an inaccurate calibration plot, whereas AIC stepwise selection and BIC stepwise selection produced calibration plots with similar accuracy (Figure 1). Therefore, the test error produced by the two methods was used to select the final model.

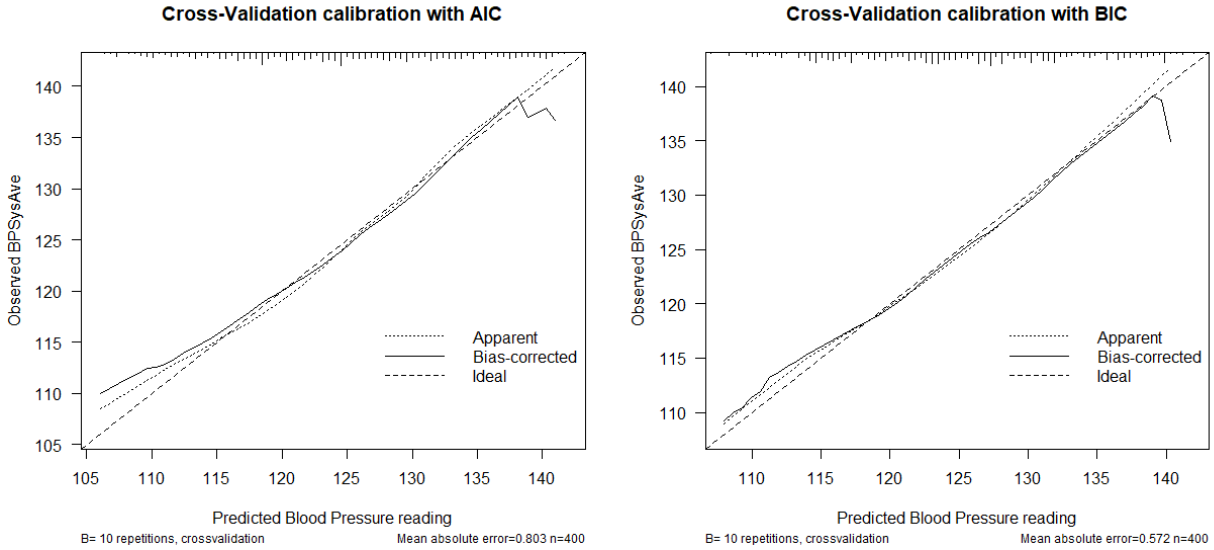


Figure 1

The test error produced by the different methods are included in the table (Figure 2). From the table, we see that LASSO selected a model with the least test error but due to model inaccuracy and inability to select more than one variable, it was eliminated. The test error of the model produced by BIC stepwise selection produced a lower test error than the AIC stepwise selection model. Therefore, the BIC model was selected for the final model as it produced a simple model, fairly accurate calibration plot and the second lowest test error. For the purposes of the study, smoking status as a variable was included in the model produced by BIC stepwise selection which is now the Final model (Figure 2).

Mean Squared Error	
AIC stepwise selection	263.1321
BIC stepwise selection	255.1556
LASSO selection	247.8412
Final model	255.2461

Figure 2(Test error)

The results of the model diagnostics concluded that the normality and homoscedasticity assumptions were not met with the usage of qq plot (Appendix 3) and residual plots (Figure 3).

To meet the model assumptions to a satisfactory level, a box-cox transformation was used in the final model, termed as the transformed model. From the residual plot of the final model, cook's distance points, dfbetas and dffits calculations, concludes that influential points are present. After conducting variance stabilizing box-cox transformation, the number of influential points is reduced as seen in the residual plot of the transformed model and the normality assumption was met.

The transformed model showcased no variable with higher than five for the vif factor, less number of influential points, outliers and less variance for the predicted values.

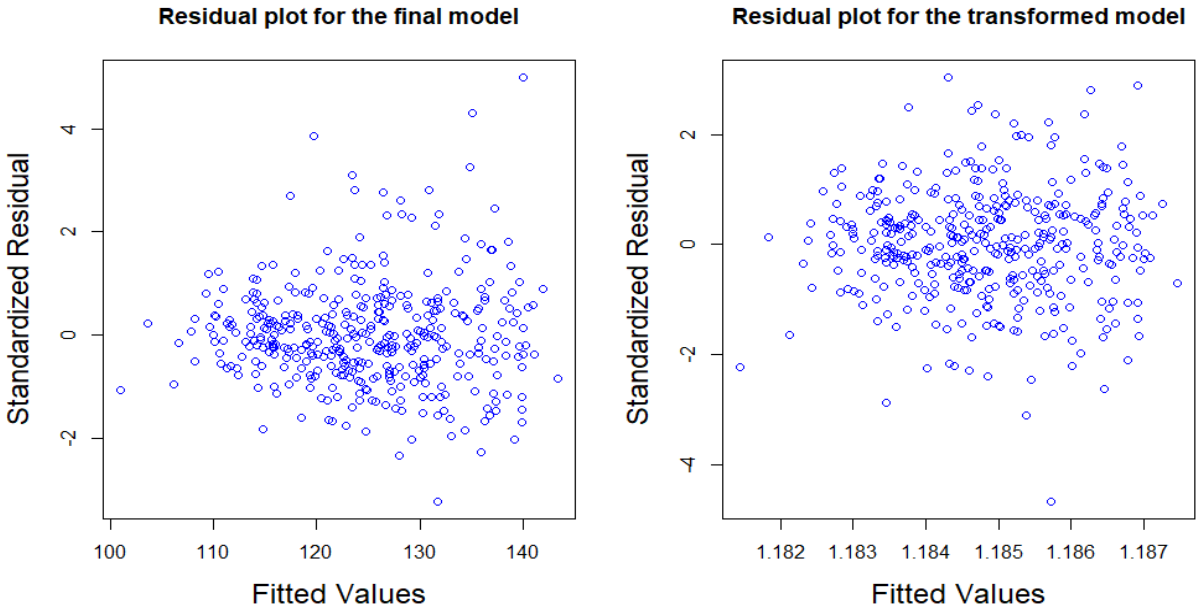


Figure 3

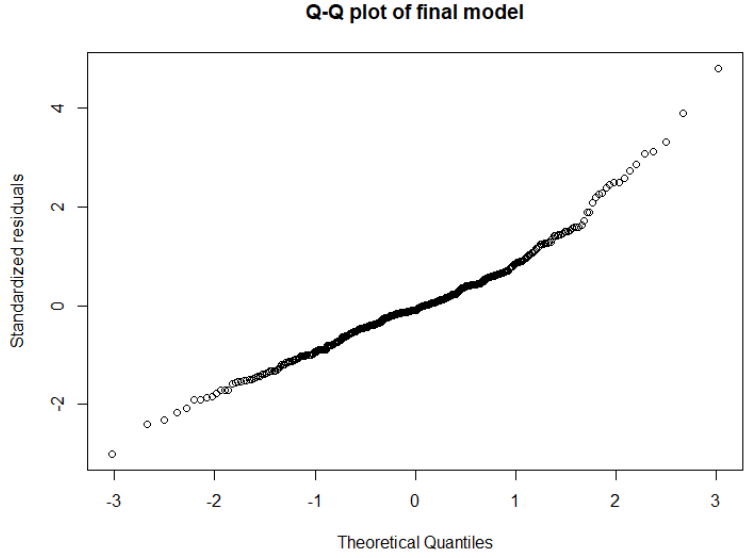
Discussion

From the summary table (Appendix 3) the effect of smoking status being yes decreases combined systolic blood pressure reading by -0.01965268052 units when all the other variables are held constant, that is all the estimates of the other variables are held to 0. This effect of smoking status is not accurate as the p-value of smoking effect is not significant in the model. The other variables selected in the final model include Age, Poverty, Sleeping trouble and Weight with their estimates shown in the summary table (Appendix 3).

The mean and median values of participants in the study who do not smoke showcases a higher blood pressure level than participants who smoke (Appendix 1). Similarly, the number of non-smoking participants is 416 compared to smoking participants 327. The higher mean and median value of non-smoking participants as well as the unequal number of participants for the two groups might be one of the causes that lead to the non-significant result of the study. Therefore, further studies need to be conducted for the effect of smoking on blood pressure levels in order to gain more accurate results than this study.

Appendix

Appendix 1



Appendix 2

```
Call:
lm(formula = ((BPSysAve^lambda - 1)/lambda) ~ Age + Poverty +
    Weight + SleepTrouble + SmokeNow, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0100991 -0.0012796  0.0000295  0.0012877  0.0065697

Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)      1.181e+00  6.926e-04 1704.607  < 2e-16 ***
Age              6.321e-05  6.848e-06    9.230  < 2e-16 ***
Poverty         -1.702e-04  7.341e-05   -2.319 0.020907 *
Weight           2.038e-05  5.927e-06    3.438 0.000648 ***
SleepTroubleYes -6.453e-04  2.360e-04   -2.734 0.006534 **
SmokeNowYes     -1.073e-04  2.426e-04   -0.442 0.658417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002235 on 394 degrees of freedom
Multiple R-squared:  0.2215,    Adjusted R-squared:  0.2116
F-statistic: 22.42 on 5 and 394 DF,  p-value: < 2.2e-16
```

Appendix 3

References

Centers for Disease Control and Prevention. (2017, September 15). *Nhanes - about the National Health and Nutrition Examination Survey*. About the National Health and Nutrition Examination Survey. Retrieved from https://www.cdc.gov/nchs/nhanes/about_nhanes.htm#intro