FINAL PROJECT

STA302

ALI KRISHT

1003376544

The National Health and Nutrition Examination Survey (NHANES) consists of a number of studies intended to evaluate the wellbeing and nourishment status of grown-ups and youngsters in the United States.

This survey inspects a broad delegate test of ~5,000 individuals for every round, with 2-year informational indexes accessible for examination. NHANES package includes study variables such as SurveyYr and ID, demographic variables such as gender, age, marital status, HH income; physical measurements like weight, length, BMI and BPsysAve. Health variables such as Direct Chol, total Chol, etc., and lifestyle variables such as physical activity, smoker or not, and others.

The original dataset has 76 variables, but we created a small data set called 'small.nhanes' that contains only 17 variables, and only those over the age of 17 were selected to be a part of the study. We then created a data set called 'train' that contains 400 observations, randomly selected from the 5000 individual's data set we had. We were interested in examining the effect of SmokeNow on the combined systolic blood pressure reading and identifying the variables that are best for predicting them.

The analysis was done by creating the model diagnostics. We first ran a multiple regression BPsysAve as the dependent variable with respect to the other 16 variables. We obtained the table shown below, highlighting the main variables and their coefficients:

	Estimate	SD. estimate	T value	p-value
Intercept	228.15320	76.30719	2.99	0.00298 **
Gendermale	4.89494	2.47180	1.980	0.04842 *
Age	0.44355	0.06406	6.924	$2e^{-11}$ ***
Poverty	-2.59380	1.09006	-2.380	0.0178 *
Weight	0.60317	0.43459	1.388	0.16603
Height	-0.77814	0.4473	-1.739	0.08280
BMI	-1.54493	1.27091	-1.216	0.22419
PhyActiveYes	-1.19446	1.87725	-0.636	0.52500
SmokeNowYes	-0.13537	1.91207	-0.071	0.94360

Fig1. The summary table of my model.

Holding all other variables constant, a unit increase in age has led to a 0.443 unit increase in BPsysAve. The p-value being $2e^{-11}$, this meant age was a statistically significant variable. However, when looking at the variable SmokeNowYes, we noticed an inverse relation between BPsysAve and SmokeNowYes. According to ceteris paribus, a one unit increase in SmokeNowYes was associated with a 0.135 decrease in BPsysAve. The p-value being large, (0.943) indicated that the variable was not significant at all. The value of $R^2 = 0.258$ meant that only about 26% of the variance for a dependent variable is explained by an independent variable in a regression model.

We then went on to test for outliers. First, we computed the QQ-plot. In Fig1 below, we noticed that the dots deviated when it reached the t = 2 quantile, and it moved away from the blue dotted line.

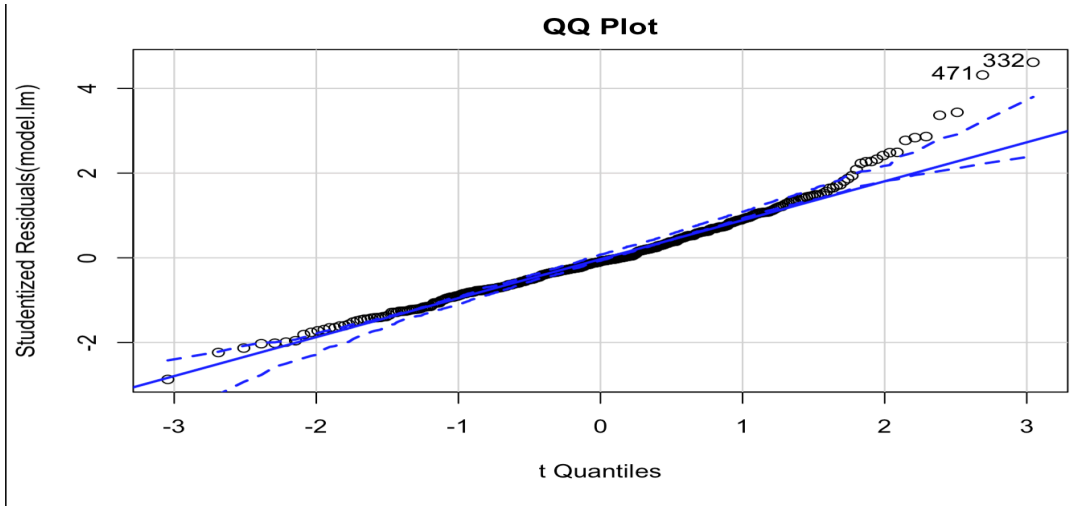
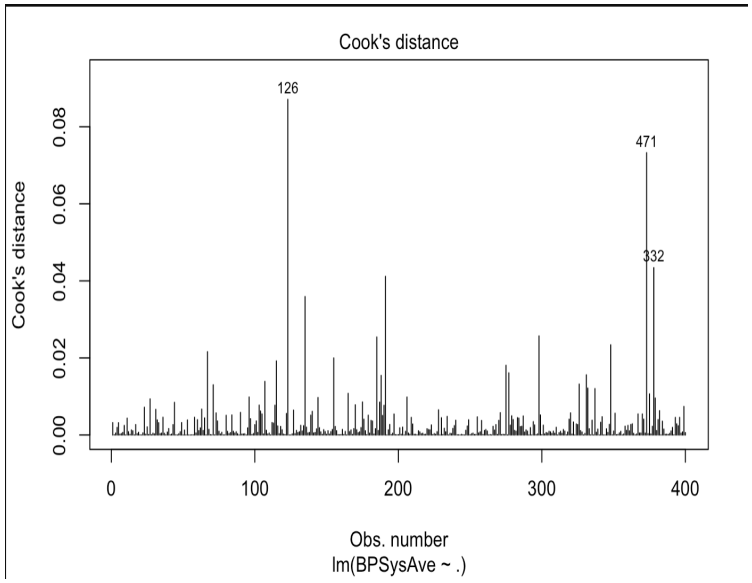


Fig2. This graph shows the relation between the standardized residual with respect to t quantile.



To confirm the result, we generate cooks distance plot (Fig.3). From this plot, we saw that there was a lot of influential points that would change the direction of the regression line (i.e. 126, 332, 471). These points were considered outliers and can be removed them from the regression to prevent the change in slope of the regression line.

Fig3. This diagram shows the relation between cooks' distance and the observation numbers (lm (BPSysAve ~.)).

To make the model suitable for analysis, we found the transformation of the model and obtained power = -1. We then ran the regression again with $(BPsysAve)^{-1}$, it showed that the value of ^{R2} has increased to 0.26 (26%), and the variable SmokeNowYes had a positive relation with BPsysAve. Moreover, the regression suggested that a 1 unit increase in SmokeNowYes is associated with a 2.141e^-5 increase in BPsysAve.

For the standardized residual in fig 3 below, we noticed that the curve looked like a flat U shape. After performing transformation, we again generated both the qq-plot and the standardized residual (Fig.5). We noticed the line is less deviated and is almost at par with the fitted line –

similarly to the standardized residual we noticed that the curve was flattened, to be the same as the fitted blue line.

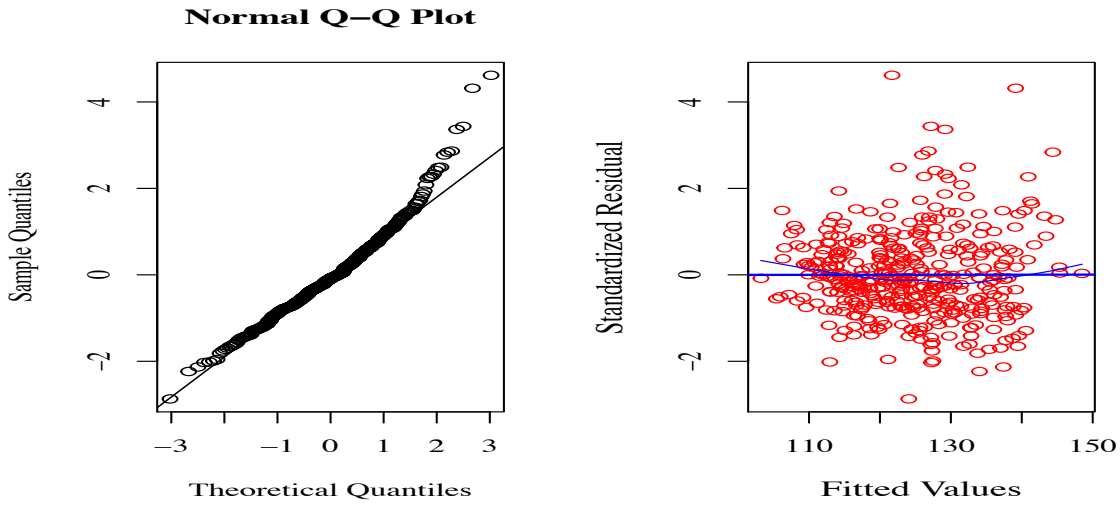


Fig3. The second graph shoes the relation between standardized residual and fitted value.

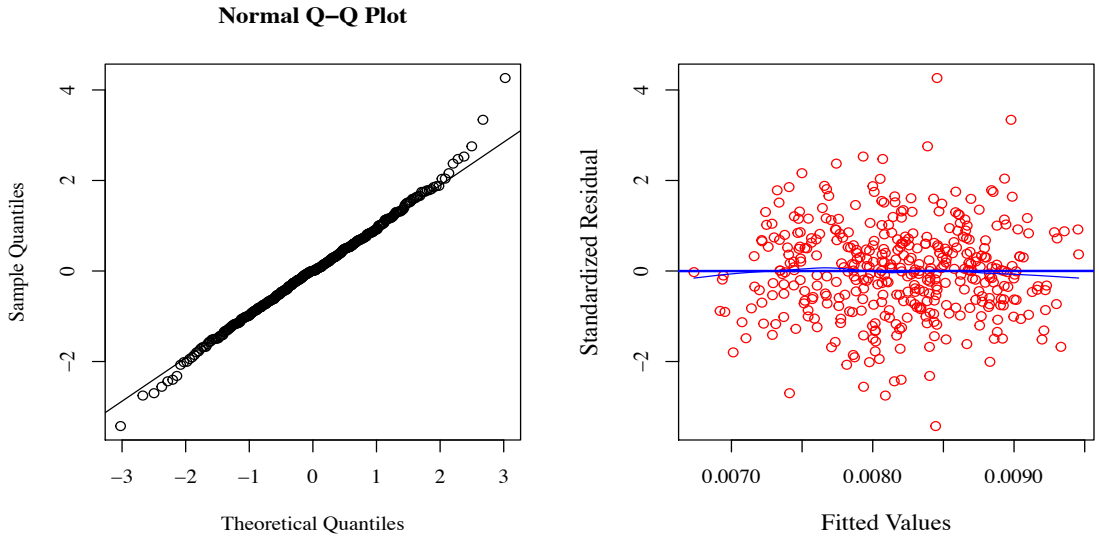


Fig5. This graph shows the relation between the standardized residual with respect to t quantile after transformation.

Knowing that if GVIF > 5 it is preferable for the variables to be removed, the GVIF table obtained allowed us to remove 4 variables: HHIncome, weight, height, and BMI. We ran the regression again with the remaining 12 variables.

We observed that after removing the 4 variables, the betas of some variables had increased. For instance, the intercept of SmokeNow was -0.135, and after the removal it increase to -1.0008.

After obtaining the transformation of the model, we began variable selection techniques to find the best model. First, we began with stepwise variable selection; a procedure based on the information criteria rather than the adjusted *R.* After stepwise variable selection, it showed that only age is a significant variable. However in this analysis, the goal was to see the relationship between SmokeNow and BPsysAve. So, two models were done: one with only age and the other with both age and SmokeNow. Upon adding the variable SmokeNow the value of adj R^2 decreased from 0.1557 to 0.1556. For this analysis, this meant that SmokeNow is not a significant variable and it may very well be removed from the regression.

Second, we move on to backward selection: the process of starting with all predictors, to later delete one predictor at a time; this is equivalent to removing the predictor with the largest p-value. The data obtained showed that both stepwise and backward selection gave us the same values which is that SmokeNow is not a significant variable.

Lastly, we performed forward selection, a process of beginning with no predictors, to later add one predictor at a time such that the resulting model will have the lowest information criterion. All three of the models will select the model with the lowest AIC/BIC. After this selection was done, 10 variables were obtained that are age, race3, education, poverty, maternal status, depressed, sleephrnight, sleeptrouble, physactive, and SmokeNow. Then two models were created: one with smoke and one without, to see the effect of SmokeNow variable on this regressing. It turned out that the variable SmokeNow is not a significant variable

We used these 3 different kinds of selection to find AIC, BIC and LASSO. AIC and BIC are both punished probability rules, they are utilized for picking best indicator subsets in relapse, as well as for looking at none sense models - which standard measurable tests cannot do. The AIC or BIC for a model is normally written in the structure [-2logL + kp], where L is the probability work, p is the number of boundaries in the model, and k is 2 for AIC and log(n) for BIC. LASSO is a relapse examination strategy that performs both variable choice and regularization, to upgrade the expectation exactness and interpretability of the factual model it produces.

We begin by LASSO: after finding the best value of lambda = 0.01831, and noticing that as lambda increase, mean-squared error (MSE) also increase, we confirmed that only poverty has an effect on systolic blood pressure. We then move to AIC, we obtained that both poverty and age are the significant factors. As for BIC, we obtained that only age is significant.

After finding the significant variables using the 3 methods (AIC, BIC, and LASSO) we had to do cross validation calibration. We did this and found their predictor errors and see which one is the best model out of them.

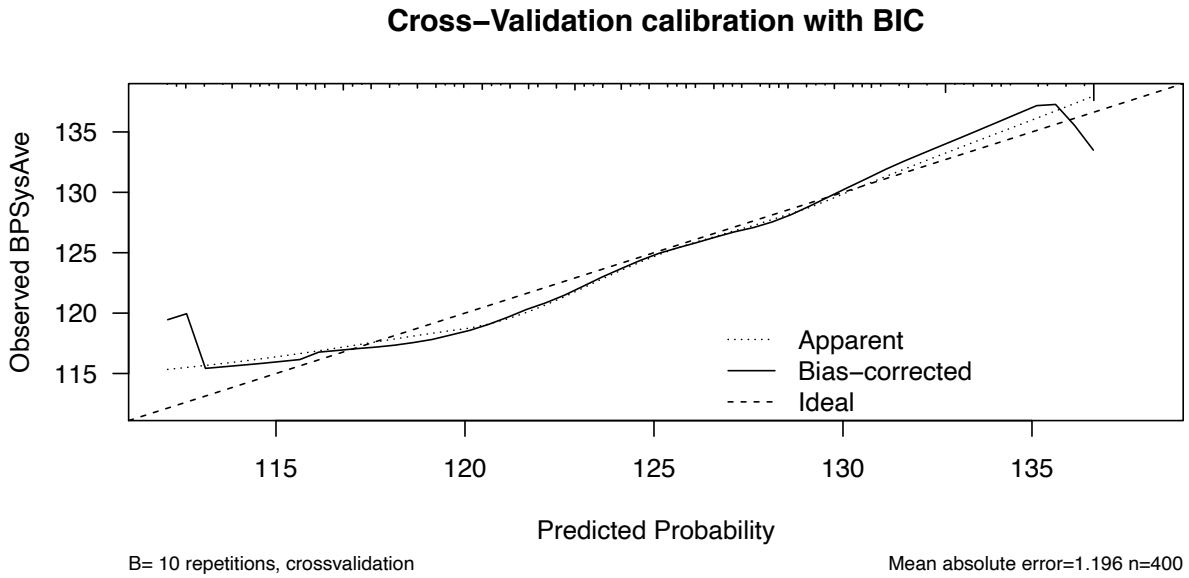


Fig6. This graph shows the relation between observed BPSysAve and the predicted probability using BIC method.

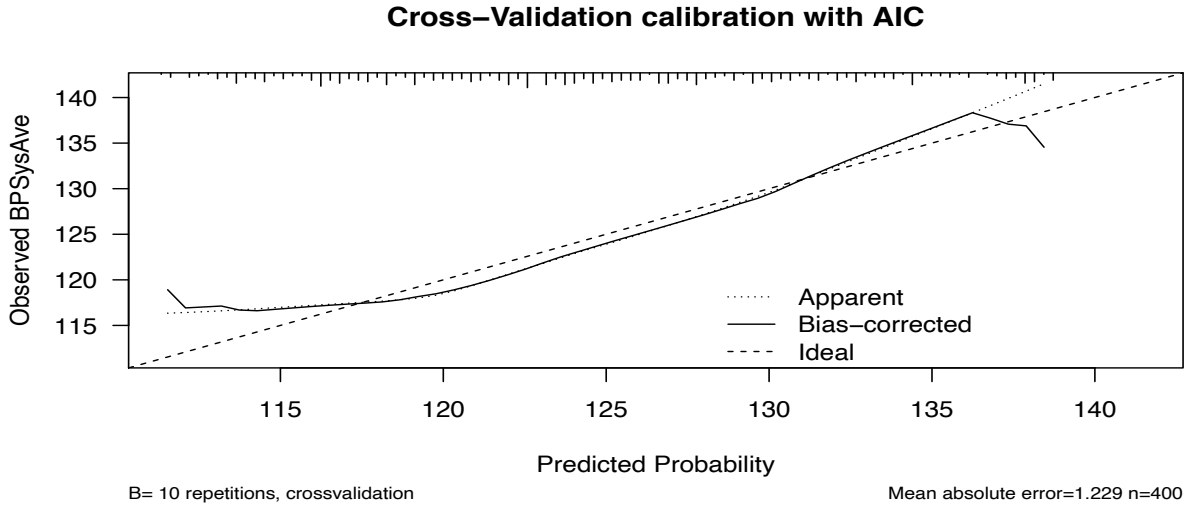


Fig6. This graph shows the relation between observed BPSysAve and the predicted probability using AIC method.

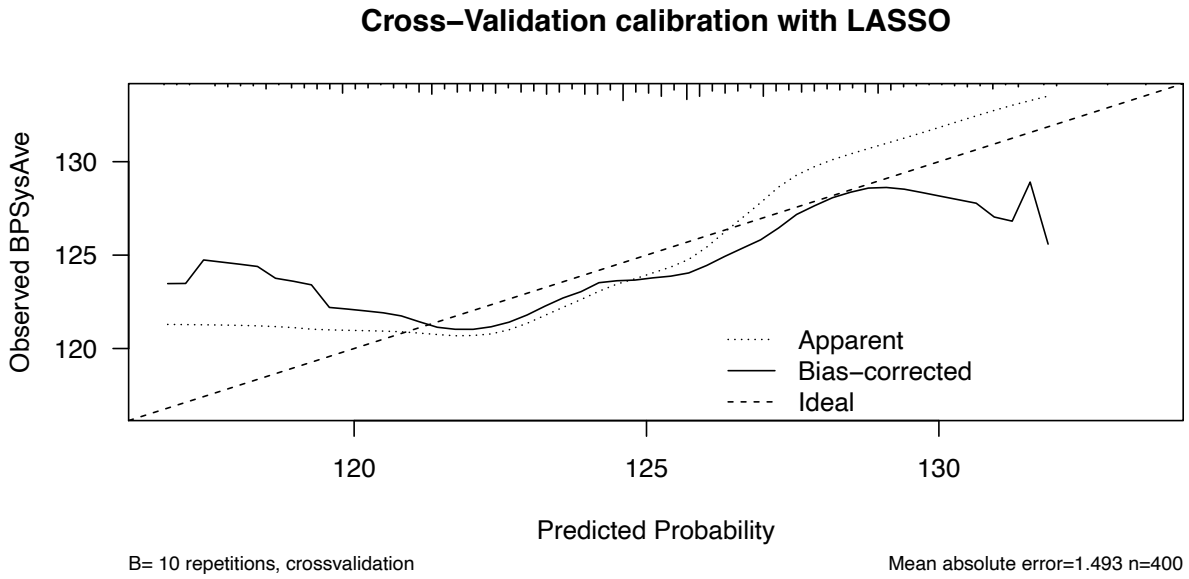


Fig6. This graph shows the relation between observed BPSysAve and the predicted probability using LASSO method.

From the 3 cross-validation graphs above, we can clearly see that using LASSO will not give us the best model, since both the bias-corrected line and the dotted apparent line are way off the ideal one. But we still have AIC, and BIC. We can see that they are both similar to each other, but AIC is better since at the beginning, in BIC, there is a deep steep, and both the bias-corrected line and apparent, slightly deviated from the ideal line. Unlike AIC, where the line begins with a gradual decrease and then it's pretty similar the ideal line.

To check if that is the correct model, we also found their predicted error, and as it turns out the AIC had the lowest error, equal to 248.46. This further confirmed the notion that AIC is the best model.

The result from this analysis indicated that the variable SmokeNow is not an important variable, given our best model AIC - which states that age and poverty are the significant variables. Now, this makes sense because if a person in below the poverty line, most of the variables are going to be affected (education, weight, BMI, depression, etc.,). Poverty will also affect the variable of interest SmokeNow, as they won't be able to buy cigarettes in the first place.

In all, we can boldly state that age and poverty represent the primary variables that will minimize the predicted error of the model, therefore making it the best model.