IS SYSTOLIC BLOOD PRESSURE ASSOCIATED WITH AGE AND SMOKING? A POPULATION-BASED STUDY OF AMERICAN ADULTS

Rachel Hyunbin Oh

INTRODUCTION

Blood pressure is read in two numbers, systolic over diastolic, and measured in millimeters of mercury. Systolic blood pressure (SBP) represents the pressure at the maximum part of your heartbeat when the heart chambers contract to push blood through your blood vessels. High blood pressure is defined as SBP of 130 or higher.[1] It is a major public health problem around the world and the leading risk factor for global disease burden, caused nearly half a million deaths in the United States in 2017.[2] Therefore, it is important to know what causes high SBP.

On average, SBP rises with age.[3] While a certain amount of blood pressure increase is unavoidable as we age, blood pressure health can still be maintained by following the same lifestyle recommendations as younger people. However, the relationship between smoking and BP are equivocal, as some studies show a positive association while others show that there is no relationship. In this study, we are going to investigate the association between age and smoking and SBP in adults in the U.S.

METHODS

Study population

This analysis was conducted using data from NHANES, which includes a series of cross-sectional nationally representative health examination surveys. Data are collected from a representative sample of the civilian noninstitutionalized U.S. population, by in-home personal interviews and physical examinations.

Data

A representative sample includes 743 Americans older than 17 years. 400 observations were randomly selected from the data and this set is going to be referred to as train data. The rest of the data will be used as a test set (343 observations).

Model violations/diagnostics

According to Cook's Distance, there was no influential observation in the train data. However, when used DFFITS and DFBETAS, there were some influential observations, so the common

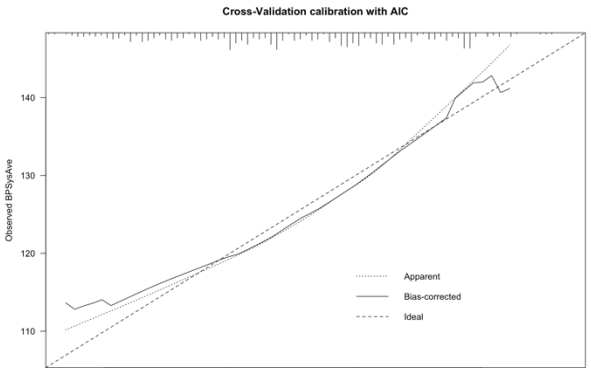
influential observations detected both by DFFITS and DFBETAS were excluded from the train data. The updated train data that now has 387 observations is going to be used for variable selection.

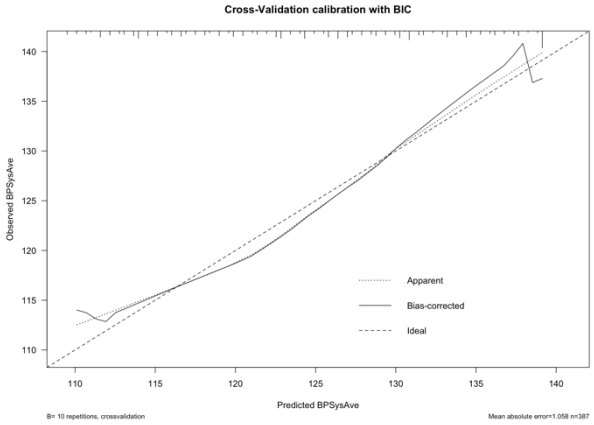
Variable selection

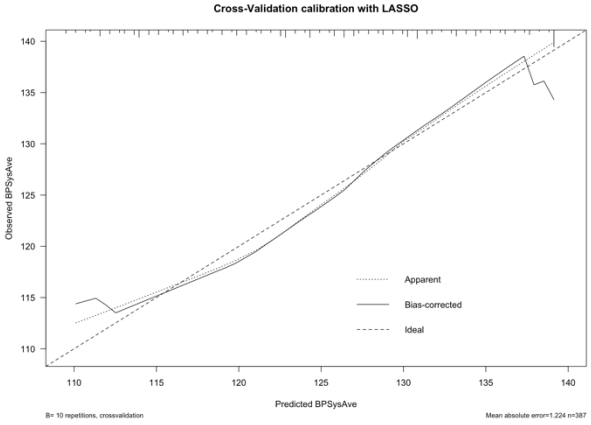
Variable selection based on AIC selected three predictors, 'Age', 'Race3' and 'MaritalStatus'. For BIC and LASSO, only one variable, 'Age' was selected. Although we are mainly interested in the effect of 'SmokeNow' on 'BPSysAve', this variable was not selected in any of the methods above. In order to check the relationship between 'SmokeNow' and 'BPSysAve', we first have to figure out if we can fit a Simple Linear Regression model by checking the correlation between 'SmokeNow' and other variables. However, 'SmokeNow' had high correlation with some variables such as 'EducationSome', so SLR could not be progressed. Hence, the final model would only include 'Age' as a predictor, but we will also see the results of a model including 'SmokeNow'.

Model validation

10-fold cross validation and prediction performance of AIC, BIC and LASSO based selection was done. The prediction error using the test data was 232.7677 for AIC and 224.7722 for both BIC and LASSO. The lowest prediction error value among these was 224.7722, which is from BIC and LASSO that selected 'Age' as the only predictor.







As you can see from the graphs above, for AIC based, the line is not as close to the 45-degree line as the ones for BIC and LASSO based, which again conveys that including 'Age' as the only predictor is the best option.

RESULTS

We are going to see the summary table for two models, the first one (left) with two predictors, age and current smoking (SmokeNow), and the second one (right) with just one predictor, age.

For the first model (two predictors), the following hypotheses are tested:

- H0: There is no linear association between BPSysAve and Age and SmokeNow.
- Ha: There is a linear association between BPSysAve and Age and SmokeNow.

	BPSysAve						BPSysAve					
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*	*df*	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*	*df*
(Intercept)	99.76	2.86	94.14 – 105.38	34.91	**<0.001**	384.00	99.62	2.40	94.90 – 104.35	41.43	**<0.001**	385.00
Age	0.49	0.05	0.40 – 0.58	10.50	**<0.001**	384.00	0.49	0.04	0.41 – 0.58	11.18	**<0.001**	385.00
SmokeNow [Yes]	-0.14	1.65	-3.38 – 3.09	-0.09	0.930	384.00						
Observations	387						387					
R^2 / R^2 adjusted	0.245 / 0.241						0.245 / 0.243					

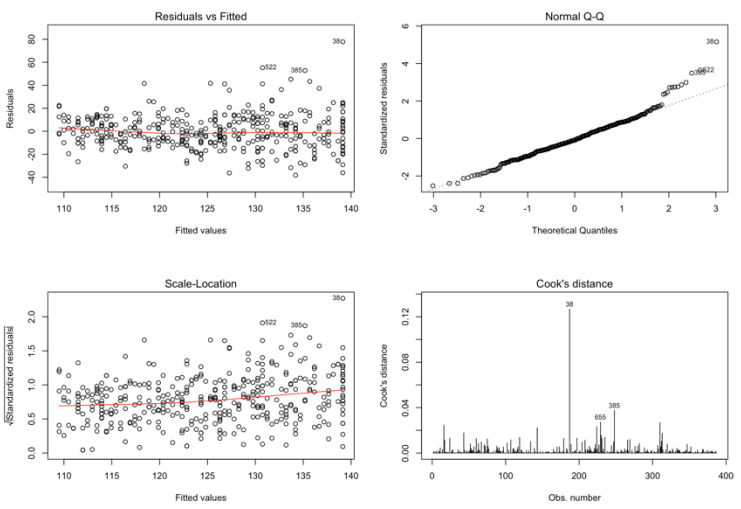
- When testing the null hypothesis for Age, we reject the null hypothesis (t = 10.50, df = 384, p-value < 0.001). For a one-unit increase in Age, on average, BPSysAve increases by 0.49.
- When testing the null hypothesis for SmokeNow, we fail to reject the null hypothesis (t = -0.09, df = 384, p-value = 0.930). For a one-unit increase in SmokeNow, on average, BPSysAve decreases by 0.14.

However, for the second model (one predictor), the null hypothesis that there is no linear association between BPSysAve and Age is rejected (t = 11.18, df = 385, p-value < 0.001). Thus, we conclude that there is a linear association between SBP and age. These results once again convey that it is a suitable choice to determine the final model with just one predictor, age, and claim that SmokeNow has very little or no relationship with SBP.

Goodness of Final Model

The final model fitting is just the first part of the story for regression analysis since this is all based on certain assumptions:

1. Independence: Y is independent of errors.
2. Linearity: The relationship between X and Y is linear.
3. Homoscedasticity: The variance of residual is the same for any value of X.
4. Normality: For any fixed value of X, Y is normally distributed.



The plot of Residuals vs Fitted values is useful for checking the assumptions of independence, linearity and homoscedasticity. To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0. To assess if the homoscedasticity and independence assumptions are met, we look to make sure that there is no pattern in the residuals and that they are equally spread

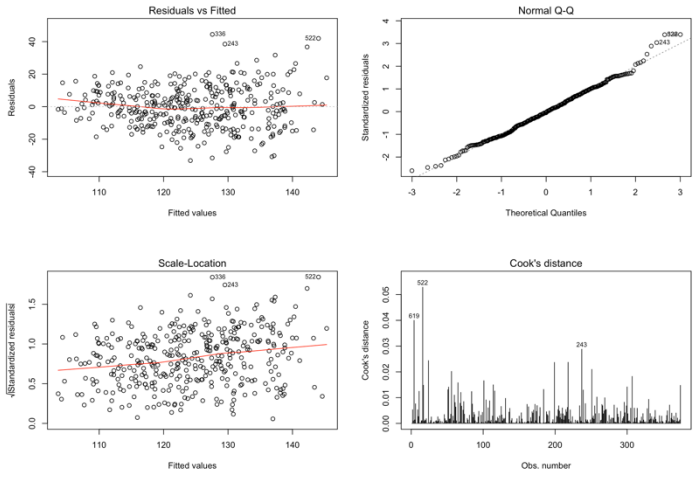
around the y = 0 line. By looking at the first graph above, we see no pattern and the red line is fairly flat. The points except for some observations are equally spread around y = 0 line. Therefore, we conclude that independence, linearity and homoscedasticity assumptions are met for the final model, but we need to further investigate some observations.

The normality assumption is evaluated using a Normal Q-Q plot. Observations except for the last few lie along the 45-degree line in the QQ-plot. Therefore, the removal of these influential observations will help making the normality assumption to be met.

The third plot is a scale-location plot and is useful for checking the assumption of homoscedasticity. In this particular plot we are checking to see if there is a pattern in the residuals. Since we see no pattern, it shows that homoscedasticity assumption is met.

The fourth plot is of Cook's distance, which is a measure of the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances requires investigation. In our case, observation 38 has larger Cook's distance than other data points in Cook's distance plot. Hence, we are now going to examine model diagnostics to decide whether to leave out this observation or not.

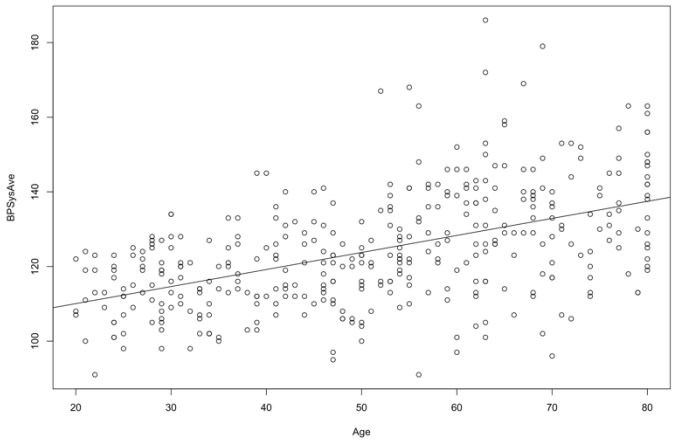
According to Cook's distance we calculated, there was no influential observation. However, there were 13 common influential observations detected both by DFFITS and DFBETAS, so these observations have been removed from train data to help meet the assumptions for linear regression model. The number of observations in train data has now decreased from 387 to 374. The graphs below show the effect of removing the influential observations.



Overall, the points are more spread around the y = 0 line in the first plot, well on the diagonal line in the second plot, still have no pattern in the third plot, and lastly, the average Cook's distance is closer to 0.

DISCUSSION

Using the final model after handling diagnostics, we conclude that age was a significant predictor of performance on the test of systolic blood pressure as 'Age' and 'BPSysAve' parallelly increased (shown in graph below). Nevertheless, there were some cases where



BPSysAve was much higher than the average of it for higher age, and also some where BPSysAve was much lower than the average of it for lower age.

However, it is also concluded that current smoking was not a risk factor of high systolic blood pressure as there was very little or no relationship between smoking status and systolic blood pressure.

The final model is limited that not all inputs of our main interest have been included in the model to check the direct relationship. Multicollinearity was also an obstacle to fit a simple linear regression model for current smoking. Therefore, we could only roughly say that there was little or no relationship between current smoking status and combined systolic blood pressure reading but could not assure with numbers and graphs.

---

[1] National Institute on Aging. High blood pressure. Updated May 2, 2018

[2] Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death, 1999–2017. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed January 7, 2019.

[3] Williams B, Lindholm LH, Sever P. Systolic pressure is all that matters. Lancet. 2008;371:2219–21. [PubMed] [Google Scholar]