Does smoking has a large effect on the combined systolic blood pressure reading? How can we predict it?

Qinxi Yu

Abstract

- **Objects**: Find out the effect of actively smoking(SmokeNow) on the combined systolic blood pressure reading(BPSysAve). And identify the best variables selected to predict BPSysAve.
- **Methods**: The data is collected by US National Center for Health Statistics(NCHS) targeting the non-institutionalized civilian resident population of US. Complex survey designs are used. Multiple linear regression is used to conduct the analysis.
- **Results**: Whether one is smoking actively or not does not have a large effect on his or her BPSysAve and a slightly higher BPSysAve is found with active smokers. To predict, the best variables are gender, age, poverty and interaction(gender&age). Males, older people and people living in poverty generally has higher BPSysAve. BPSysAve of females tend to increase faster as they get older.
- **Conclusion**: The public should be warned about the harm of bad BPSysAve, especially males, old people, and people living in poverty. These groups tend to have higher BPSysAve and should pay attention to BPSysAve for the purpose of preventing unnormal BPSysAve which can lead to many diseases.

- **Keyword**: combined systolic blood pressure reading(BPSysAve), Active smoking, gender

Introduction

A high combined systolic blood pressure reading is usually associated with high possibilities of damage in the blood vessels around important organs like heart, brain. According to the American Heart Association, BPSysAce has become a major risk factor for cardiovascular disease for those over 50.(https://www.hindawi.com/journals/crp/2011/264894/). Considering the large number of people smoking actively, it is important for the public to know the effect of active smoking on the BPSysAve and also to know which group of people should be more concerned about their BPSysAve. The purposes of this study are to analysis the effect of smoking now on BPSysAve and also to find out the most important factors that can influence BPSysAve.

Methods

Model Diagnostics

The data in use has 17 variables and 743 observations. 400 observations are randomly selected as training data and the rest is used as testing. First, residual plot is drawn to assess assumptions visually and to identify any outlier if exists. We will use cook's distance, DFFITs, DFBETA

which are widely used to measure influence of observations. After calculating the 3 measures of each observation, we will decide which points to remove. Then we will redraw the residual plot and QQ plot to check if the outliers still exist and if the assumptions are met.

Check MultiCollinearity

To check the multicollinearity of the model, correlation matrix and VIFs are calculated. As we are mainly interested in effect of active smoking on BPSysAve, the focus will be on the correlation of active smoking with other predictors and VIF of it.

Variable Selection & Model Validation

Effect of active smoking on BPSysAve

Since the estimated coefficients from post-model-selection-model is biased, relationship between active smoking and BPSysAve is analyzed before model selection. To check possible interaction effect of active smoking with other factors, interactions between active smoking and other numerical variables are added to the model, working as full model. Notice number of observations outnumbers number of predictors even after adding interactions. A reduced model is fitted with the variable SmokeNow and its interaction variables removed. Since variables in reduced model is a subset of variables in full model and we are checking the effect of the removed predictors on BPSysAve by comparing two models, we will apply partial F-test to analyze.

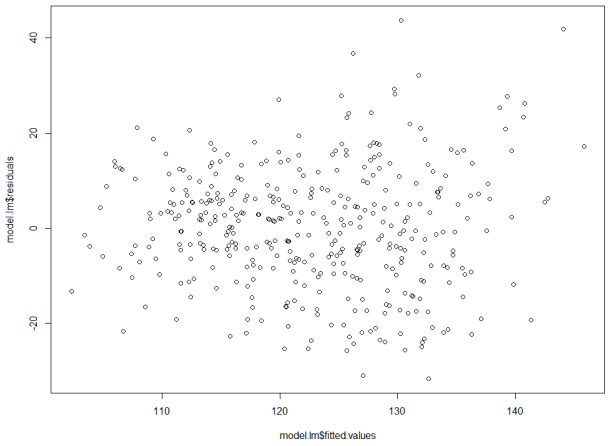
Select variables that are best for the prediction

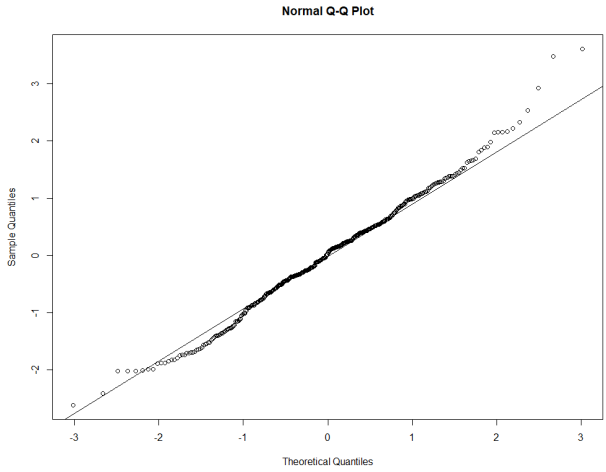
Three common variable selection methods are applied. Criterions AIC, BIC are used to compare the model. We add/delete one variable according to AIC/BIC after each addition/deletion until it is unable to add/delete further variables, which is known as stepwise selection. So that two models based on AIC and BIC are given. The third model is obtained by shrinking parameter estimates to zero, which is known as LASSO. To select from 3 models, we will split the training data randomly into 10 parts and fit model using 9 parts ,then check the prediction accuracy on the validation dataset by matching predicted value and actual value, known as cross validation. Also, we will calculate prediction error on the actual test data using each model. Matching better in cross validation and having smaller prediction error means the model works better for prediction.

Results

Diagnostics

As we can see, there are outliers in the dataset. Checking based on cook's distance cannot find influential points, but checking based on DFFITs and DFBETA indicate 22 and 23 influential observations. It means these observations are not very influential to least square estimates but are influential to fitted values and coefficient estimates. Since the purpose of the analysis is to find the relationship between active smoking and BPSysAve and find model that is best for





prediction, we will remove influential observations indicated by both DFFITs and DFBETA. Then we redraw the residual plot (shown above). This time outliers are removed and all points are scattered around 0. There is no obvious separation of clusters of residuals from the rest. Hence we can verify linearity and homoscedasticity. For normality assumption, the residuals are slightly more spread out for very large and small fitted values. Also on QQ plot (shown above), points at both ends wiggle slightly. Because the deviation is slight, we will ignore the violation and also verify the normality assumptions.

Multicollinearity

Variable SmokeNow has a small VIF and SmokeNow does not have strong correlations with any other variables. Although some of other variables have high VIF, for the purpose of prediction, we will not remove any variable.

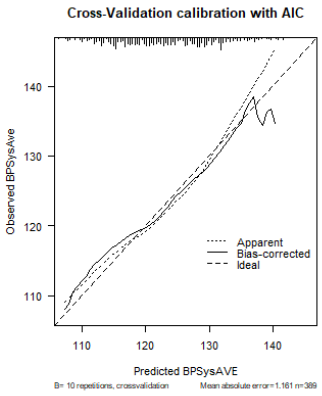
Effect of active smoking on BPSysAve

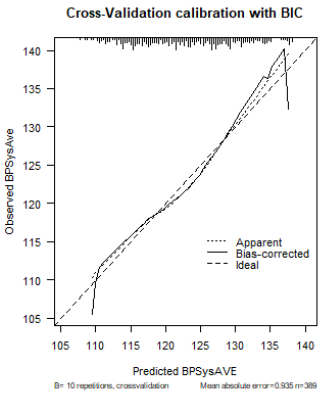
According to the partial-F test shown below , almost the same amount of variability are explained by reduced model and full model, implying that active smoking may not be necessary.

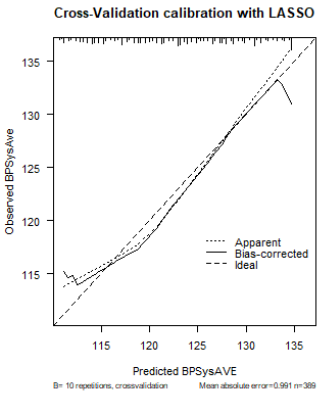
```
Analysis of Variance Table

Model 1: BPSysAve ~ Gender + Age + Race3 + Education + MaritalStatus +
    HHIncome + Poverty + Weight + Height + BMI + Depressed +
    SleepHrsNight + SleepTrouble + PhysActive
Model 2: BPSysAve ~ Gender + Age + Race3 + Education + MaritalStatus +
    HHIncome + Poverty + Weight + Height + BMI + Depressed +
    SleepHrsNight + SleepTrouble + PhysActive + SmokeNow + SmokeNow *
    Age + SmokeNow * Poverty + SmokeNow * Weight + SmokeNow *
    Height + SmokeNow * BMI + SmokeNow * SleepHrsNight
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    352 62488
2    345 61640  7    848.33 0.6783 0.6905
```

SmokeNow does not have a strong effect on BPSysAve. By previously fitted model, people who are smoking actively tend to have slightly higher BPSysAve.

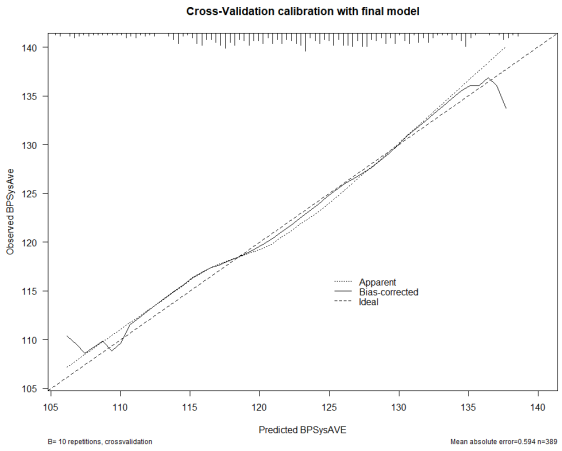
Variable Selection & Model Validation







The shown cross validation calibration plots of three models selected based on AIC, BIC, and LASSO all match well. The model selected based on BIC has the smallest prediction error. Since we are interested in prediction, we proceed to work with the model selected based on BIC, which has predictors of gender, age, and poverty. Notice gender is a binary categorical variable. We will fit the models with added interactions between gender and other variables to check if the model fits better. R^2, R^2adjusted, AIC, AICc, and BIC are calculated to compare models. All criterions except R^2 agree with the selection of the same model. Considering R^2 tend to be larger for model with more predictors, which is the case here, we will select the model indicated by other 4 criterions. Hence, predictors of our final model are gender, age, poverty, and interaction(gender&age).



Again based on final model, we redo cross validation(shown) and calculate prediction error to check its prediction accuracy. We can see that although the calibration is still slightly deviating at both ends but the calibration fits well in general. The prediction error is 267.58 which is smaller than the original prediction error 288.17. Residual-fitted_value plot and QQ plot are also redrawn. Although there is still a little violation of normality assumption at both sides of the plot, the overall assumptions are still satisfied and we can hence ignore the slight violation. Summary of the final model shown.

```
> summary(model.final)

Call:
lm(formula = BPSysAve ~ . + Age * Gender, data = train[, which(colnames(train) %in%
    c(sel.var.bic, "BPSysAve"))])

Residuals:
    Min      1Q  Median      3Q     Max
-32.982  -9.359   0.086   8.545  55.639

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    95.50545    3.59903  26.536  < 2e-16 ***
Gendermale     14.86387    4.42381   3.360 0.000858 ***
Age             0.55326    0.06374   8.680  < 2e-16 ***
Poverty        -1.05041    0.41773  -2.515 0.012326 *
Gendermale:Age -0.22442    0.08292  -2.706 0.007103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.54 on 384 degrees of freedom
Multiple R-squared:  0.2346,    Adjusted R-squared:  0.2266
F-statistic: 29.42 on 4 and 384 DF,  p-value: < 2.2e-16
```

Discussion

Our final data does not include

SmokeNow as predictor. Both the analysis and the final data agree on whether one is smoking active does not have a large effect on the BPSysAve. Although one fitted model shows that people who is actively smoking tend to have slightly higher BPSysAve, this may not be true in realistic. However smoking should still be avoided as much as possible for the consideration of one's other health conditions. According to our analysis, in general, BPSysAve of males tend to be about 14.8 higher than BPSysAve of females. Possibly caused by fact that males are more likely to have bad habits like long-term smoking or drinking. A female's BPSysAve is likely to be increased by about 0.55 as she gets one year older, this increase is however lower for a male, which is about 0.33. Older people or people living in poverty tend to have higher BPSysAve. The public, especially the mentioned groups should be warned about harm of BPSysAve and pay more attention. Promotions of things like healthy eating habits or lower working pressure might help. Notice we ignored the slight deviating of validation for very large and very small predicted value. This means although our model can give good predictions in most cases, the prediction may not be very accurate when the predicted value is larger than around 135 or smaller than around 107.