Is systolic blood pressure reading related to gender, age, poverty, weight, sleep trouble, and smoking habit? American population-based study

Jiangyue Mao

1003928039

I. Introduction

This project aims to explore mostly related variables to the change of systolic blood pressure reading (BPSysAve) and the best predictors for combined systolic blood pressure accurate reading in which the effect of smoking (SmokeNow) is to be identified. The dataset drives from 'NHANES' survey data collected by the US National Center for Health Statistics (NCHS). The analysis of the effects of the variables will manifest the behaviors or body properties contributing most to systolic blood pressure to maintain a good health.

II. Methods

1. Variable Selection

In stepwise selection, based on AIC, the variables of "Gender", "Age", "Poverty", "Weight", and "SleepTrouble" are selected in the direction of backward and both, and all the variables are selected in forward direction, while based on BIC, the variables of "Gender" and "Age" are selected in the direction of both and backward, and all variables in the forward selection. To elaborate the necessary variables, Lasso selection is performed by applying cross validation, showing only the coefficient of "Age" is greater than 0, leaving the variable "Age" selected in Lasso method. We also establish new models by adding SmokeNow to each model, checking if the models can produce better inferences and predictions. By fitting these models to the training dataset, we do comparison to see if the new ones fit the data better. Table 1 shows although the original model might fit the data better with the lowest AIC, AICc, and BIC, the model of BPSysAve ~ Gender + Age + Poverty + Weight + SleepTrouble (denoted by model.1) has a higher adjusted R squared, with other properties only slightly higher than the original model. With SmokeNow predictor added to model.1, AIC, AICc, and BIC even decrease, and further model validation is needed to check the properties of the models to determine the best model(s) for inferences and predictions respectively.

	Adjusted R Squared	AIC	AICc	BIC
Original model	0.2208034	2170.758	2172.351	2242.612
BPSysAve ~ Age	0.1973340	2220.551	2222.145	2292.406
BPSysAve ~ Gender + Age	0.2112868	2212.530	2214.124	2284.385
BPSysAve ~ Gender + Age + Poverty+ Weight + SleepTrouble	0.2288483	2200.489	2202.083	2272.344
BPSysAve ~ Age + SmokeNow	0.1954749	2220.470	2222.064	2292.325
BPSysAve ~ Gender + Age + SmokeNow	0.2098506	2212.249	2213.843	2284.104
BPSysAve ~ Gender + Age + Poverty+ Weight + SleepTrouble + SmokeNow	0.2279321	2199.947	2201.541	2271.802

Table 1. Adjusted R Squared, AIC, AICc, and BIC of models after variable selections

i) Model inferences

To check the data-fitting ability of models, properties are checked from the following perspectives: coefficient, T-test, F statistic, partial F test, VIF, etc. Firstly, we calculate VIF from the original model, and find that the GVIF^(1/(2Df)) of "Weight", "Hight", and "BMI" are greater than 5, suggesting some predictors needed to be excluded from the model due to the high multicollinearity among predictors. Secondly, the summary(model) indicates the p-value for "Gender", "Age", and "SleepTrouble" are significant, and all models except the original one have a p-value <2e-16 and larger F-statistics, signaling better inferences. Thirdly, Partial F test is performed to reduced models and check the interactions of some variables with the SmokeNow. Model.1 presents the largest p-value, demonstrating the significance of the variables, and should be preserved. Model.1 being of the smallest F-statistic (around 0.8) reflecting both the reduced model and the full model explain a similar amount of variability, so the additional predictors may not be necessary. Next, VIF for all predictors are around 1, demonstrating the minor multicollinearity among predictors. Moreover, Appendix A shows most heavy smokers around the age of 40-60 are at a higher poverty level and at greater risk of suffering from extreme systolic blood pressure compared to non-smokers, but with relatively lower wights compared to non-smokers. The high multicollinearity between SmokeNow and other variables provides another reason for not selecting model with SmokeNow. Overall, model.1 serves as the most appropriate model for data fitting, producing the most accurate estimates for the coefficients of variables based on training dataset.

ii) Model prediction/validation

To explore the model prediction outcomes, we obtain the mean of prediction errors (MPE) by fitting the test dataset to each model and calculate the difference between the true value of BPSysAve in test dataset and the predicted BPSysAve. Then, by shrinkage methods we employ ridge penalty to see if the penalized version of the original model can produce better prediction. However, the MPE of penalized model is 310.7266, higher than all other models. Appendix B shows the lowest prediction error with model.1, and model.1 with "SmokeNow" predictor has little increase in MPE. Furthermore, we perform cross validation to see if the predicted outcomes fit the true values in test dataset. Fig.1 shows the deviation of some predicted BPSysAve from the true values, but the predicted model.1 outcomes with the smallest mean absolute error are still the best among AIC models to fit the true values. Moreover, cross validation of models based on BIC and Lasso are performed, and Appendix C shows predicted values generated from BIC based model can best fit the true values, with the smallest mean absolute error, even smaller compared to model.1. Since the graph exhibits only minor differences between model.1 and BPSysAve ~ Gender + Age model, and the Adjusted R Squared for model.1 is larger compared to the original model, Adjusted R Squared for BPSysAve ~ Gender + Age model being even smaller, we hold that model.1 provides the best predictions.

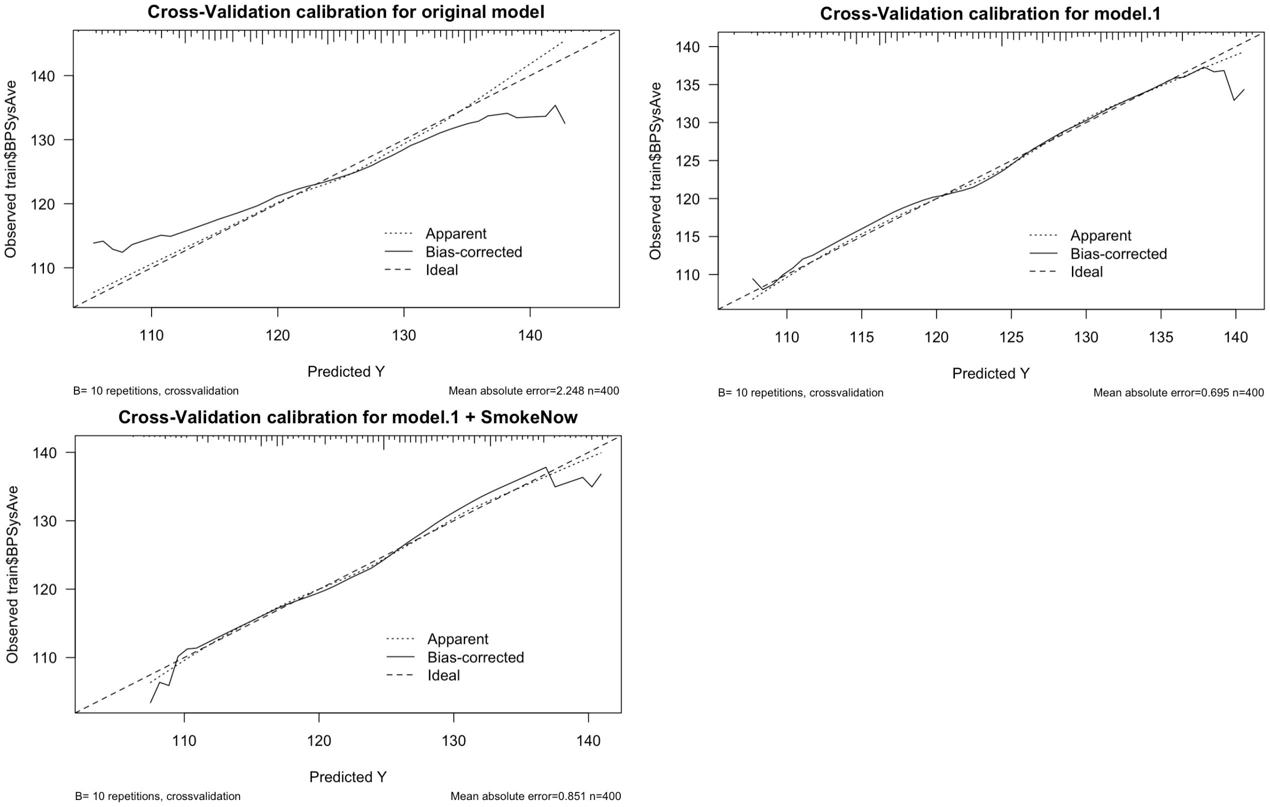


Figure 1. Cross validation for models selected based on AIC and the original model

2. Results

The analysis demonstrates the final model for both best inferences and predictions should be model.1. The coefficient of gender male of the final model is 5.59, the largest among all coefficients, indicating in comparison to the female individuals, male individuals are 5.59 times more likely of developing high blood pressure in average. The coefficients of weight, sleep trouble, poverty and age show similar effects on blood pressure. However, diagnostic check is needed to determine the final model.

III. Diagnostic check for original model and the final model

To judge check model assumptions and find outliers in the model, diagnostic check for the original model and final model is needed. Although some outliers are detected from hii, Defitts and Dfbeta, they are not necessarily to be removed, since Cook's Distance detects no outliers. QQ plot reveals some violation of Normally distributed residuals, whereas Standardized residuals plot shows the residuals are spread around a horizontal line without distinct patterns, exhibiting a nearly linear relationship between predictors and outcomes. And a nearly linear relationship between predicted outcomes and true values can be judged from BPSysAve vs. fitted values plot (Fig. 2).

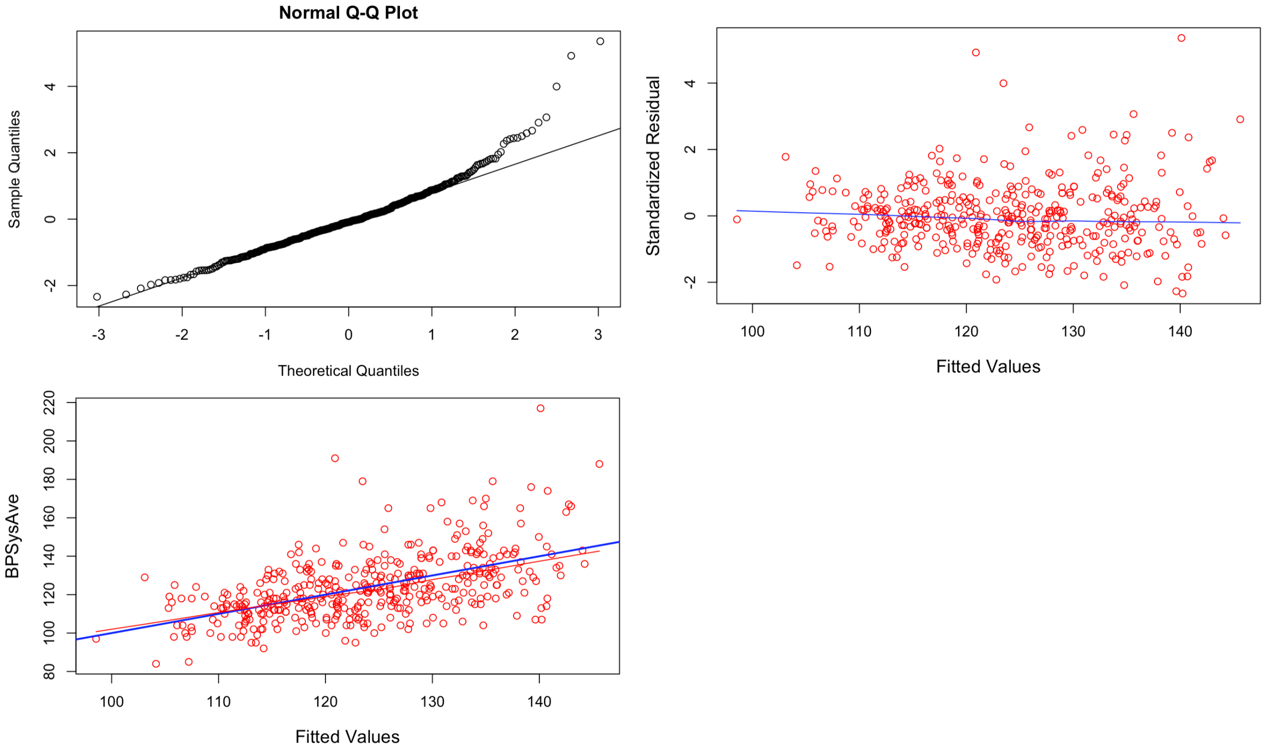


Figure 2. Normality of standard residuals, standardized residuals vs. fitted values, and predicted outcomes vs. true values of Y of the original model.

Next, diagnostic check is performed for the selected model, which has much less outliers judging from Dfbeta, and no outliers from Cook's Distance, Diffits and hii, indicating the model reduces outliers effectively and predictions and inferences are more accurate. Judging from the plots, no huge difference can be seen in diagnostics compared with the original model, basically meeting all assumptions, with a little violation on Normality assumptions of residuals, but does not affect the performance of the model much in all (Fig.3).

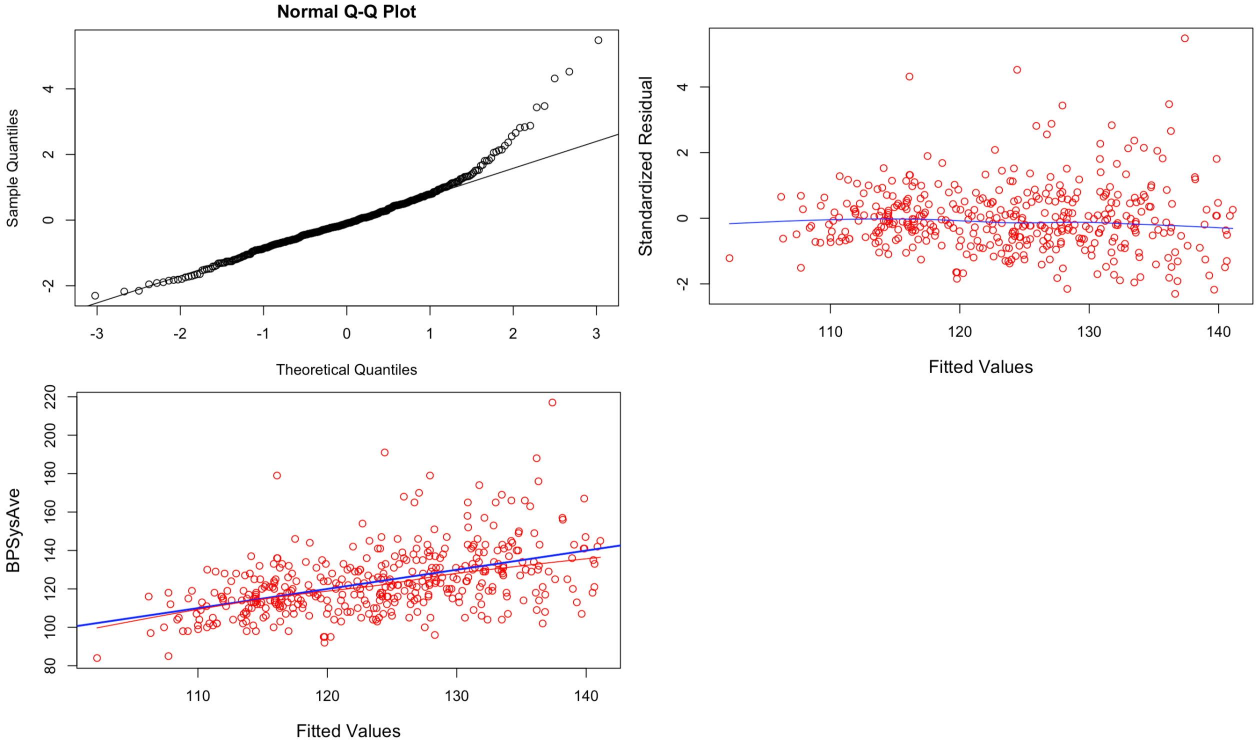


Figure 3. Normality of standard residuals, standardized residuals vs. fitted values, and predicted outcomes vs. true values of Y of the selected model.

In conclusion, meeting all assumptions, the final model is model.1.

IV. Description of data

	Sum	Mean	Standard deviation
Gender — Male	435	0.5854643√	—
Gender — Female	308	—	—
Age	—	50.67026	17.22767
Race —Black	87	—	—
Race —White	506	0.6810229√	—
Race —Mexican	55	—	—
Race —Hispanic	42	—	—
Race —Asian	29	—	—
Race —Other	24	—	—
Education—High School	178	—	—
Education—College Grad	153	—	—
Education—Some College	242	0.3257066√	—
Education—8th Grade	56	—	—
Education—9 - 11th Grade	114	—	—
MaritalStatus—NeverMarried	89	—	—
MaritalStatus—LivePartner	369	0.4966353√	—
MaritalStatus—Married	130	—	—
MaritalStatus—Divorced	84	—	—
MaritalStatus—Widowed	57	—	—
MaritalStatus— Separated	14	—	—
HHIncome—0-4999	20	—	—
HHIncome—5000-9999	26	—	—
HHIncome—10000-14999	51	—	—
HHIncome—15000-19999	57	—	—
HHIncome—20000-24999	62	—	—
HHIncome—25000-34999	93	—	—

HHIncome—35000-44999	81	—	—
HHIncome—45000-54999	60	—	—
HHIncome—55000-64999	46	—	—
HHIncome—65000-74999	41	—	—
HHIncome—75000-99999	70	—	—
HHIncome—more 99999	136	0.1830417√	—
Poverty	1933.33	2.602059	1.632478
Weight	—	83.65397	19.52197
Height	—	170.2999	9.566823
BMI	—	28.80525	6.174117
BPSysAve	—	124.3378	17.72028
Depressed—None	548	0.7375505√	—
Depressed—Several	128	–	—
Depressed—Most	67	–	—
SleepHrsNight	—	6.751009	1.383839
SleepTrouble—Yes	254	0.3418573	0.4746516
SleepTrouble—No	489	0.6581427√	
PhysActive—Yes	355	0.4777927	0.4998431
PhysActive—No	388	0.5222073√	
SmokeNow—Yes	327	0.4401077	0.4967343
SmokeNow—No	416	0.5598923√	
√: Situation for majority people			

Table 2. Variable Summary

According to Table 2, males constitutes most of the interviewees (58%) of which the average age is around 50. The income of approximately 20% is very high (> 99999 dollars annually), below poverty guidelines. The mean of wight and BMI exhibits overall overweight (BMI >25), less than 7 hours average of sleeping in all interviewees. But the mental health condition of interviewees is relatively good, the majority not smoking frequently now and not depressed or sleep troubled.

V. Scalability of the final model

After selecting the final model, we check the inferences and predictions of the final model with test dataset. Although Adjusted R squared is a little lower (0.205) than that of the original model, AIC, AICc, and BIC are much lower than that of the original model, being 1911.9149782, 1913.7865378, 1981.1563958 respectively, indicating the scalability of the final model, not only working for training dataset but also for validation dataset. The significant p-value (2.288e-16) for the final model demonstrate that the model fits for test dataset.

VI. Discussion

1. Final model interpretation and significance

Blood pressure with men is higher than that with women at similar ages[1]. The increase in blood pressure with age is mostly associated with artery structural changes[2].The huge poverty impact on systolic blood pressure may be attributed to poor medical condition. In addition, obesity leads to artery stiffness, affecting stolid blood pressure, while sleep impacts overall cardiovascular health. Thus, the five selected variables in the final model are most related to affect the change of systolic blood pressure, and provides the most accurate predictions based on the dataset. But SmokeNow is not included for its lack of significant effect on inferences and predictions.The larger number of males, the overall old ages, less amount of sleep and higher weight all lead to a high combined systolic blood pressure (124>120).

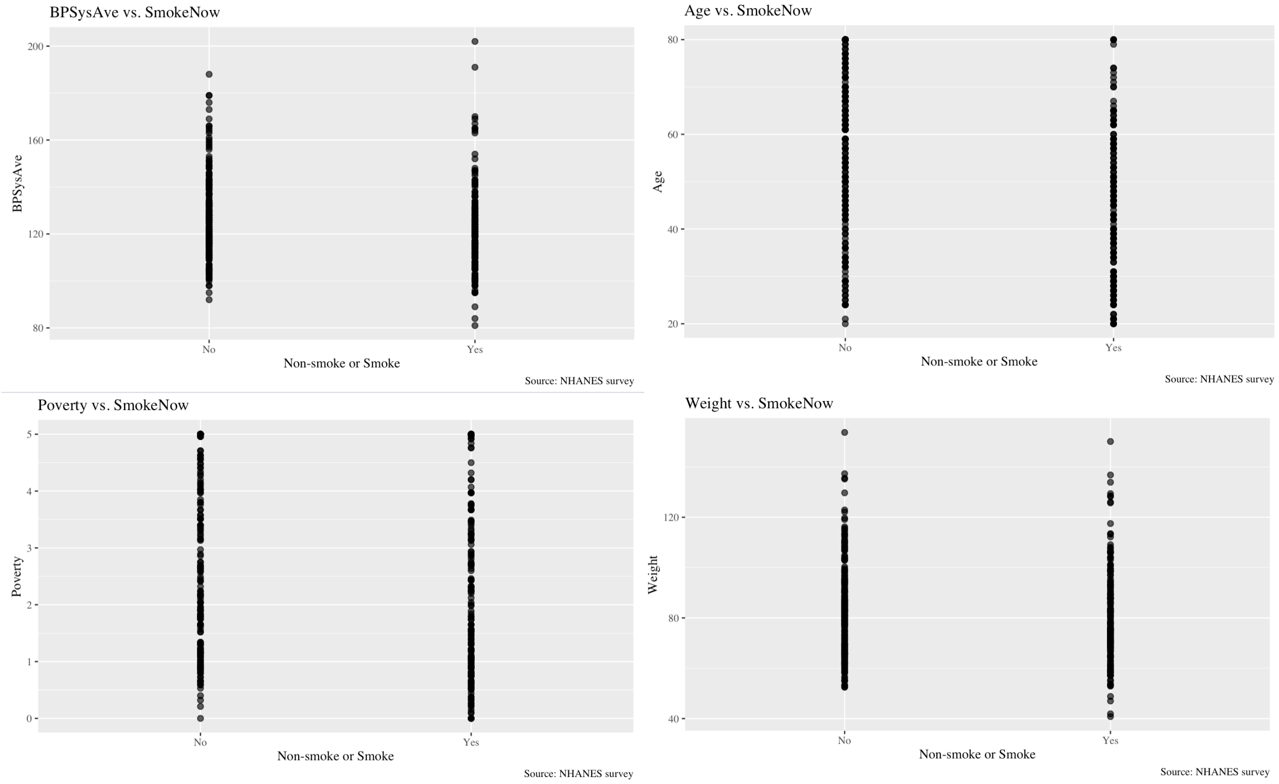
2. Limitations of analysis and potential

Firstly, the dataset is adapted for educational purposes not suitable for research, with more variables and complex issues to affect blood pressure. Secondly, we only fit multiple linear regression models here, not including non-linear models that may improve prediction accuracy (leading to a low Adjusted R squared). Thirdly, the sample size is not large enough to come up with an all-rounded conclusion. However, this model can explore the effects of five predictors in the model for preliminary systolic blood pressure studies.

[1] Jane F Reckelhoff, "Gender Differences in the Regulation of Blood Pressure." *Hypertension* 37, no. 5 (2001): 1199-208. doi:10.1161/01.hyp.37.5.1199.

[2] E Pinto, Blood pressure and ageing. *Postgrad Med J*. 2007;83(976):109-114. doi:10.1136/pgmj.2006.048371.

APPENDIX A


BPSysAve vs. SmokeNow

Age vs. SmokeNow

Poverty vs. SmokeNow

Weight vs. SmokeNow

Source: NHANES survey

Source: NHANES survey

Source: NHANES survey

Source: NHANES survey

Figure 1. MPE of each model based on test set.

APPENDIX B

	MPE
Original model	289.322
BPSysAve ~ Gender + Age + Poverty + Weight + SleepTrouble (AIC)	275.2037
BPSysAve ~ Gender + Age (BIC)	276.501
BPSysAve ~ Age (Lasso)	280.5042
BPSysAve ~ Age + SmokeNow	280.7134
BPSysAve ~ Gender + Age + SmokeNow	276.9964
BPSysAve ~ Gender + Age + Poverty+ Weight + SleepTrouble + SmokeNow	275.6115

Table 1. Cross validation for models selected based on BIC and Lasso.

APPENDIX C

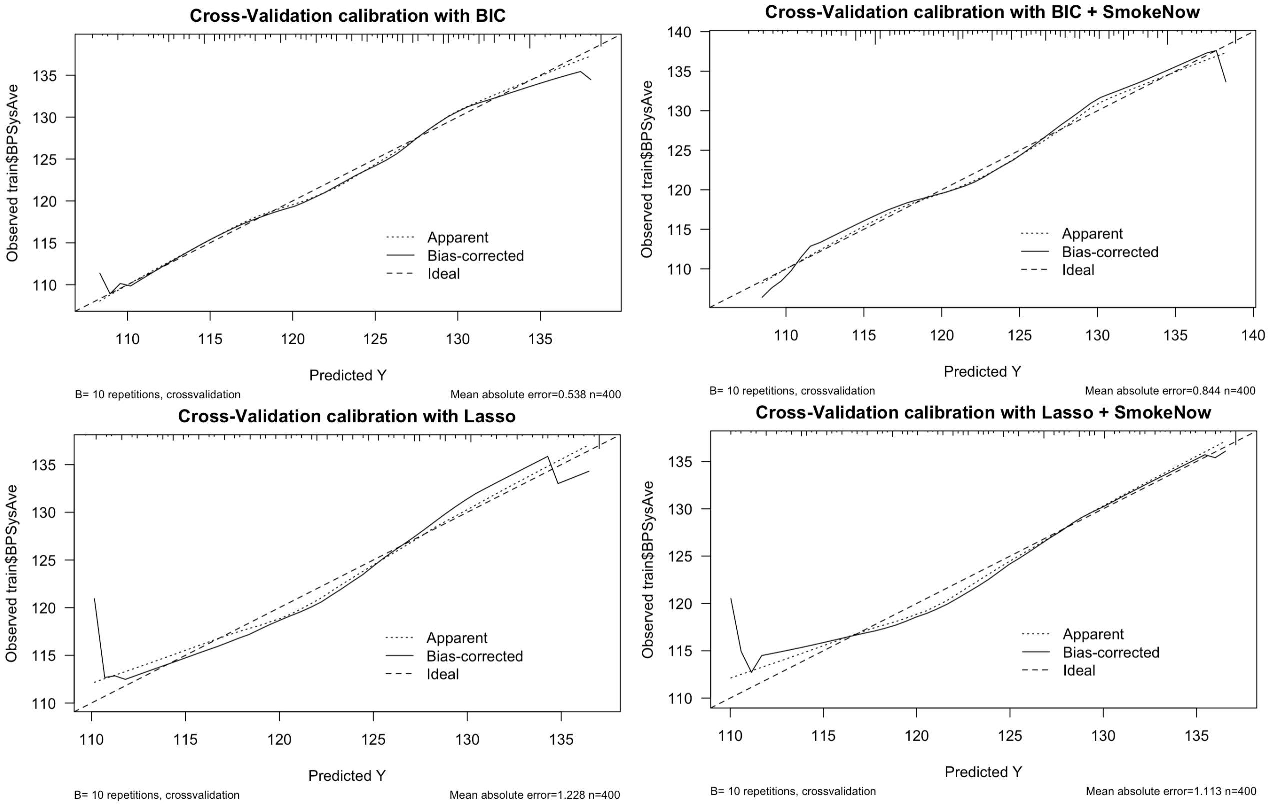


Figure 2. Relationship between SmokeNow and other variables

References

1. Reckelhoff, Jane F. "Gender Differences in the Regulation of Blood Pressure." *Hypertension* 37, no. 5 (2001): 1199-208. doi:10.1161/01.hyp.37.5.1199.
2. Pinto, E. Blood pressure and ageing. *Postgrad Med J*. 2007;83(976):109-114. doi:10.1136/pgmj.2006.048371.