Smoking and other significant contributors to hypertension from NHANES

Anton Sugolov

Introduction

Hypertension is a common sign of poor cardiovascular function and overall health, with smoking being an avoidable, significant contributing factor (Primatesta *et al.* [1]) to increased blood pressure and heart disease. Covariates including BMI, alcohol consumption, and physical wellbeing also affect (Kastarinen *et al.* [2]) blood pressure. The cross-sectional 2012 *National Health and Nutritional Examination Survey* dataset [3] is used to study the effect of current smoking status on average systolic blood pressure through a fitted linear model.

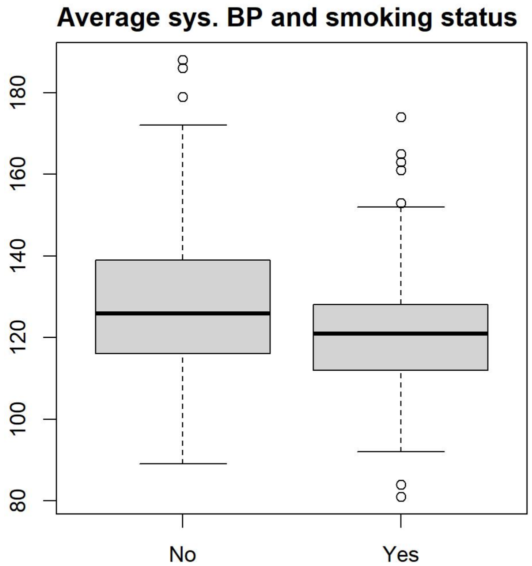

Figure 1: Distribution of average systolic blood pressure in current smokers ($n = 179$) and non-smokers ($n = 221$) in training data.

Methods

The 2011-2012 NHANES ($n = 743$) cross-section was partitioned into training ($n = 400$) and test ($\tau = 343$) subsets. An initial paired t-test shows a statistically significant ($p = 2.96 \times 10^{-4}$) difference in means between smokers and non-smokers in the training data (figure 1), with a higher mean in non-smokers. An initial linear model on the entire NHANES, including gender, age, race, education, marital status, household income, poverty, weight, height, BMI, hours of sleep, reported sleep trouble, current physical activity and current smoking status as predictors was fitted. Race, BMI, and hh. income were removed due to high relation with other variables including poverty and height/weight. A full model using the remaining predictors confirmed that assumptions for a linear model were met. As quantified by DFFITS and DFBETAS, there were no observations having extraordinary influence on the fitted model, or value of predictors. Using different variable selection methods: a punitive information criterion aiming to find the true predictors (BIC), another aiming to best describe the high-dimensional data (AIC), and LASSO shrinkage methods, three candidate models were found. Both LASSO and BIC selection gave a model with only age being a significant predictor, while AIC based variable selection yielded gender, age, marital status, weight, height, and physical activity as best describing the variation.

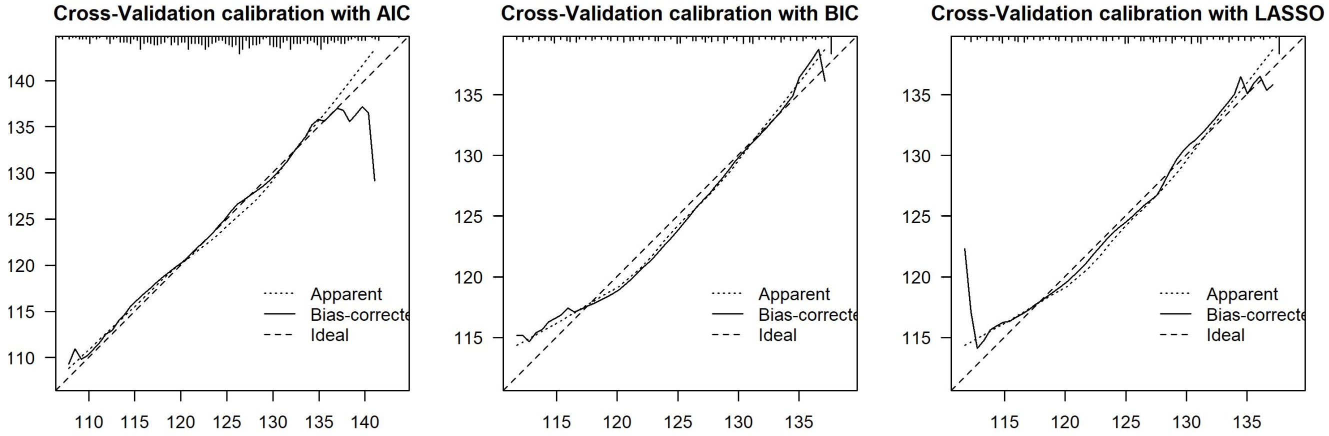

Figure 2: Apparent, bias-corrected, and perfect predictions for average systolic blood pressure in validation datasets during cross-validation.

All models were cross validated on the training set of $n = 400$ with batches of size $k = 20$. From figure 2 and table 3, the AIC based model performs much better during cross validation, having more accurate predictions for a wider range of values and lower mean square error. Validating the model on the test data, the AIC based model had the lowest mean squared error of 278.014 with the BIC and LASSO based models, only using age as a predictor, had a MSE of 285.1103. As a result, the AIC predictors are chosen for our final model. S

Results

The reduced model chosen after variable selection includes gender, age, marital status, weight, height, and physical activity as valuable predictors. Current smoking status is included in the final model in order to study its effect, but it is not significant at the $\alpha = 0.05$ level, while age and weight are. The model predicts current smoking status to decrease systolic blood pressure on average, but is not significant. Performing a partial F-test, the final model explains the same amount of variation as the starting model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.1672	18.3176	6.72	0.0000
Gendermale	3.3483	2.0507	1.63	0.1033
Age	0.4330	0.0564	7.67	0.0000
MaritalStatusLivePartner	-2.5564	3.3913	-0.75	0.4514
MaritalStatusMarried	-1.4812	2.5595	-0.58	0.5631
MaritalStatusNeverMarried	4.3968	2.9615	1.48	0.1384
MaritalStatusSeparated	-4.7796	7.2048	-0.66	0.5075
MaritalStatusWidowed	0.5354	3.5513	0.15	0.8802
Weight	0.0854	0.0422	2.02	0.0438
Height	-0.1603	0.1111	-1.44	0.1499
PhysActiveYes	-3.0777	1.6400	-1.88	0.0613
SmokeNowYes	-0.7327	1.7072	-0.43	0.6680

Table 1: Estimates of predictors, variance, and their significance in the final model, $R^2 = 0.2192$.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Full	388	88967.13				
Final	379	88211.11	9	756.02	0.36	0.9529

Table 2: Partial F-test for the full and final model. The low F statistic indicates that the full model explains a similar amount of variance in average systolic blood pressure as the full model.

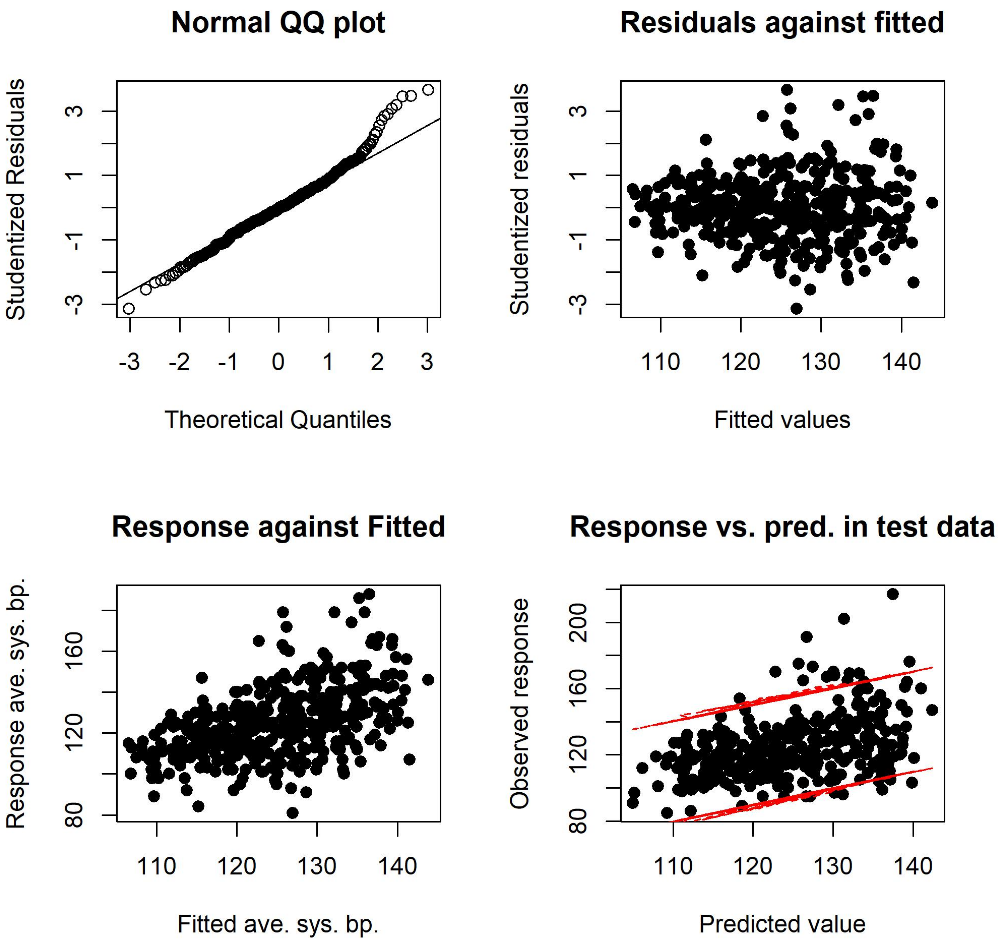

Figure 3: Normal QQ plot, residuals against fitted, and response against fitted and predictions from test data for the final model. Response vs. predictions from the test data are plotted with 95% pred. interval.

The normal QQ-plot and residuals against fitted verify that the normality assumption is met, with some deviation for further values, and that there is no non-linear pattern in the residuals. The response plotted against the fitted values in the model is similar to the response against the test predicted values, indicating a lack of overfitting and an appropriate final model. Most predictions on the test data fall within the 95% prediction interval. DFFITS and DFBETAS calculations yield that there are no significant observation, and the VIF of the model is <5 for all predictors (see table 4). The model is seen to be an appropriate fit.

Discussion

In our model, current smoking status was not found to be a significant predictor of increased average systolic blood pressure. The most significant were age and weight, which are consistent with previous literature [2]. The variables this dataset have limitations that may have caused smoking to be an insignificant predictor. Previous findings have found smoke to affect diastolic blood pressure [2], while systolic blood pressure is the studied response within the dataset. As well, smoking status is defined as having smoked more than 100 cigarettes within a year, which does not account for heavy smoking, prolonged smoking, and ex-smokers [4]. Age may be very significant in this dataset, since there is no distinction between prolonged older smokers and non-smokers. As well, smoking is more popular with older adults, which may introduce confounding in the model [5]. In addition, ages above 80 are categorized as 80, and may provide less accurate prediction for highly prolonged smoking. Further studies with a dataset addressing these flaws should be performed.

References

1. Primatesta, P, Falaschetti, E, Gupta, S, Marmot, M. G. & Poulter, N. R. Association between smoking and blood pressure: evidence from the health survey for England. en. *Hypertension* **37,** 187–193 (2001).
2. Kastarinen, M. *et al.* Trends in lifestyle factors affecting blood pressure in hypertensive and normotensive Finns during 1982-2002. en. *J. Hypertens.* **25,** 299–305 (Feb. 2007).
3. National Health and Nutrition Examination Survey Data.
4. Omvik, P. How smoking affects blood pressure. *Blood Press* **5,** 71–77 (1996).
5. Fiore, M. C. Trends in cigarette smoking in the United States. The epidemiology of tobacco use. *Med Clin North Am* **76,** 289–303 (1992).

Appendix

Vars.	Mean abs. error	Mean sq. error	Quantile of abs. error
AIC	0.564	1.194	0.952
BIC	1.009	1.557	1.933
LASSO	0.925	2.927	1.570

Table 3: Mean absolute error, mean square error, and quantile of absolute error for cross validation of AIC, BIC, and LASSO variable selected models.

	Final model variance inflation factor
Gendermale	1.80
Age	1.72
MaritalStatusLivePartner	1.81
MaritalStatusMarried	2.85
MaritalStatusNeverMarried	2.49
MaritalStatusSeparated	1.12
MaritalStatusWidowed	1.89
Weight	1.22
Height	2.00
PhysActiveYes	1.17
SmokeNowYes	1.26

Table 4: Variance inflation factor for all variables included in final model. Note that all are < 5.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	250.90	250.90	1.09	0.2962
Age	1	23475.98	23475.98	102.38	0.0000
MaritalStatus	5	2121.39	424.28	1.85	0.1022
Weight	1	994.99	994.99	4.34	0.0379
Height	1	559.30	559.30	2.44	0.1192
PhysActive	1	765.76	765.76	3.34	0.0684
SmokeNow	1	42.23	42.23	0.18	0.6680
Residuals	388	88967.13	229.30		

Table 5: Analysis of variance for the final model.