Midterm

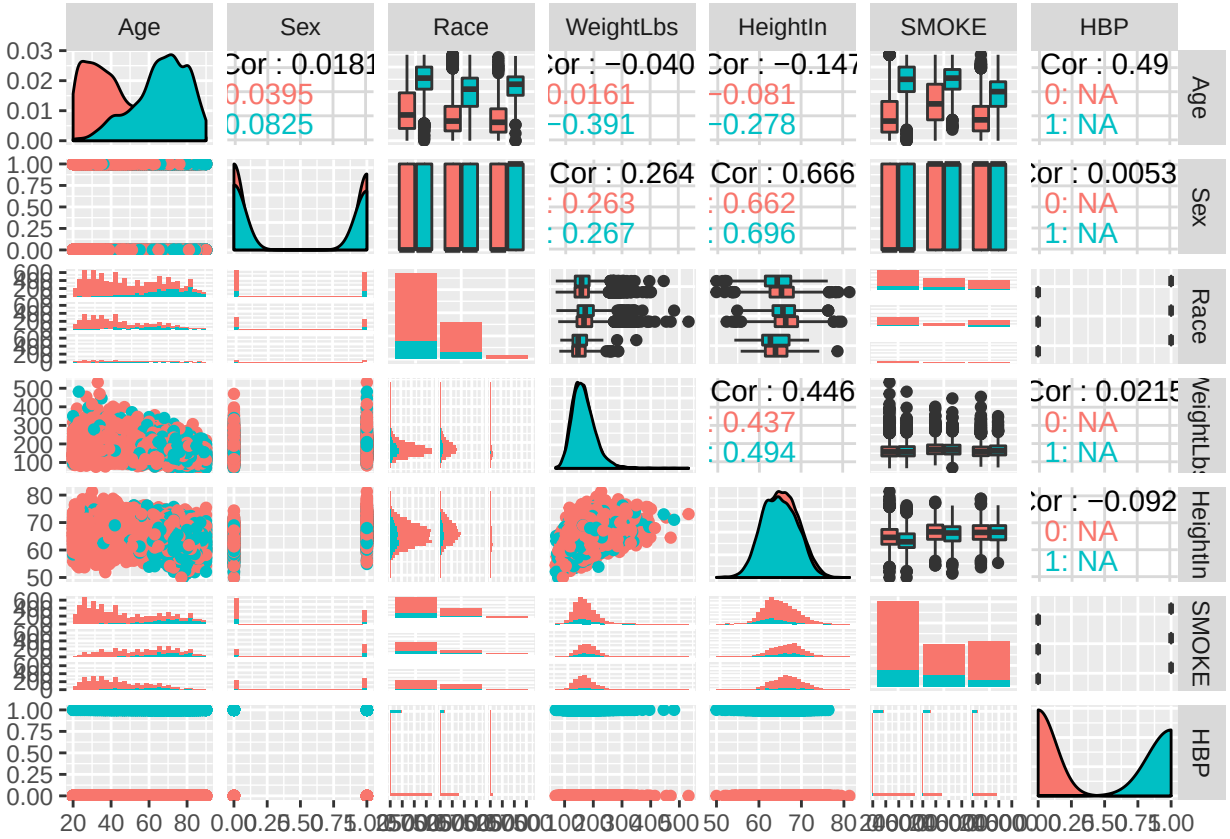
Zhijian Liu

Project 2: Smoking and Blood Pressure

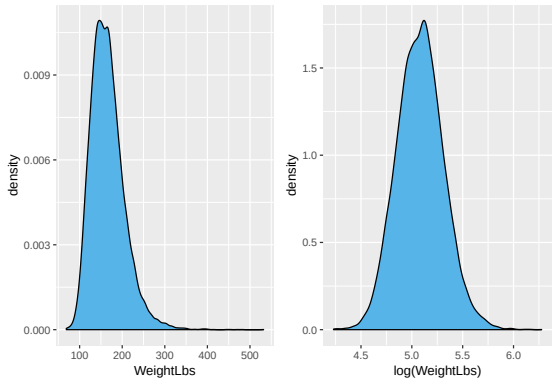
1. Data overview

In this project, we are interested in the relationship between Smoking and High Blood Pressure. Here, `HBP`, High Blood Pressure, will be the response variable, and Smoking, `SMOKE`, will be one of the predictors. Also the interaction between `Smoking` and other predictors, such as `Age`, `Sex`, `Race`, `WeightLbs` and `HeightIn`, is also of our interest.

In the dataset, there are 16441 observations and 8 variables, one of which is Respondent Identification Number, not of our interest. I start with an overview of the data.

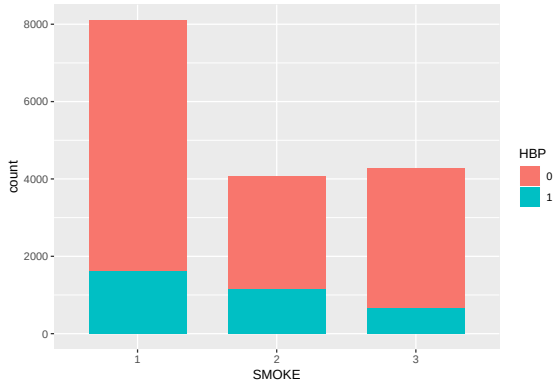


I notice that the distribution of `WeightLbs` is skewed. A transformation on it could help to bulid a model with a better fit. So I apply logarithm to this variable. After the transformation, the distribution is roughly normal distributed.



There is some relationship between predictors. For instance, `WeightLbs` have positive relationship with `HeightIn`. It is a natural relationship, since the taller a person is, the heavier he/she would be. But since our interest is to investigate the relationship between `SMOKE` and `HBP`, I will not dig into the relationship between other predictors. Because, the relationship between other predictors will not significantly affect the effect of `SMOKE` on `HBP`.

Initially, we can have a look at a plot to have a sense of the association between `SMOKE` on `HBP`. The boxplot shows no obvious difference between group 2 in `SMOKE`, people used to smoke, but not any more, and group 3, people still smoke. But the proportion of peole who has high blood pressure in group 1 is apparently lower than that of other two groups. So, I would now consider that people who had smoked are more likely to have high blood pressure.



2. Model

2.1 Model Selection

To initialize a model, I will put all 6 variables i have into the model for the reason that there are a large number of observations in the dataset. Our response variable `HBP` has 0, 1 values, which follows a Bernoulli distribution. So a logistic regression model will be applicable. Futhermore, prior knowledge tells that blood pressure itself is normally distributed in the population. In other word, the binary outcome, HBP, depends on a hidden normal distributed variable. It will be proper to use *Probit* link function in the logistic model. Here, the initial model is:

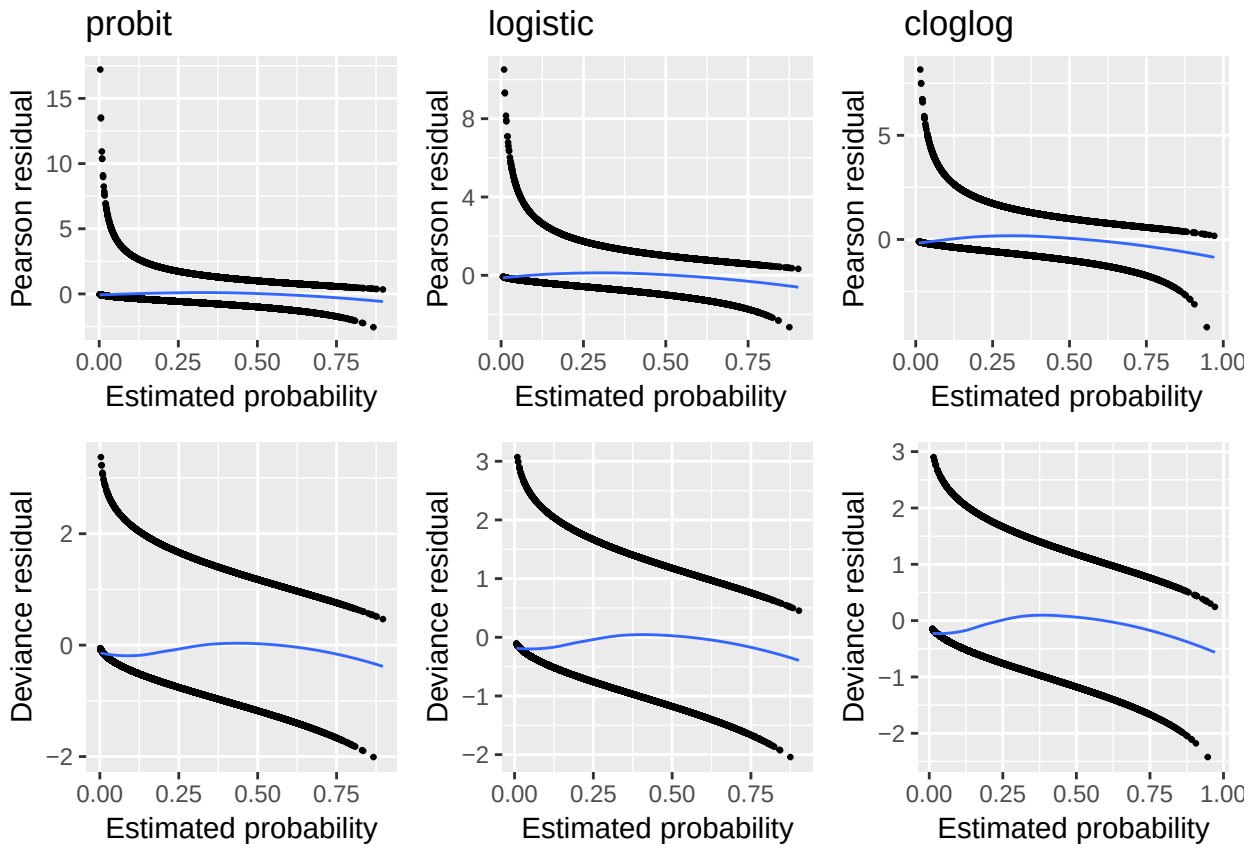
$$\begin{aligned}\Phi^{-1}(\pi) =&\beta_0 + \beta_1(Age) + \beta_2(Sex = 2) + \beta_3(Race = 2) + \beta_4(Race = 3) + \beta_5(WeightLbs)+\\ &\beta_6(HeightIn) + \beta_7(SMOKE = 2) + \beta_8(SMOKE = 3)\end{aligned}$$

Then I utilize stepwise model selection algorithm using AIC as criteria to determine the best subset of the initial model. The initial model itself turns out to be the best subset. Hence, I keep all the predictors.

2.2 Diagnostics

2.2.1 Residual plot

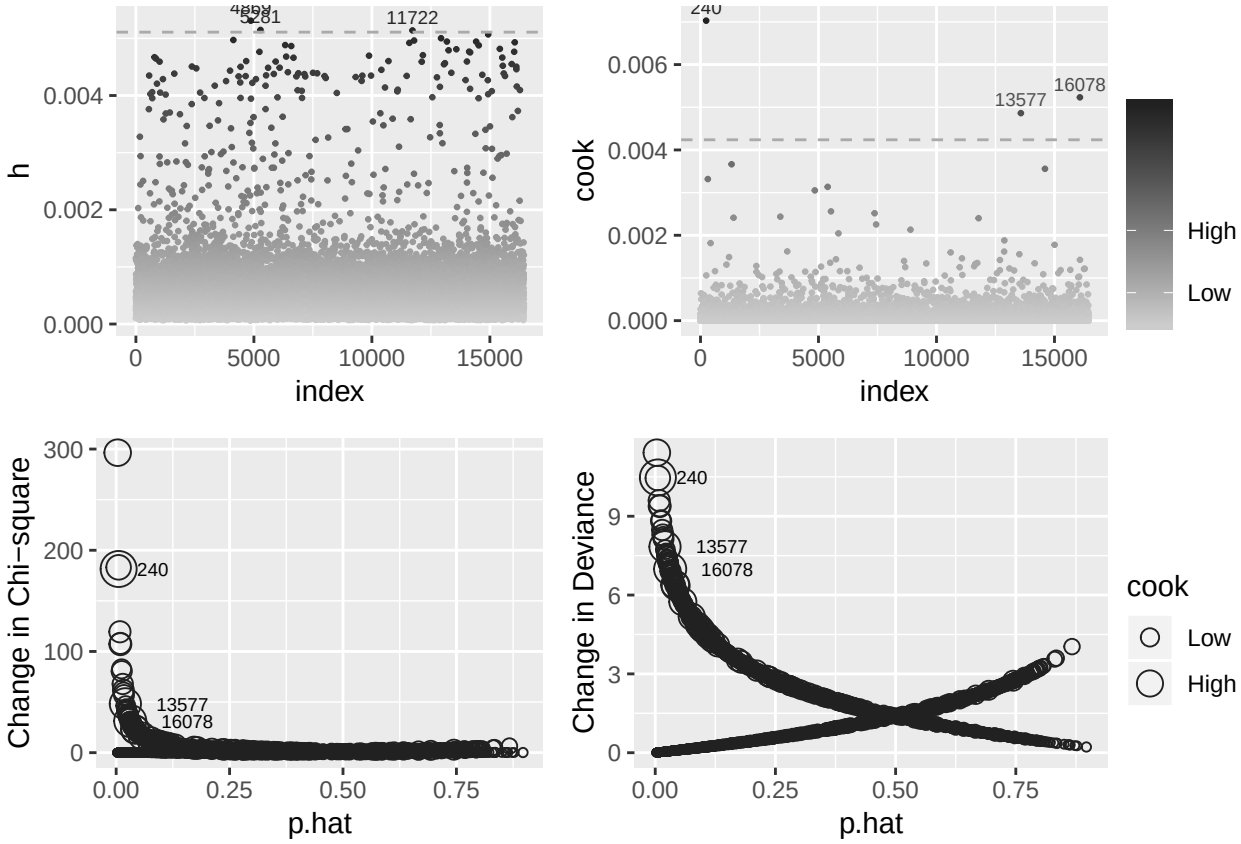
I check the adequacy of the model with residual plots.



The plots confirm that my probit model performs better than the models using logit link and complementary log-log link. The roughly flat lowess line validates this correct model.

2.2.2 Outliers and Influential Cases

Outliers could significantly undermine the quality of a model. Here I try to detect the influential outliers.



The plots of different statistics share a consistent suggestion that the 240^{th} observation is the most influential cases to the model. I remove this observations from the data set to improve the fitness of the model.

2.2.3 Goodness of fit test

After modifying the data, I fit the model again and use Hosmer-Lemeshow test to test the goodness of fit of this model. The p-value of the test is closed to 0, which indicates lack of fit of the data. So adding flexibility to the model could help to improve the model.

X^2	Df	P(>Chi)
85.86	8	3.22e-15

2.3 Interaction terms

It is of interest for this study to evaluate whether the effect of `SMOKE` on `HBP` is the same for different gender, race, or age. A more flexible model that includes interaction term can be considered, especially the sample size is large in this case. To determine whether I should include an interaction term in the model, I implement a Likelihood Ratio Test (LRT) for the full and reduced model:
Reduced: $\Phi^{-1}(\pi) = \beta_0 + \beta_1(Age) + \beta_2(Sex = 2) + \beta_3(Race = 2) + \beta_4(Race = 3) + \beta_5(WeightLbs) + \beta_6(HeightIn) + \beta_7(SMOKE = 2) + \beta_8(SMOKE = 3)$
Full: Reduced model + 2^{nd} order interaction terms between `SMOKE` and other 5 variables.
H_0: No interaction between `SMOKE` and other 5 variables.

H_A: At least 1 of the other 5 variables have interaction with `SMOKE`.

G2	p-value
20.58525	3.39e-05

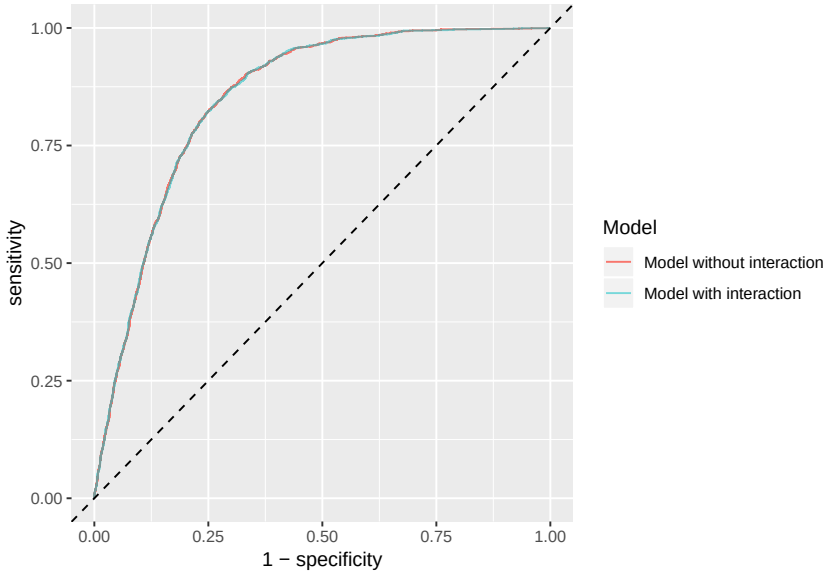
The p-value of LRT is closed to 0, so I reject the null hypothesis and conclude that at least 1 of the other 5 variables have interaction with `SMOKE`. But which variable(s) have interaction with `SMOKE`? I am not sure. So next, I simultaneously run 5 LRTs to detect significant interaction. In each LRTs, I add one interaction term, `SMOKE` with 1 of the other 5 variables, to the same reduced model. Meanwhile, to control the confidence level of the whole family, Bonferroni method is applied. Thus, the significance level for each level is $0.05/5 = 0.01$.

	Age:SMOKE	Sex:SMOKE	Race:SMOKE	log(WeightLbs):SMOKE	HeightIn:SMOKE
G2	0.7056880	6.5748302	8.8215923	2.1024162	5.6481634
p.value	0.7026868	0.0373503	0.0121455	0.3495152	0.0593631

It turns out that all these LRTs have p-values > 0.01. In other words, none of the interaction terms is significant when other terms are in the model. But to keep the consistency with the previous LRT, I retain the interaction term `(Race)*(SMOKE)`, which has the lowest p-value closed to 0.01.

2.4 Validation

Given the large sample size,it is reasonable to try cross validation to test the performance of the new model with interaction term. I randomly seperate the data into 2 equal size sets, training set and testing set. Then I fit the model in the trainging set, and measure the performance of the model in the testing set. I use the area under ROC curve as the measurement.



	Model without interaction	Model with interaction
AUC	0.8512489	0.8512416

The ROC curve of two models under validation method are identical. Also, their AUC, area under the curve, is also very closed to each other. That means, the addition of the interaction term does not make contribution to the performance of the model. To avoid the redundancy of the model, it is advisable to drop this interaction term.
As the result, the finalized model is the same as the initial model:

$$\begin{aligned}\Phi^{-1}(\pi) =&\beta_0 + \beta_1(Age) + \beta_2(Sex = 2) + \beta_3(Race = 2) + \beta_4(Race = 3) + \beta_5(WeightLbs)+ \\ &\beta_6(HeightIn) + \beta_7(SMOKE = 2) + \beta_8(SMOKE = 3)\end{aligned}$$

3. Conclusion

High blood pressure is associated with smoking. Keeping other factors constant, the people who smoke are more likely to have high blood pressure than those who do not smoke or who used to smoke but stopped. This association does not significantly vary among different age, race, etc.
In the following output, i relevel `SMOKE` as the baseline.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.229	0.3795	-13.78	3.489e-43
Age	0.04525	0.000829	54.58	0
Sex	0.08586	0.03671	2.339	0.01935
Race2	0.2761	0.02996	9.216	3.075e-20
Race3	0.05636	0.08141	0.6924	0.4887
I(log(WeightLbs))	0.8241	0.06577	12.53	5.034e-36
HeightIn	-0.03498	0.005112	-6.842	7.797e-12
SMOKE1	-0.09466	0.03423	-2.765	0.005686
SMOKE2	-0.1241	0.03721	-3.336	0.0008491

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	16877 on 16439 degrees of freedom
Residual deviance:	12375 on 16431 degrees of freedom