Investigation of High Blood Pressure using Logistic Regression

Xiaoyun Qin (xq2189)

Summary

The goal of this project is to study the effect of different factors of human including their weight, age, sex, race, body weight and whether they are smoking in whether they get a high blood pressure. Moreover, we will give suggestions to prevent high blood pressure based on the results we got.

Our dataset also has a unique feature since it also includes the survey information: the PSU (SDPPSU6), the stratum (SDPSTRA6) and the statistical weight (WTPFHX6). Specifically, stratum or stratification is the process by which the sampling frame is divided into subgroups or strata that are as homogeneous as possible using certain criteria. Within each stratum, the sample is designed and selected independently. And the PSUs are typically census enumeration areas. Statistical weights are adjustment factors applied to each case in tabulations to adjust for differences in probability of selection and interview between cases in a sample. (Croft, Trevor N., Aileen M. J. Marshall, Courtney K. Allen) We will build a design-based model to deal with these three variables as response variable using svyglm function in R. At the same time, we will also build a model without considering the survey information by assuming the data are collected from a simple random sample.

Both models show that when age, body weight and serum cholesterol level increase, the probability of getting high blood pressure will increase. While taller people has less chance of getting high blood pressure. The chance of getting High blood pressure are also different by sex and race. Smoking level is not a significant factors of high blood pressure when taking the survey information into consideration while it has significant influence on high blood pressure if we assume the data is from a simple random sample. This indicate the necessary to consider the survey design (stratification and clustering) and statistical weight to get a more accurate result.

Introduction

High blood pressure is prevalent no matter in developing or developed country. It is a long-term medical condition in which the blood pressure in the arteries is persistently elevated. Although high blood pressure does not have significant symptoms, long term high blood pressure will cause severe problems including stroke, heart failure and kidney disease. Our analysis is important because we will find the exact probability change when we increase one unit in weight, age or height. More importantly, we will find the influence of smoking according to whether they have smoked and whether they smoked over 100 cigarettes by fitting a logistic regression.

We selected the data from National Health and Nutrition Survey in which 16 variables are included. Among these variables, we choose whether people having high blood pressure or not (HBP) to be our response variable to study the risk factors of high blood pressure. The dataset also contains the information about subjects' age, sex, race, weight, height, serum cholesterol, whether one has smoked over 100 cigarettes in their life and whether one smoking now.

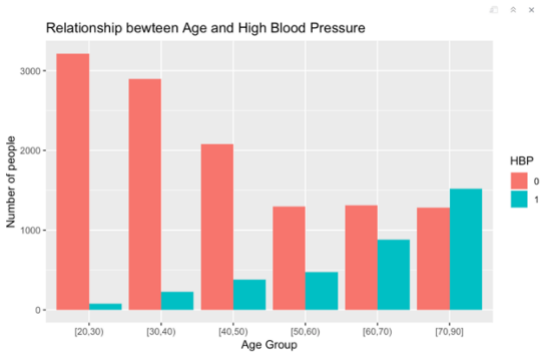
Data Description

We use the average systolic BP and the average diastolic BP are the indicator of high blood pressure. We define our response as a binary response variable. If a subject's average systolic BP is higher than 140 or the average diastolic BP is higher than 90, we will define this subject have high blood pressure, otherwise we define this subject as not having high blood pressure.

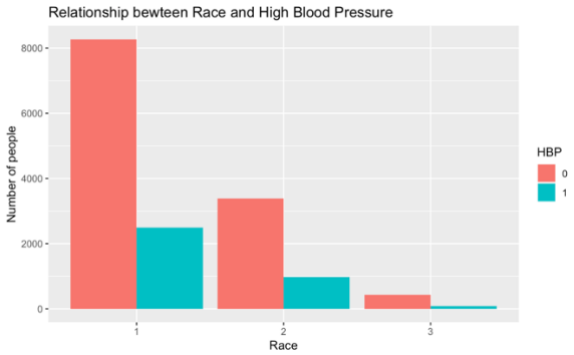
Our covariate includes both numerical variables and categorical variables. Numerical variables include body weight in pounds (BMPWTLBS), standing heights in inches (BMPHTIN), serum cholesterol (TCP) in mg per 100 ml and age in years (HSAGEIR). Categorical variables include sex (HSAGEIR), race (HSSEX), whether one has smoked over 100 cigarettes in their life and whether one smoking now (SMOKE).

Our dataset also contains the survey information: the PSU (SDPPSU6), the stratum (SDPSTRA6) and the statistical weight (WTPFHX6) because it is not from a simple random sample. The dataset has over 17,000 subjects' records to represent over 177 million adults in the United State at that time and each subject has a weight to show how much population they represent. In our analysis, we will take these weights into consideration.

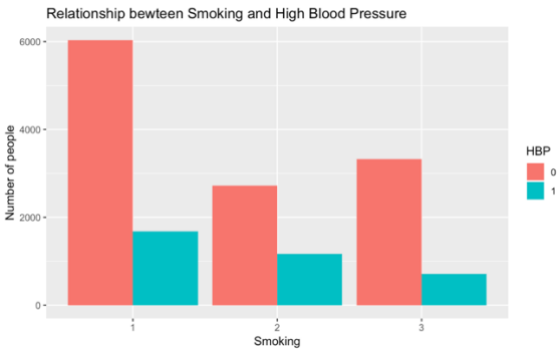
Exploratory Data Analysis



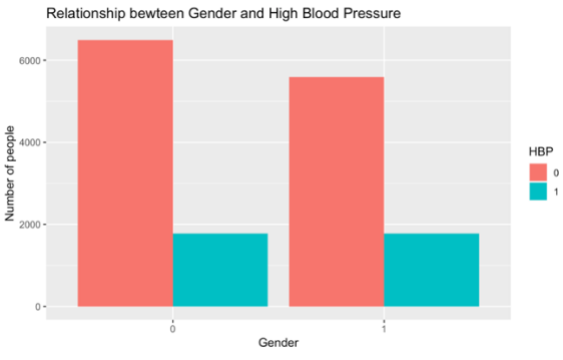
Graph 1. Relationship between Age and HBP



Graph 2. Relationship between Race and HBP



Graph 3. Relationship between Smoking Levels and HBP



Graph 4. Relationship between Gender and HBP

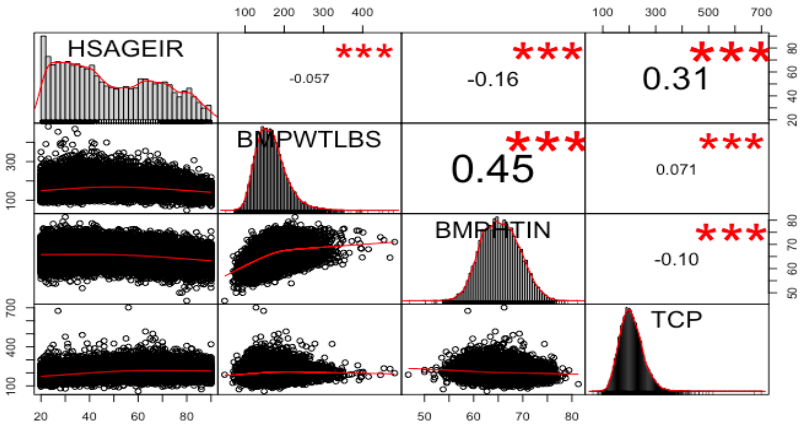
Finding 1: The range of subjects' age is from 20 to 90. We cut the age range into 6 subgroups and it is obvious that older people have a higher chance of getting high blood pressure. (See graph 1)

Finding 2: There are three race groups: white (1), black (2), other (3). 68.83% subjects are white people, 27.9% are black and 3.27% are people from other race. The proportion of high blood pressure respondents among three race are close to each other, which shows it may not

have significant difference between race in the chance of having high blood pressure. (See graph 2)

Finding 3: SMOKE has three levels: level 1 corresponding to people smoked less 100 cigarettes, level 2 corresponding to people abandon smoking but has smoked over 100 cigarettes, level 3 corresponding to people continue to smoke. 49.33% subjects are in level 1, 24.86% subjects are in level 2 and 25.81% subjects are in level3. The proportion of not getting high blood pressure is particular high in level 1. Comparing with the people in level 2, people in level 3 have a larger proportion of having high blood pressure. (See graph 3)

Finding 4: Among the subjects, 52.92% are male and 47.08% are female. Male seem to have higher chance of having high blood pressure. (See graph4)



Graph 5. Distribution of continuous covariate

Finding 5: The distribution of each variable is shown on the diagonal. The distribution of standing height is very close to normal distribution. The distribution of body weight and serum cholesterol are left skewed. There is no significant trend of age distribution. The bivariate scatter plots with a fitted line are distributed on the bottom of the diagonal. The scatter plot of age and body weight, age and standing height, age and serum cholesterol, body weight and serum cholesterol, standing height and serum cholesterol show no trend between the two variables. But the scatter plot between body weight and standing height is not random, which implies that the two variable are correlated. We decided not to remove one of the two variables, because using one of the two variables couldn't represent all of the information of respondents' standing height and body weight. The value of the correlation and the significance level as stars are on the top of diagonal. We decided keep to use this information because the correlation values are not very high. Although all of the correlation levels are significant, this can be due to our large data size. (See graph 5)

Finding 6: We find several covariates in the model is not linear in logit. We used the transformation method that suggested by Hosmer, D.W. and Lemeshow in the applied logistic regression textbook. In detail: Graph6 shows that logit^{-1} (HSAGEIR) is not in sigmoid function(S-shape) and appropriate transformation is to include two terms: HSAGEIR + HSAGEIR^3 (see graph7). Graph 8 shows that logit^{-1} (BMPWTLBS) is not in sigmoid function(S-shape) and the appropriate transformation is to include: Log (BMPWTLBS) and then scale this variable (see graph9). Graph 10 shows that logit^{-1} (BMPHTIN) is in sigmoid function(S-shape) so there is no need to do transformation. Graph 11 shows that logit^{-1} (TCP)

is not in sigmoid function(S-shape) and the appropriate transformation is to include log (TCP) and then scale this variable (see graph 12 in Appendix 1). (Hosmer, D.W. and Lemeshow, S)

Analysis

Statistical analyses of survey data that take the survey design (stratification and clustering) and statistical weights into consideration are generally called 'design-based'. When such features are ignored and the data are handled as if the arose from a simple random sample, the resulting statistical analyses are turned 'model-based'. We will compare the results generated from the design-based and model-based procedures.

1. Build logistic regression with survey information

The final estimated model with survey information is

$$\text{Logit}(\pi_{\text{HBP}}) = -2.1 + 0.66\widehat{\text{Age}} + 0.56 * \text{I(Black race)} + 0.17 * \text{I(Not Black or White)} + 0.52\widehat{\text{Weight}} - 0.25\widehat{\text{Height}} + +0.41 * \text{I(Sex)} + 0.31\widehat{\text{TCP}}$$

Where:

$$\widehat{\text{Age}}: (\frac{\text{Age} - \overline{\text{Age}}}{\sigma_{(\text{Age})}})^3 + \frac{\text{Age} - \overline{\text{Age}}}{\sigma_{(\text{Age})}}, \text{with } \sigma_{(\text{Age})} = 19.23 \text{ here}$$

$$\widehat{\text{Weight}}: \frac{\log(\text{weight}) - \overline{\log(\text{weight})}}{\sigma_{\log(\text{weight})}}, \text{with } \sigma_{\log(\text{weight})} = 0.228 \text{ here}$$

$$\widehat{\text{height}}: \frac{\text{height} - \overline{\text{height}}}{\sigma_{\text{height}}} \quad \text{with } \sigma_{\text{height}} = 3.88 \text{ here}$$

$$\widehat{\text{TCP}}: \frac{\log(\text{TCP}) - \overline{\log(\text{TCP})}}{\sigma_{\log(\text{TCP})}}, \text{with } \sigma_{\log(\text{TCP})} = \quad 0.215 \text{ here}$$

The information about estimates are as follow:

	Estimate	Std. Error	Odds.Ratio	Pr(>|t|)	OR.Lower.95	OR.Upper.95
(Intercept)	-2.1005	0.0599	0.1224	0	0.1088	0.1377
HSAGEIR	0.6602	0.0206	1.9351	0	1.8586	2.0147
DMARACER2	0.5695	0.0798	1.7674	0	1.5114	2.0667
DMARACER3	0.1714	0.2508	1.1869	0.4982	0.726	1.9403
BMPWTLBS	0.5178	0.0363	1.6784	0	1.5631	1.8021
BMPHTIN	-0.2473	0.0522	0.7809	0	0.705	0.865
HSSEX1	0.4083	0.0956	1.5043	0.0001	1.2474	1.8141
TCP	0.3109	0.0416	1.3646	0	1.2577	1.4806

Table 1. The results of model with survey information

The meaning of each covariates is illustrating in table 2 as below:

Covariates	Meaning
HSAGEIR	Our transformed Age variable in years
DMARACER2	Dummy variable for black people
DMARACER3	Dummy variable for not black or white people
BMPWTLBS	Our transformed weight in pounds
BMPHTIN	Our transformed standing height in inch
HSSEX1	Dummy variable for male.
TCP	Serum Cholesterol in mg per 100 ml

Table 2. The meaning of covariate

The results of analysis show that most of predictor variables, age, sex, gender, race, height, weight and serum cholesterol level have significant influence on response variable, using a significance level of 0.05. Smoking level is not significant using any usual level of confidence (0.1, 0.05, 0.01) as you can see in appendix 2 Table 6, so we exclude this factor. A Wald test statistics table is also attached in the appendix 2 Table 5. Here is the interpretation of each variable in detail.

Notes: After we scale the continuous variables, different variables become comparable on their effect on HBP.

Also, If we took log for one or more of our covariate and holding other covariate the same, we will have: $\text{logit}(\pi(x_2)) - \text{logit}(\pi(x_1)) = \beta * \log(\frac{x2}{x1})$, with that to say, as long as the percent increase in the predictor variable is fixed, we will see the same difference in logit of probability, regardless where the baseline is. For example, we can say that for a 10% increase in TCP, the difference in the logit of HBP will be always $\beta * \log(1.1)$.

If we scale after log, just as what we do for TCP and weight, we can say that when we increase one standard deviation of log(x), the probability will changed by $\frac{e^{\beta}}{1+e^{\beta}}$. Or in other words, $\log(x_2) - \log(x_1) = \sigma \Leftrightarrow x_2 = x_1 * e^{\sigma}$, when we increased x by (e^{σ} -1), the probability will changed by $\frac{e^{\beta}}{1+e^{\beta}}$.

The results of analysis show that an increasing in age will increases a person's probability of having high blood pressure problem in a nonlinear trend. That is to say, when we fix all other factors, a unit increase in people's age will increase the probability of getting HBP relatively smaller when people are young compared to they getting elder. High blood pressure is the most common chronic condition among older adults. Elder are exposure to higher risk of having high blood pressure and we recommend people to keep a regularly physical examination as they age.

Race affect the odds of people having high blood pressure. As we can see from the results, the difference between white and black people in having HBP are significant. The odds of black people having high blood pressure problem is a number between 51% to 106% higher than white people having the same problem when holding other factors fixed. However, at any usual level of confidence (0.1, 0.05, and 0.001), the odds of having high blood pressure is not statistically different between white people and people neither as White nor Black. We will need to do more research on why black people having higher risk of having high blood pressure.

Obesity contributes to high blood pressure. When a person's log weight increase by one standard deviation (0.228) while holding other factors fixed, in other words, when weight increased by $e^{0.228}$-1 = 25.6% percent, the odds of having high blood pressure change by a multiplicative factor between 1.56 to 1.8, which means the odds would increases a number between 0.56 and 0.8. Taller person has less opportunity of getting HBP. We are 95% confident to say that when we increase one standard deviation of height (3.88 inches), the odds of getting HBP will decrease by a number between 13% to 29% while other holding other predictor factors same.

Combined with results from height and weight, it shows that we need to control our weight to reduce the probability of having high blood pressure. This meet our expectation that overweight is unhealthy and we need to control our weight to have a better life. Under the situation that we are unable to get taller when we are adult, we could exercise more to control our weight to help us stay away from hypertension.

Gender has a significant influence on the probability of people having HBP. We have 95% confidence to say that the estimated odds of a male having high blood pressure is a number between 1.24 and 1.81 times the estimated odds of a female having high blood pressure while other predictor factors are same. This shows male are more likely to have hypertension than female and they should pay more attention to their blood pressure.

Serum Cholesterol has significant influence the odds of having HBP as well. $e^{0.215}$-1 = 24% increase in serum, the odds of having high blood pressure will increase by a number between 26% and 48% given they have the same status on other factors.

Goodness-of-Fit

We use the prediction accuracy to evaluate our model performance. The accuracy of prediction on the whole data set is 78.86%, which shows our model is reasonable to the data.

2. Build logistic regression without survey information

The estimated model without survey information is:

$$\begin{aligned}\text{Logit}(\pi_{\text{HBP}}) = &-1.99 + 0.34\widehat{\text{Age}} + 0.31 * \text{I(Sex)} + 0.43 * \text{I(Black race)} + 0.082 \\ &* \text{I(Not Black or White)} + 0.37\widehat{\text{Weight}} - 0.24\widehat{\text{Height}} + 0.086 \\ &* \text{I(Smoke more than 100 cigarettes but quit smoking)} + 0.1417 \\ &* \text{I(Smoke more than 100 cigarettes and still smoking)} + 0.33\widehat{\text{TCP}}\end{aligned}$$

Where $\widehat{\text{Height}}, \widehat{\text{Weight}}, \widehat{\text{TCP}}, \widehat{\text{Age}}$ are transformed in a completely same way as previous model.

Without survey information:

	Estimate	Std. Error	Odds.Ratio	Pr(>|t|)	OR.Lower.95	OR.Upper.95
(Intercept)	-1.9899	0.0488	0.1367	0	0.1242	0.1504
HSAGEIR	0.3398	0.0081	1.4046	0	1.3826	1.427
HSSEX1	0.3093	0.0617	1.3625	0	1.2073	1.5377
DMARACER2	0.4252	0.0501	1.5298	0	1.3866	1.6878
DMARACER3	0.0818	0.1352	1.0852	0.5452	0.8326	1.4144
BMPWTLBS	0.3744	0.0256	1.4541	0	1.383	1.5288
BMPHTIN	-0.2379	0.0333	0.7883	0	0.7384	0.8415
SMOKE2	0.0855	0.0532	1.0892	0.1078	0.9815	1.2088
SMOKE3	0.1417	0.0571	1.1523	0.0131	1.0302	1.2888
TCP	0.3251	0.0231	1.3841	0	1.323	1.4481

Table 3. The results of model without survey information

The meaning of each covariate is illustrating in table 4 as below:

Covariates	Meaning
HSAGEIR	Our transformed Age variable in years
DMARACER2	Dummy variable for black people
DMARACER3	Dummy variable for not black or white people
BMPWTLBS	Our transformed weight in pounds
BMPHTIN	Our transformed standing height in inch
SMOKE2	Smoke more than 100 cigarettes in life and stop smoking
SMOKE3	Smoke more than 100 cigarettes in life and still smoking
HSSEX1	Dummy variable for male.
TCP	Serum Cholesterol in mg per 100 ml

Table 4. The meaning of covariate

In a very similar way for another model while smoking level has significance influence on high blood pressure when we don't take survey information into consideration. We use the deviance anova test to test if the influence of smoking level is significance and the results shows that we should reject the null hypotheses and conclude that smoking levels have significance influence in high blood pressure. (See Table 7) A Wald test statistics table is also attached in the appendix 2 Table 8. We can see that:

The results of analysis show that age has a nonlinear effect on increasing a person's probability of having high blood pressure problem as previous model.

The odds of black people having high blood pressure problem is a number between 38% to 68% higher than white people having the same problem when holding other factors fixed. However, at any usual level of confidence (0.1, 0.05, and 0.001), the odds of having high blood pressure is not statistically different between white people and people neither as White nor Black.

Gender has a significant influence on the probability of people having HBP. We have 95% confidence to say that the estimated odds of a male having high blood pressure is a number between 1.21 and 1.54 times the estimated odds of a female having high blood pressure while other predictor factors are same.

When a person's log weight increase by one standard deviation (0.228) while holding other factors fixed, in other words, when weight increased by $e^{0.228}$-1 = 25.6% percent, the odds of having high blood pressure change by a multiplicative factor between 1.38 to 1.52, which means the odds would increases a number between 0.38 and 0.52. Besides, we are 95% confident to say that when we increase one standard deviation of height (3.88 inches), the odds of getting HBP will decrease by a number between 16% to 27% while other holding other predictor factors same.

When we assume that the data is from simple random sample, smoke or tobacco use might be considered a risk factor. Compared to people who smoke less than 100 cigarettes in life, the odds of having high blood pressure in people who smoked over 100 cigarettes in life and still smoke now increase a number between 3% and 28.9% while holding other factors fixed. The difference in odds of people having high blood pressure is not significance between people who smoke less than 100 cigarettes in life and people who smoke over 100 cigarettes in life but quit smoke now.

3. Comparison of Two Methods

In both cases, the sex, race, height, weight and TCP are all significant in 0.05 level. After including the survey information, in which we considering the probability of each person be chosen in their own population, the significance of smoke status become less important in determining their probability of getting HBP as can be seen on the table above. In a 0.05 level, the smoke factor is significant in the model without survey information while it is not in the model with survey information. While these survey information is necessary if the survey was done in a significant different region with significant different populations.

Conclusion

In this research, we study the factors that may influence high blood pressure using a dataset that contains over 17,000 records that contains the information about people's age, sex, race, height, weight, smoke status, average systolic blood pressure, average diastolic blood pressure, and serum cholesterol. This analysis concluded with a high statistical significance that the age, sex, race, height, weight, and serum cholesterol influence the odds of people having high blood pressure. While we conclude that further research need to be done in whether smoking is a significant factor in high blood pressure.

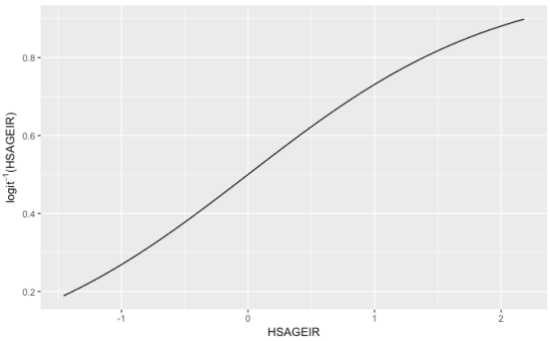
We also have some suggestions for people who have such concerns in getting high blood pressure. Male, black, elder and overweight people should concern more with their healthy, especially on the blood pressure. We should all pay attention to our serum cholesterol level. Although smoking does not have significant effect on getting HBP in our survey information model, we should still avoid it since it still has an overall negative effect on our health and it will increase 2% in probability of getting HBP.

Reference

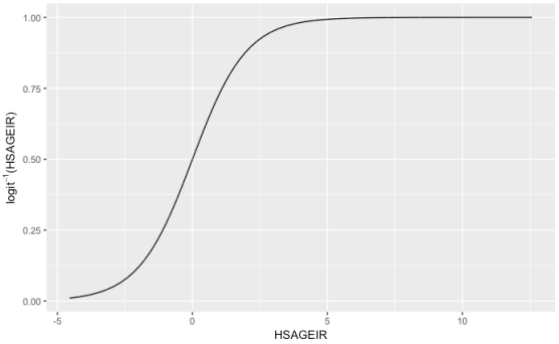
[1]Croft, Trevor N., Aileen M. J. Marshall, Courtney K. Allen, et al. 2018. Guide to DHS Statistics. Rockville, Maryland, USA: ICF.

[2]Hosmer, D.W. and Lemeshow, S. (2000) Applied logistic regression. 2nd Edition, John Wiley & Sons, Inc., New York

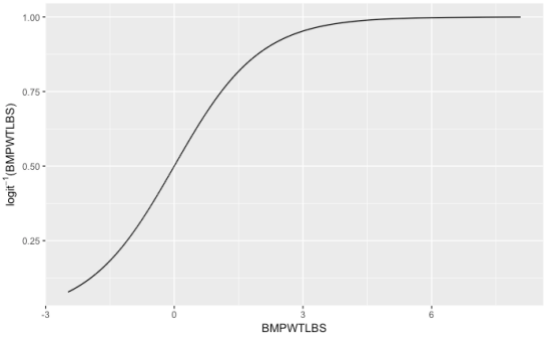
Appendix 1 - Graphs in Exploratory Data Analysis Part



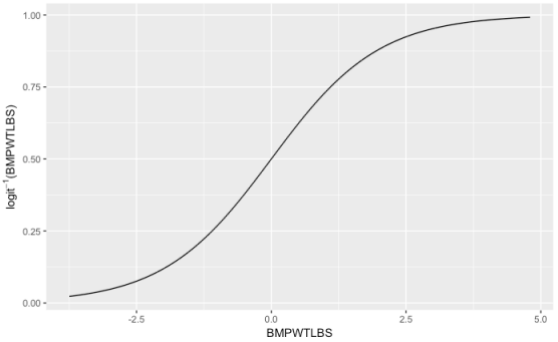
Graph 6 Age before transform



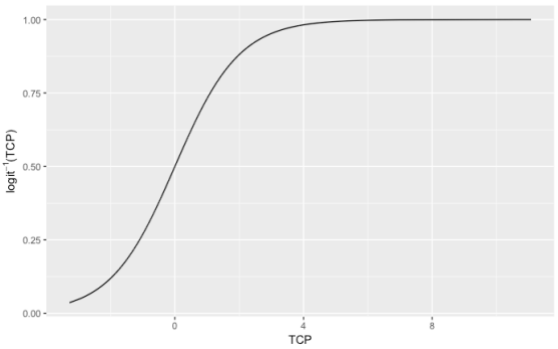
Graph 7 Age after transform



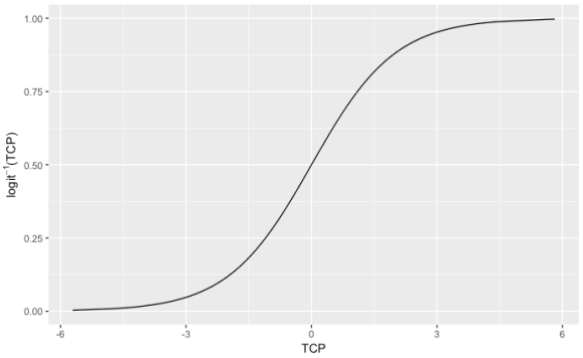
Graph 8 Weight before transform



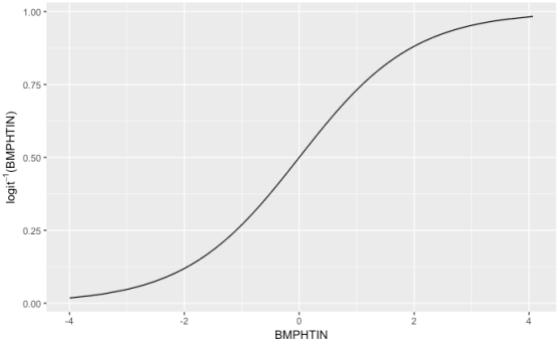
Graph 9 Weight after transform



Graph 10 Height before transform



Graph 11 TCP before transform



Graph 12 TCP after transform

Appendix 2 - Tables in Analysis Part

	Df	Chisq	Pr(>Chisq)	
HSAGEIR	1	1133.5257	2.20E-16	***
DMARACER	2	48.0954	3.60E-11	***
BMPWTLBS	1	231.3662	2.20E-16	***
BMPHTIN	1	22.929	1.68E-06	***
HSSEX	1	19.5558	9.77E-06	***
TCP	1	56.6775	5.14E-14	***
SMOKE	2	1.7917	0.4083	

Table 5. Wald test for model with survey information

	Estimate	Std. Error	Odds.Ratio	Pr(>|t|)	OR.Lower.95	OR.Upper.95
(Intercept)	-2.0904	0.0709	0.1236	0	0.1076	0.1421
HSAGEIR	0.6677	0.0198	1.9498	0	1.8754	2.027
DMARACER2	0.5565	0.0807	1.7446	0	1.4894	2.0436
DMARACER3	0.1643	0.2455	1.1786	0.5071	0.7284	1.907
BMPWTLBS	0.5253	0.0345	1.6909	0	1.5803	1.8094
BMPHTIN	-0.2478	0.0518	0.7805	0	0.7052	0.8638
HSSEX1	0.4201	0.095	1.5221	0.0001	1.2635	1.8337
TCP	0.3121	0.0415	1.3662	0	1.2596	1.4819
SMOKE2	-0.093	0.094	0.9112	0.3283	0.7578	1.0955
SMOKE3	0.0425	0.0979	1.0434	0.6662	0.8613	1.2641

Table 6. The results of model with survey information with SMOKE

	Resid.Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	15635	13080				
2	15633	13070	2	10.413	0.00548	**

Table 7. Deviance Anova Test of Smoke

	Df	Chisq	Pr(>Chisq)	
(Intercept)	1	1663.4136	2.20E-16	***
HSAGEIR	1	1774.5506	2.20E-16	***
HSSEX	1	25.132	5.35E-07	***
DMARACER	2	71.9668	2.36E-16	***
BMPWTLBS	1	214.1983	2.20E-16	***
BMPHTIN	1	50.9242	9.60E-13	***
SMOKE	2	6.6828	0.03539	*
TCP	1	198.8465	2.20E-16	***

Table 8. Wald test for model without survey information