Group 2 Coronary Heart Disease Final Project

Andrea Chung and Kelly Wentzlof

2022-12-11

Introduction

Cardiovascular disease, also known as heart disease, is a group of heart conditions that include diseased vessels, structural problems, and blood clots. It is the leading cause of death among men and women in the United States. Through this research, we will be looking at coronary heart disease (CHD), also known as coronary artery disease (CAD). CHD is the most common type of heart disease and often leads to a heart attack. This type of heart disease occurs when plaque builds up in the walls of the arteries that supply blood to the heart. Due to the plaque buildup, the coronary arteries begin to narrow and eventually be unable to transport enough oxygen-rich blood to the heart.

Previous research have found several factors that may increase one's chances of getting CHD including: unhealthy eating habits, smoking, high systolic blood pressure, old age, family history of CHD, low density lipoprotein cholesterol levels, and performance of type A behaviors. While there are many factors that can lead to CHD, often people with CHD do not show any signs or symptoms of the disease until they have a heart attack. Therefore, it is important to analyze the factors that could lead to CHD as a way to prevent it. By finding a meaningful relationship between risk factors and CHD, it can help future health care providers with prognosis and prevention of CHD. For these reasons, in this project, we intend to answer the following research question:

- What is the relationship between type A behavior and coronary heart disease after accounting cumulative tobacco, systolic blood pressure, and age?

For this project, we will be using data set obtained from Kaggle. The data was taken from a larger data set used in South African Medical Journal in 1983 and consists of sample of 462 male-only observations from a heart disease risk region in South Africa. The data set has a variable which indicates whether the observation has CHD (chd = 1) or did not (chd = 0). Approximately one-third of the sample had CHD.

Additionally, the data set has 9 predictor variables (low density lipoprotein cholestorol, adiposity, family history of heart disease, obesity, alcohol, systolic blood pressure (SBP), cumulative tobacco consumption, type A behavior, and age). After researching CHD, we determined that the main risk factors that are typically highlighted are systolic blood pressure ("sbp"), cumulative tobacco consumption ("tobacco"), and age ("age"). Therefore, we decided to keep these 3 predictive variables, as well as the additional predictor variable, type A behavior ("typea"), and our desired response variable as CHD ("chd").

We are particularly interested in the relationship between type A behavior and CHD because this relationship has been investigated in previous research, however, it has been investigated without consideration of the other main risk factors of CHD. Additionally, previous research has found that type A behaviors only play a role in CHD in certain countries and within certain cultures, so it will be interesting to see the significance of the relationship between this predictor variable and CHD when applied to people in South African region.

Summary of Data

There are total of 462 samples for each variables. Among the total sample population, 34.62% had CHD and 65.37% had no CHD (See Table 1). The average SBP for non-CHD was 135.46 and for CHD was 143.74 resulting in a difference of 8.28 mmHg (See Table 1). The average amount of cumulative tobacco consumed by individuals in their lifetime indicated that there was a large difference between CHD and non-CHD participants. For non-CHD individuals, the average cumulative tobacco consumed was 2.63 kg when the average for CHD was 5.52 kg (See Table 1). Overall, the average tobacco consumption for people with CHD was more than two times the average tobacco consumption of people without CHD. Additionally, there was a big difference in mean age between people with CHD and people who did not have CHD. The average age of non-CHD individuals was roughly 39 and the average age of participants with CHD was about 50 (See Table 1). According to the South African Medical Journal that this data was taken from, people took the self-report Bortner Short Rating Scale to measure type A behavior. The participants who had scores that were on the upper two-fifths of the score range, which is higher than or equal to score of 55, were classified as a person exhibiting type A behavior. The participants whose scores were less than 55 were classified as a person not exhibiting type A behavior. The probability of having type A behavior was higher when one had CHD resulting in a difference of 0.09 (See Table 1).

CHD	Mean SBP	Mean Cum. Tobacco	Mean Age	Mean Prob. Type A Behavior
0	135.46	2.63	38.85	0.42
1	143.74	5.52	50.29	0.51

Table 1: Quantitative Distribution Analysis on Variables According to CHD

Spearman's Correlation Coefficient

When we looked at a pairwise correlation plot of all of the predictor variables and response variable using ggpairs() function, we were able to observe that the variables did not have linear relationships between each other. Therefore, Spearman's rho correlation was more appropriate to use than Pearson's r.

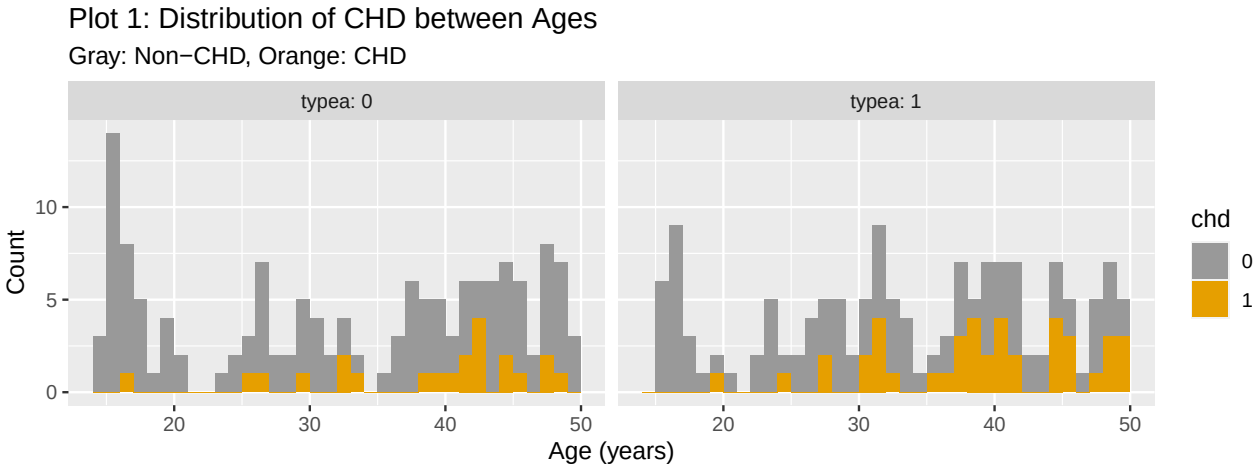
Variables	Correlation Coefficient
Age vs. CHD	0.367
Tobacco vs. CHD	0.323
SBP vs. CHD	0.172
type A vs. CHD	0.082

Table 2: Correlation Coefficients of Variables with CHD

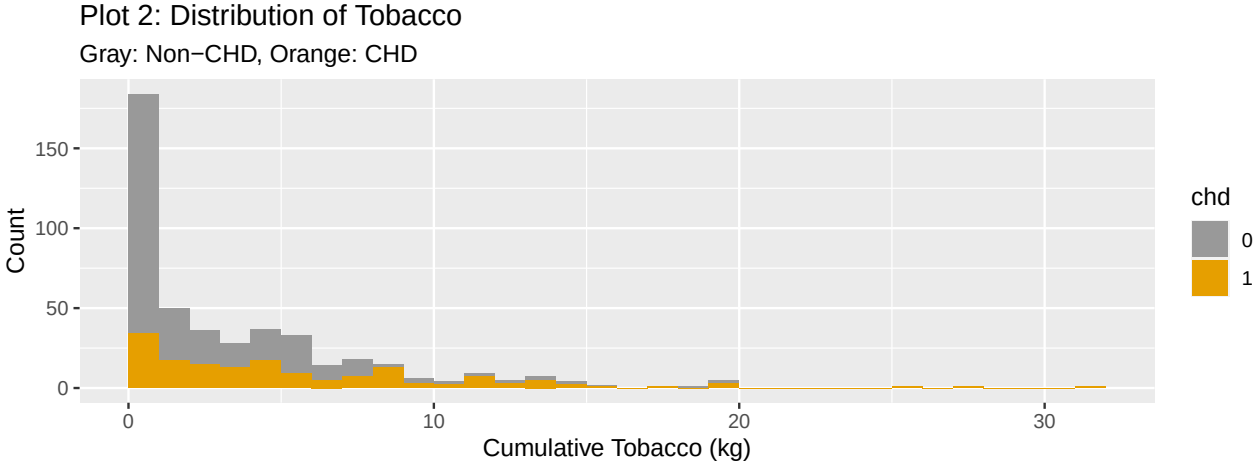
From the table, we can see that all of the predictor variables have a positive correlation with the response variable, CHD. This indicates that (1) older individuals are more likely to have CHD, (2) people who smoke more are more likely to have CHD, (3) people with higher SBP are more likely to have CHD, and (4) people who perform type A behaviors are more likely to have CHD. Also, we can see that age and tobacco appear to have the highest correlation with CHD out of all of the predictors.

Exploring the Data

Age and tobacco seem to have the strongest correlations with the response variable, CHD, so we will take a closer look at the distributions of CHD across these variables.

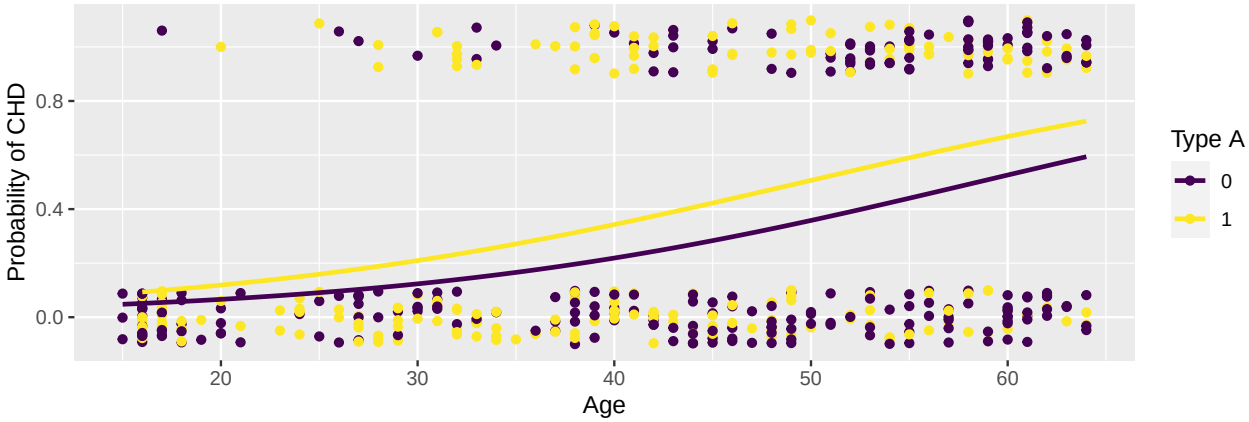


Looking at Plot 1 above, we can observe that, in general, the number of people diagnosed with CHD increases as the age increases. Also, we faceted by type A behavior to look at the distribution further by accounting the type A behavior. Looking at the plot, we can see that there appears to be more people with CHD when they perform type A behavior as indicated by the larger area of the plot that is colored orange.



Looking at plot 2, the distribution is strongly right-skewed, suggesting a transformation. While, the preferred transformation would be a log transformation, we cannot conduct a log transformation because of the large number of people have indicated that they smoke 0 kg of tobacco. For now, we will continue without a transformation. We will focus on our variable of interest, type A behavior.

Plot 3: Probability of CHD across Age and Type A Behavior



To explore the apparent effect of type A behavior on CHD, we looked at jittered plot of CHD across age and type A behavior and plotted logistic regression fits for both people with type A behavior and people who do not conduct type A behavior. We used logistic regression fits due to the binary nature of the response variable, CHD. Looking at the plot above, we can perceive that people who carry out type A behavior are slightly more likely to have CHD than those who do not even after accounting for age. More specifically, it appears that there is a larger difference between two fitting lines after the age of 30. From this we were able to indicate that there is relationship between type A behavior and CHD that should be further investigated.

Modeling

Initial Model

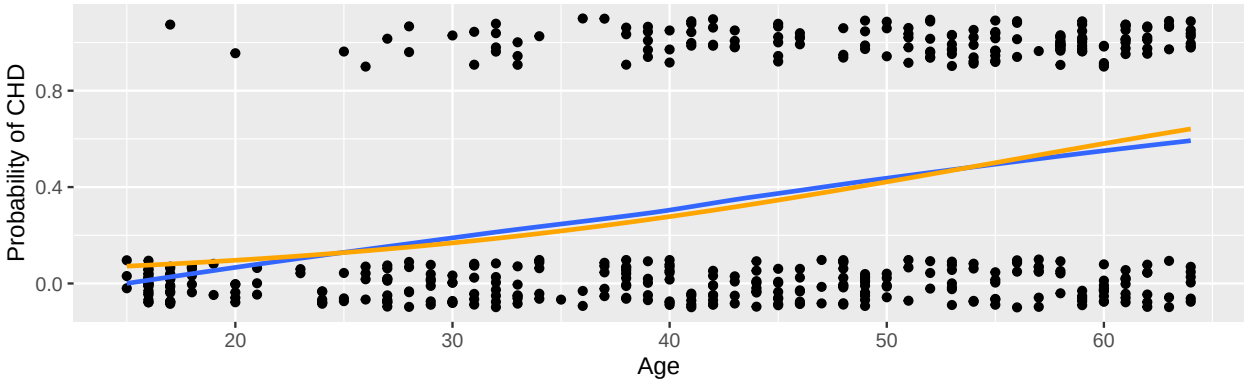
Since age and cumulative tobacco consumption emerged to have the highest correlation with CHD in the exploration of our data set, we started our initial model by only including tobacco and age as predictor variables. First, we plotted both loess curve and logistic regression line to see which method was more appropriate for our model.

CHD vs. Age

We initiated the comparison of loess and logistic regression fits by looking at the model with age as the sole predictor.

Plot 4: CHD vs. Age with Fittings

Blue: Loess, Orange: Logistic Regression



The loess curve fits linear models "locally" to the data, so that the fit at any given point is a regression fit. The logistic regression model allows us to model the probability of an event that has binary levels of response. The two fits seem to follow a similar pattern throughout the entire data set. Using the logistic regression

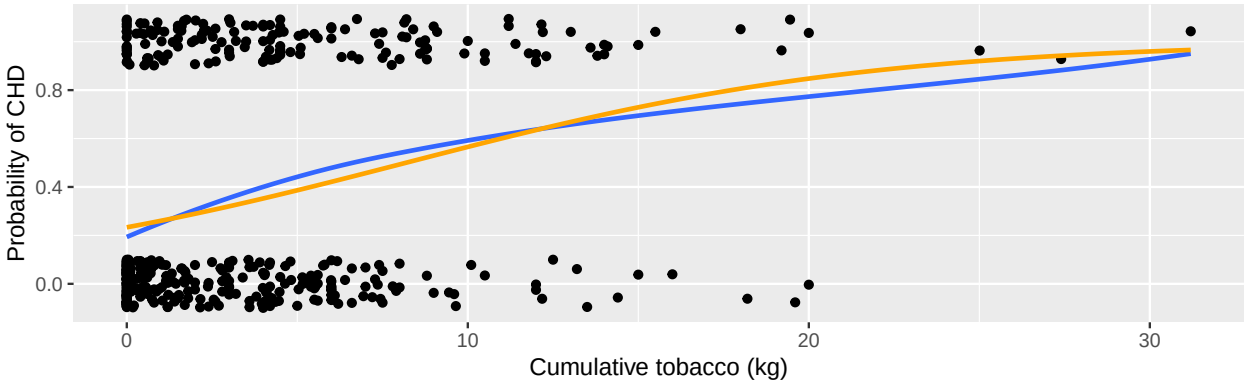
model over the loess model aids in interpretation and creates a smoother appearing model. However, we do lose some specificity in fit because of this smoothing. Also, looking at this plot, there appears to be a positive relationship between age and CHD which can be seen by the line with a positive slope.

CHD vs. Tobacco

For further exploration, we tried the same approach but using tobacco as the sole predictor.

Plot 5: CHD vs. Tobacco with Fittings

Blue: Loess, Orange: Logistic Regression



The loess model and the logistic regression model appear to be relatively similar also. We can see that the loess model appears to be over-fitting around 30 kg of cumulative tobacco where we see the weird bump at the beginning of the plot which is eliminated when using the logistic regression. Thus, we thought that employing the logistic regression may be best in this situation.

Including both Tobacco and Age

After looking at tobacco and age variables individually, we also investigated whether CHD depended on age and tobacco simultaneously.

Plot 6: CHD vs. Tobacco and Age

There is a lack of data in the top left, i.e., we have no observations with age 0-35 and cumulative tobacco greater than 10 kg. This is most likely to do with the fact that younger individuals may not have as much access to tobacco products or their smoking habits did not have enough time to grow to a larger extent. We can see a trend, that someone who is older and who smokes more tobacco is somewhat more likely to be diagnosed with CHD because of there being more orange points in the top right corner, however, we can not distinguish clear pattern where the majority of the data lies.

Since we did not discover big distinction between loess curve and logistic regression curve and spotted possible

over-fitting of loess curve with tobacco variable, we decided to use logistic regression for the model. Due to the binomial nature of CHD response variable in the data set, we assume that logistic regression will be appropriate for our analyses.

As an initial model, we fit an additive model with no interaction between the two predictors and CHD as the response variable: $chd \sim age + tobacco$

Here we have the coefficient estimates from the simple additive model:

Intercept	-3.375
Age	0.054
Tobacco	0.079

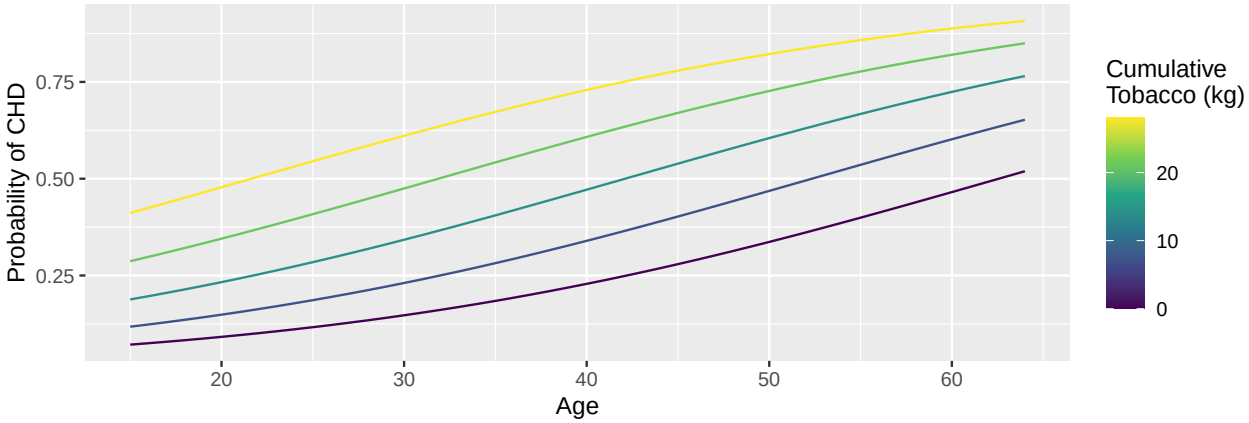
Table 3: Coefficient Estimates of Variables from Initial Model

The model states:

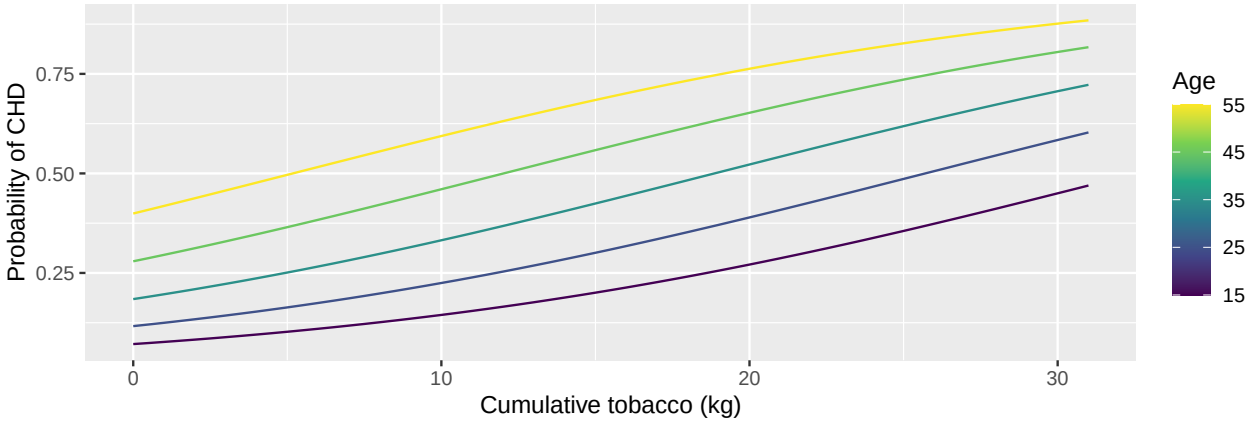
$$logit[P(chd|age, tobacco)] = -3.375 + 0.054 \times age + 0.079 \times tobacco$$

We will visualize the fit by drawing multiple curves representing different values of one of the predictors. For the first plot, we will plot the probability of CHD as a function of age for a few values of tobacco. Then, we will find prediction curves for tobacco for a few different ages.

Plot 7: Probability of CHD vs. Age for Different Amount of Cumulative Tobacco



Plot 8: Probability of CHD vs. Cumulative Tobacco for Different Ages



From both of these plots, we conclude that the older an individual gets and more cumulative tobacco they smoke, the higher the probability is that they get CHD. Additionally, we see that the lines are parallel indicating that no interaction appears to be necessary between age and tobacco when predicting CHD. To

ensure this, we compared akaike information criterion (AIC) scores between two models: model with no interaction and model with interaction between tobacco and age variables. Looking at the AIC scores of the models, we have noted that interaction was not needed.

Model	AIC Score
Model with No Interaction	521.39
Model with Age, Tobacco Interaction	522.62

Table 4: AIC Score Comparison of Two Models

Adding more predictor variables

After we determined the basis of our initial model which was an additive model with the main predictors, age and tobacco, we added the other predictor variables, type A behavior and SBP.

Intercept	-4.525
Age	0.055
Tobacco	0.078
SBP	0.006
Type A = 1	0.613

Table 5: Coefficient Estimates of Variables with SBP and Type A added to Initial Model

All of these predictors have positive coefficient estimates which aligns with previous research on CHD indicating that the increase of all these variables increases the likelihood of CHD.

AIC Comparison (Two-Way Interaction)

After adding more predictor variables, we used AIC to determine whether we should include these variables in our model and whether there exists any relevant and useful two-way interactions between SBP and type A behavior. Three-way and other higher-order interactions were not considered because they would result in a model that is extremely difficult to interpret and to fit.

Model	AIC Score
Model with No Interaction	516.68
Initial Model + 'age:sbp'	518.61
Initial Model + 'age:typea'	518.68
Initial Model + 'tobacco:sbp'	518.68
Initial Model + 'tobacco:typea'	518.36
Initial Model + 'tobacco:sbp' + 'sbp:typea'	519.30

Table 6: AIC Score Comparison of Different Models

Looking at the table above (Table 6), we found that all two-way interactions were unnecessary and did not help the model perform better. However, we were able to note that adding SBP and type A is helpful since the AIC score of model with no interaction with SBP and type A variables added was lower than model with only tobacco and age, which was 521.3855 (Table 4).

Moreover, we tried fitting a GAM model with our variables and data to see if adding smooth terms improved our model.

Model	AIC Score
Logistic Regression Model with no Interaction	516.68
GAM with s(age)	517.05
GAM with s(tobacco)	516.68
GAM with s(age) and s(tobacco)	517.06

Table 7: AIC Comparison between Initial Model and GAM Model

The table above (Table 7) shows the comparisons of AIC scores between the models. We can observe that smooth terms on the two main predictors did not improve our model. Thus, we decided to continue with logistic regression model including the four predictor variables with no interaction as our final model.

Final model and Fits

Final model

The final model was an additive logistic regression model containing age, tobacco, systolic blood pressure, and type a behavior:

$$logit[P(chd|age, tobacco, sbp, typea)] = -4.525380 + 0.055055 \times age + 0.077523 \times tobacco + 0.612987 \times typea$$

Intercept	-4.525
Age	0.055
Tobacco	0.078
SBP	0.006
Type A = 1	0.613

Table 8: Coefficient Estimates of Variables from Final Model

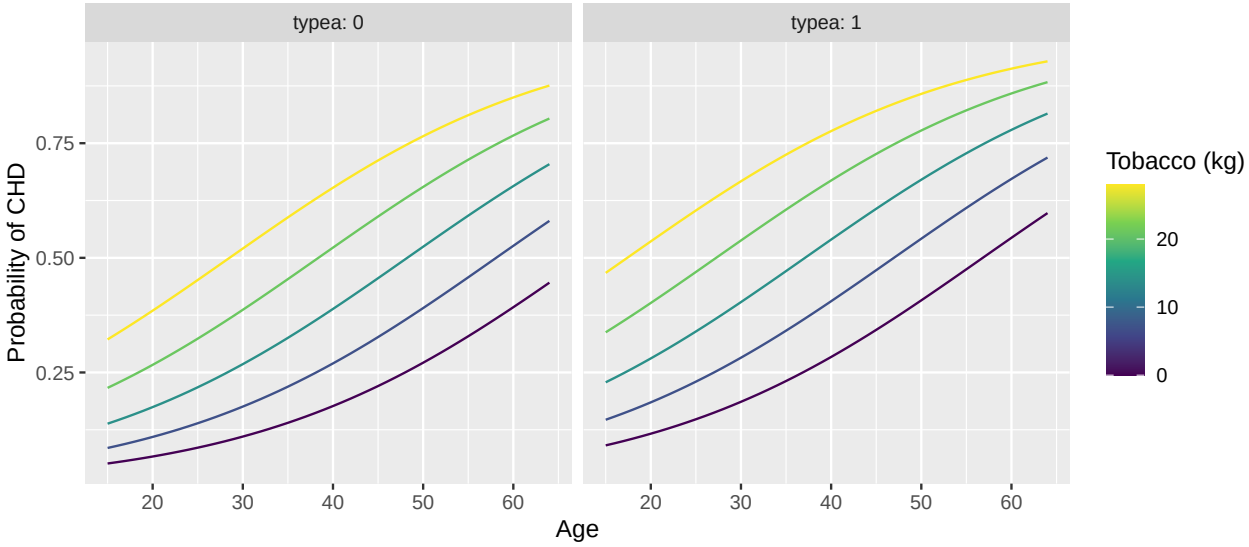
All of the estimations are positive indicating that the older the individual is, the more tobacco they consume, the higher their systolic blood pressure is, and if they conduct type A behavior, the more likely it will be that the individual has CHD. Also, it is important to note the magnitude of the type A behavior predictor variable coefficient compared to the other predictors.

Fits of the model

Let's take a look at the fit of this model. First, we will fix SBP at it's median because SBP does not appear to have a high correlation with CHD nor does it appear to have a large effect within our current model. Then, we will draw tobacco curves for a few values of age.

Plot 9: Probability of CHD vs. Age accounting Tobacco and Type A Behavior

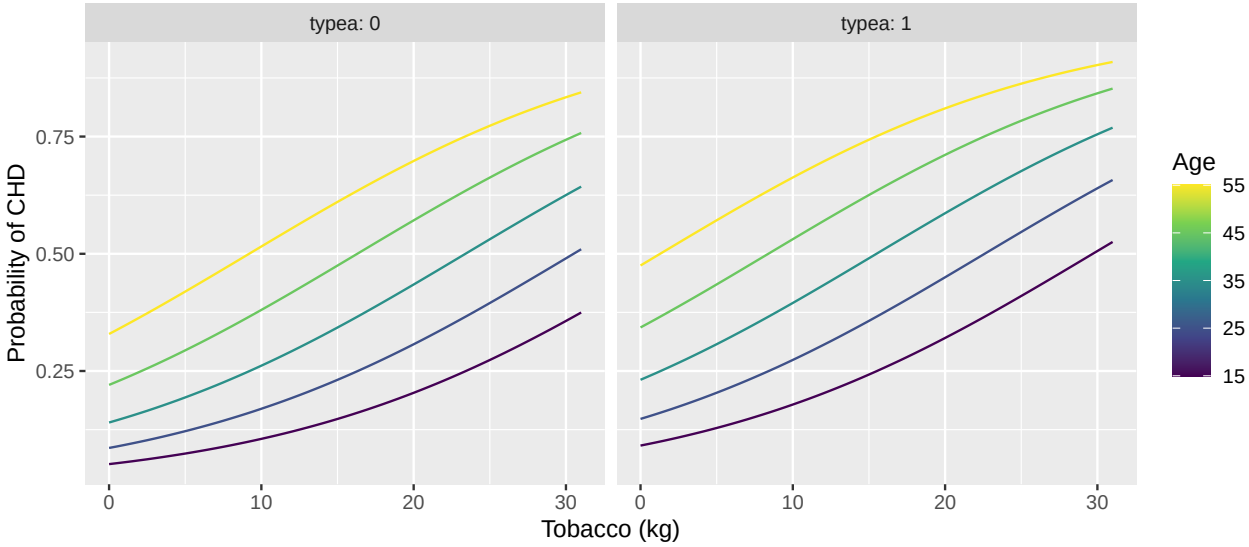
Fixed SBP at median



Again, we will fix SBP at its median, but now draw age curves for a few levels of cumulative tobacco consumption.

Plot 10: Probability of CHD vs. Tobacco accounting Age and Type A Behavior

Fixed SBP at median



Looking at both of these plots, it appears that after considering age and tobacco while fixing SBP at its median, type A behavior does appear to play a role in increasing the probability of having CHD. More specifically, it appears that after accounting for age and tobacco, if a person has type A behavior, the probability of CHD increases at a faster pace which is indicated by the steeper slopes in the plot 10. Also, when the amount of tobacco consumption and age are held constant (at the y-intercept), the probability of having CHD is higher.

Conclusions

In this study, the relationship between type A behavior and CHD after accounting cumulative tobacco, SBP, and age was explored. After experimenting with various models including gam, loess, and logistic models and

investigating the need for interactions, we found that the best model was an additive logistic regression model with no interaction, which consists of the predictor variables: age, cumulative tobacco consumption, type A behavior, and SBP. This fit provided a possible model that could be used to help determine which factors are most important when predicting CHD. In conclusion, after accounting for age, SBP, and cumulative tobacco consumption, type A behavior plays an important role in increasing the likelihood of having CHD. In particular, people who are older, consumed more tobacco, having higher SBP, and perform type A behavior are more likely to have CHD.

Limitations

While this research could be the first step towards helping physicians understand the affects of the risk factors of CHD, there are some limitations to this research that should be noted. To start, there are many other important factors such as family history, weight, and low density lipoprotein cholesterol that could have a large effect on the likelihood of having CHD. These factors were not considered or captured in our final model because we only focused on the most important factors that previous research has found. However, there was a decently large coefficient estimate for type A behaviors found in our model that may be reduced when we consider the other factors that we have not included.

Additionally, it is important to recognize that the self-report Bortner Short Rating Scale used to measure type A behavior has been more recently investigated and it has been found to have an unacceptably low reliability score for measuring type A behavior personalities. Most psychologists and other researchers use updated personality tests or conduct face-to-face interviews to determine a person's likelihood to exhibit type A behaviors. This may be affecting our model and may affect the role that type A behaviors play in likelihood of CHD.

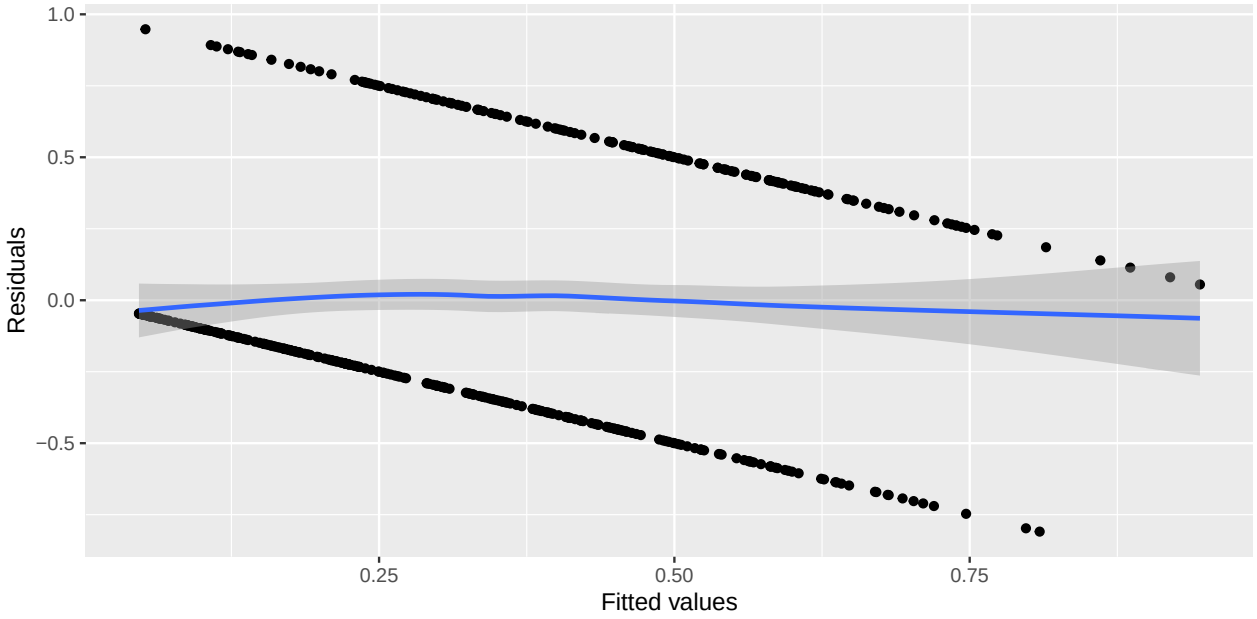
Future Work

For further studies, we could consider other variables in the data set to determine if there are better predictors for CHD that were not considered for the current study of ours. Also, our data consisted of a small population in South Africa of only male participants that was collected in the 1980s. Access to more inclusive data of people from a variety of locations across the world, ages, genders including women and non-binary individuals, and racial and ethnic identities is crucial for the continuation and exploration into research of CHD. In the future studies, using a more diverse and recent data set may help us and other researchers to create models that can better quantify the effects of various risk factors for CHD. In general, access to more data would allow us to generalize this model to a variety of populations which could help physicians around the world in prediction and prognosis of coronary heart disease.

Appendix

Now we will extract the fitted values and residuals from our final model and plot the latter against the former.

```
## `geom_smooth()` using formula 'y ~ x'
```

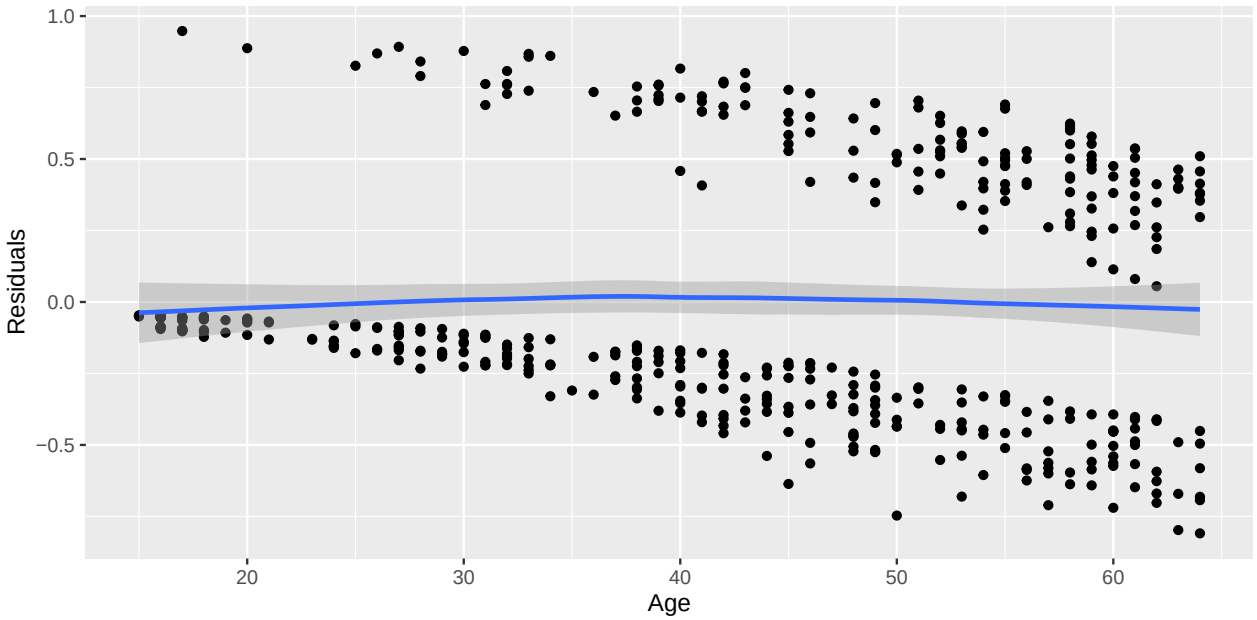


The points on the plot will always fall exactly on two lines, so look at the smooth instead. We see there's a little bit of waviness. This is fairly typical for logistic regression: there's no reason why the relationship between probability and the predictors should have that exact functional form.

Overall, this plot looks pretty good and hovers around the 0 line.

Now we will plot the residuals against age.

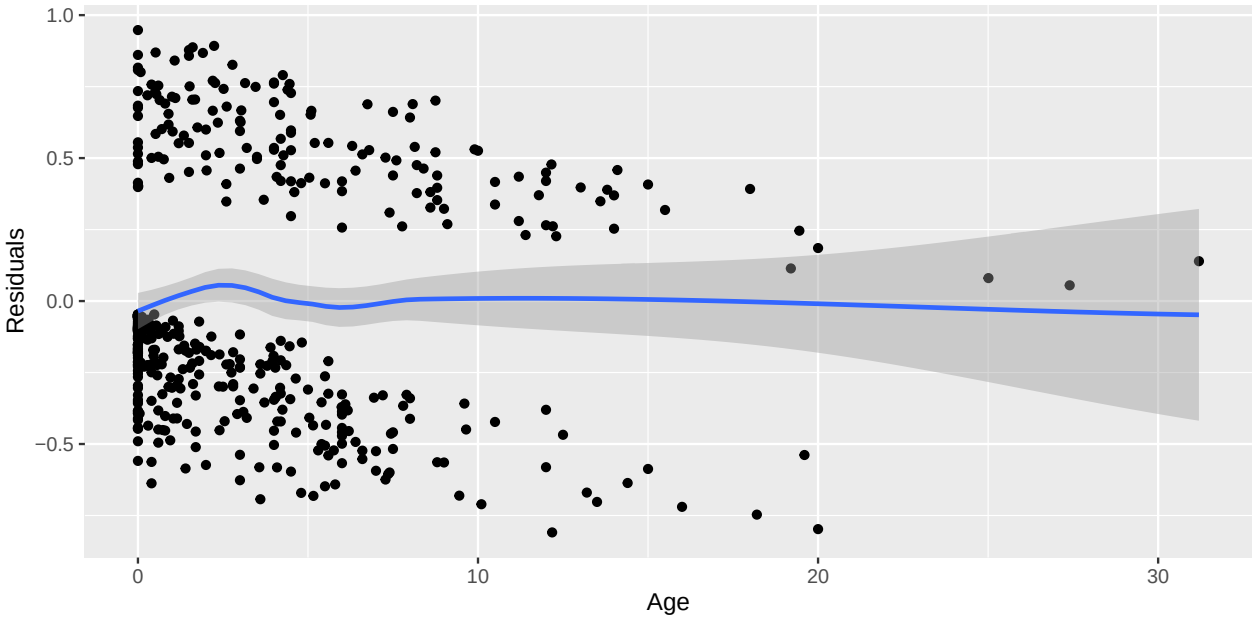
```
## `geom_smooth()` using formula 'y ~ x'
```



The plot appears to contain the 0 line in the confidence band and appears to be relatively null.

Now we will do the same with the tobacco predictor.

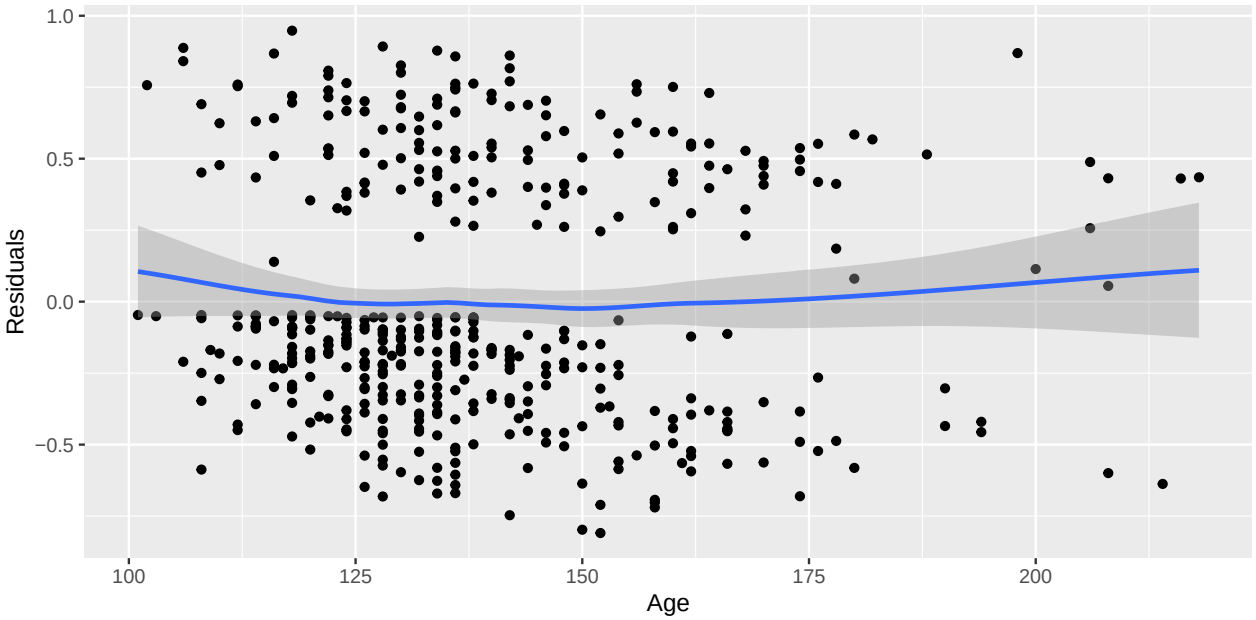
```
## `geom_smooth()` using formula 'y ~ x'
```



There does appear to be a bit of a wiggle at teh beginning of the plot, however, the zero line does appear to be enclosed in the confidence band throughout the plot. There does not appear to be that big of a lack of fit not enough to be too worrisome.

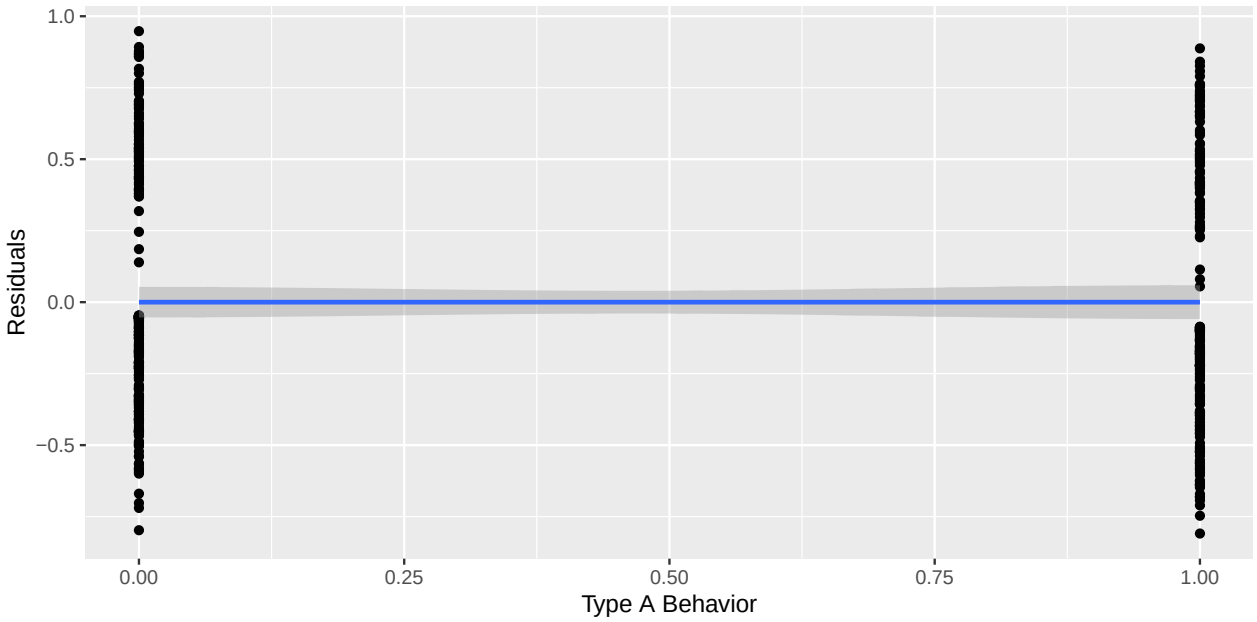
Let's continue to do this for sbp and typea even though these predictors do not have as large effect on the probability of having coronary heart disease.

```
## `geom_smooth()` using formula 'y ~ x'
```



The fit appears to be slightly too high, however, the zero line is still within the bounds of the confidence band, and the line does not wiggle too much.

```
## `geom_smooth()` using formula 'y ~ x'
```



The plot appears to contain the 0 line in the confidence band and appears to be relatively null.

Overall, our residuals of the additive logistic model appear to be null plots and appear to do well. This indicates that our model is doing relatively well.