Statistics 149, Spring 2020
Final Project: Modeling Coronary Heart Disease Risk
May 6, 2020
Authors: Seth Billiau, Katherine Deng, Karissa Huang, Sophia Li
Github Repo: `https://github.com/sethbilliau/heartdisease`

1 Introduction

In this project, we use data from the Framingham Heart Study to model the risk of future coronary heart disease (CHD) using a variety of sociodemographic and health risk measures. Our primary goal is statistical inference—although multiple models using machine learning and other methods have already been developed to accurately predict CHD risk, a focus on inferring the relationships between certain factors and CHD risk will provide the insights necessary for *prevention*.[1] Especially as heart disease claims more lives than any other cause of death in the United States, and as more and more research emerges revealing strong ties between sociodemographic factors and health outcomes, understanding which factors influence one's risk of future CHD—and how—lays the groundwork for more comprehensive and targeted prevention, and the distribution of limited healthcare resources to those at highest risk.[2]

The Framingham Heart Study is an ongoing research study involving cohorts of residents from the town of Framingham, Massachusetts. The data supplied for this project includes 4,238 individuals. The response variable, `CHD_Risk`, is a binary variable indicating whether each person has a ten-year risk of coronary heart disease. There are fifteen possible predictor variables, both quantitative and categorical, consisting of sociodemographic factors such as age and educational attainment, as well as health information such as systolic blood pressure and smoking status.

Our goal is to develop a binary response model using the given predictors that classifies individuals as having or not having a ten-year risk of CHD. Since our primary focus is inference, we evaluate and select models based on the explanatory power they have on the response and their appropriateness for the data.

2 Exploratory Data Analysis

We began our project with an exploration of the dataset. The goals of this exercise were to discover missingness, visualize the distribution of the response and predictor variables, and assess the predictor variables for obvious signs of multicollinearity.

We were fortunate to be given a dataset without missingness in the response variable, but we did observe missingness in six of our predictor variables. Out of a total of 4,238 observations, we found 1 to 388 missing values for the following predictors: `education`, `cigsPerDay`, `totChol`, `BMI`, `heartRate`, `glucose`, and `OnBPMeds`. In our modeling section, we will address this missingness by dropping missing values and by using the `na.convert.mean` method discussed in class.[3]

We continued our exploration by plotting the distribution of `CHD_Risk`, the binary response variable

[1] `http://www.onlinejacc.org/content/71/11_Supplement/A1483`
[2] `https://www.ncbi.nlm.nih.gov/pubmed/28161284`
[3] See the code appendix for exact numbers of missingness.

that represents whether or not a given patient has 10-year risk of coronary heart disease. We found the proportion of at-risk patients in our dataset to be 15%.[4]

Then, we plotted the distributions of each of our 8 quantitative predictor variables conditional on the response class. Obvious visual differences between the two conditional densities give us evidence that a given feature is probably a useful predictor of `CHD_Risk`.[5] Based on this graphic, we noted that the two conditional distributions of `age` and `sysBP` had obvious visual differences, suggesting that these might be useful predictors when modeling.

In addition, most of our predictors appeared to be approximately normally distributed. Though there is some concern of right-skewness in a few predictors (`sysBP`, `glucose`), we did not feel the need to make any transformations.

The notable exception was the `cigsPerDay` variable which contained a large volume of zeroes as non-smokers smoke 0 cigarettes per day. Given that we also had a categorical variable for smoker status, we decided to consider the interaction of `smoker:cigsPerDay` instead of `smoker` and `cigsPerDay` individually. This preserves the information given in both variables while allowing for interpretable results.[6]

We continued this analysis with our 7 categorical response variables, creating contingency tables conditional on the response class.[7] If the probabilities across the rows of our contingency tables were dissimilar, this was evidence of dissimilar conditional distributions and a potentially useful predictor. Based on this analysis, we saw that `Sex` and `Hyp` both demonstrated differences in the conditional distributions, meaning that they may be useful predictors of the response.

Finally, we did a preliminary check for multicollinearity in our predictor variables using GVIF. Finding no evidence of severe multicollinearity, we began modeling.

3 Models

We modeled the probability of a person testing positive for risk of coronary heart disease as a binary response GLM with a logit-link function. We chose this link function because it is the canonical link. The general form of our model is

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_J x_J.$$

We used an analysis of deviance to choose which predictors should be included in the vector of β_j's. In particular, we built our model predictor by predictor and used the likelihood ratio test to test nested models, beginning with the null model. However, since we were given 15 predictors, we felt that it was inefficient to check each predictor at each step, so after looking at a model where each predictor is considered independently (except for `smoker:cigsPerDay`, which we consider to be one predictor for the remainder of the analysis) we composed 'shortlists' of predictors to check at the next age.

[4] See Appendix A, Figure 1

[5] See Appendix A, Figure 2

[6] Note that this eliminates the possibility of, for example, making predictions for an individual that is a non-smoker in the smoker variable but smokes 7 cigarettes per day in the cigsPerDay variable.

[7] See Appendix B, Table 1

Because there was missing data, we built our model in two ways before deciding on a final model to use: 1) dropping all rows with any missing predictor values, 2) using *na.convert.mean*() on the dataframe, which imputes the mean of the predictor across all data points into any missing values, and adds a categorical predictor called *predictor.na* which has value 1 if the data point was originally missing and 0 otherwise.

3.1 GLM Dropping Missing Values

First, we considered dropping all rows with missing data. Originally, the dataset had 4,238 data points and after dropping rows with missing predictors we found 3,656 rows, so we lost $\frac{582}{4238} \approx 13\%$ of our data.

As mentioned above, before conducting an analysis of deviance to build the model term-by-term, we considered a model using all the predictors independently (combining `smoker` and `cigsPerDay` into `smoker:cigsPerDay`) to help inform a 'shortlist' of which predictors are more likely to be significant and therefore need to be tested, and to give us an idea of which predictors may be redundant or unnecessary. This was repeated at regular intervals during the analysis of deviance to make sure we were not missing any important predictors, but in general we only tested the addition of a few promising predictors at a time. As a preliminary step, the Wald tests on the predictors from this model seemed to imply that `age, cholesterol`, `sysBP, glucose, sex`, and `smoker:cigsPerDay` are significant predictors (p = 0.05). We found that adding each of these predictors iteratively allows us to reject the previous model using a likelihood ratio test (χ^2). A table with the deviances of our model checks may be found in the appendix.[8] We then found that `totChol` caused a significantly largest additional drop in deviance. This left us with a 6-predictor model `1 + age + sysBP + sex + smoker:cigsPerDay + glucose + totChol`. It turned out there were no other independent predictors that were significant - at the next step, `education` caused the biggest drop in deviance, but adding it to the 6-predictor model resulted in an insignificant LRT statistic (p = 0.31).

After finding our 6 individual predictors, we turned our attention to interaction terms. First, we considered pairwise interaction terms. The deviances for these models may be found in the appendix.[9] The model with the greatest drop in deviance from our basic 6-predictor model adds the interaction term `glucose:totChol`, but conducting a likelihood ratio test on the nested model reveals an insignificant test statistic. Therefore, we did not need to consider any pairwise interaction terms or beyond. Thus, our final model is comprised of the independent predictors `1 + age + sysBP + sex + smoker:cigsPerDay + glucose + totChol` using the logit link function. The model coefficients are as follows:

Intercept	age	sysBP	sexmale	glucose	totChol	smoker:cigsPerDay
-9.129843	0.05896	0.0175	0.5614	0.00728	0.002272	0.009613

3.2 Analysis of Deviance, GLM na.convert.mean

An alternative method to dealing with the missing data values is to use the `na.convert.mean` function - this performs mean imputation on predictors with missing values and adds a column called `predictor.na`, which is 1 if the predictor is missing and 0 if not. We followed similar steps

[8] See Appendix A: Table 2
[9] See Appendix A, Table 3

as in the previous section, and the model was similar until we added the last few predictors. A table with the deviance of each model considered may be found in the appendix.[10]

Our independent predictors end up being `1 + age + sysBP + sex + smoker:cigsPerDay + glucose + BMI.na + PrevStroke`. When considering interaction terms, it turns out that only `sex:sysBP` caused a significant drop in deviance (to 3201.204, p = 0.044). We did not consider triple interaction terms because at that point the model would be difficult to interpret (it seems unlikely that they would be significant anyway). Thus, an analysis of deviance when using `na.convert.mean` on the data yielded a final model with `1 + age + sysBP + sex + smoker:cigsPerDay + glucose + BMI.na + PrevStroke + sex:sysBP`. A brief summary of the coefficents for each term is as follows.

Intercept	age	sysBP	sexmale	glucose	BMI.na	PrevStroke
-8.1033	0.0655	0.0140	-0.5979	0.0078	2.0874	0.9329
smoker:cigsPerDay	sysBP:sexmale					
0.0108	0.0078					

These 2 approaches to deal with missing data resulted in different, but related, models. The GLM models both share `age`, `sysBP`, `sexmale`, `glucose`, `smoker:cigsPerDay` as the first five predictors in the analysis of deviance, and the coefficient estimates for these predictors are similar. The differences comes after these predictors as first model adds `totChol` while the second model adds `BMI.na`, `PrevStrokeYes` and the interaction terms `sysBP:sexmale`. As these differences arise late in our analysis of deviance testing and the coefficients for the additional predictors are quite small (i.e. they don't contribute much to the log-odds of testing positive for coronary heart disease), we can attribute most of this variation to those missing points. Because we only lose around 13% of our data when dropping our missing rows and still have enough predictions to make a meaningful model, we decided to proceed with the model that resulted from dropping all missing data.

3.3 GLM Model Diagnostics

Having arrived at a working model with formula `1 + age + sysBP + sex + smoker:cigsPerDay + glucose + totChol` using an analysis of deviance, we then performed diagnostics to evaluate the fit and appropriateness of this model. Based on a summary of the model fit, the ratio of residual deviance to residual degrees of freedom is $3120.5/3655 = 0.853762$, which is very close to 1, our benchmark for assessing overdispersion. Therefore, we do not see greater variability in our dataset than expected based on our binomial model, and our final model appears to be a good fit. Another numerical method for examining lack of fit in binary data is the Hosmer-Lemeshow test. The test gives large p-values for multiple values of the g parameter, so we cannot reject the null hypothesis, and therefore there is no indication of an obvious lack of fit. The Hosmer-Lemeshow test, however, is a low-power test (and it is possible that we have a poor fit even if the null hypothesis is not rejected), so we still proceed with graphical diagnostics to more closely examine our model's performance on specific observations.

First, we created a plot of binned averaged residuals against binned fitted values, included as Figure 2 in the appendix. There were no clear non-linearities or outliers in the residuals. We also created a plot of Cook's distances, included as Figure 3 in the appendix. All of our observations have Cook's distances *far* below our usual benchmark of 1 (in fact, the highest Cook's distance is just slightly

[10] See Appendix A, Table 4

over 0.025), which strongly suggests that there are no influential observations in the data. Based on numerical and graphical diagnostics, our final model seems to be a good fit.

3.4 General Additive Models

We also considered using General Additive Models to fit the data and explored whether or not some of the predictors have non-linear relationships with the response variable that can be captured using splines. We used the dataset with the dropped NAs to do this for consistency, as this is the dataset that we chose our final GLM from. To start, we plotted the response variable against each predictor.[11]

From the outset, we noted that there were few predictor variables that might have non-linear relationships with the response, including `glucose`, `cigsPerDay`, `totChol`, `sysBP`, and `diaBP`. We used ANOVA to further investigate these relationships, removing predictors one by one starting from the full model. We included the spline of `cigsPerDay` and `smoker` as a joint effect, because (as stated before) if someone does not smoke, then they would have no cigarettes per day, which demonstrates a clear relationship between the two predictor variables. In order of removal, we removed `Diab`, `OnBPMeds`, `Hyp`, `education`, `BMI`, `Heartrate`, and finally `PrevStroke`. After running all the ANOVAs, we determined that the best model is the model `s(age)`, `s(cigsPerDay, Smoker)`, `s(totChol)`, `s(sysBP)`, `s(diaBP)`, `s(glucose)`, and `sex`. We then ran an ANOVA test comparing this model with our final GLM model to see whether the new GAM model reduced the deviance significantly. Below is the output from that ANOVA.

```
    Analysis of Deviance Table

Model 1: CHD_Risk ~ 1 + age + sysBP + sex + smoker:cigsPerDay + glucose +
    totChol
Model 2: CHD_Risk ~ s(age) + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) +
    s(diaBP) + s(glucose) + sex
  Resid. Df Resid. Dev    Df Deviance Pr(>Chi)
1    3649.0     2762.5
2    3640.7     2741.3 8.274   21.172 0.007899 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We saw that the new model had a chi-squared p-value of 0.007899, which is significant at the 0.05 level, suggesting that the splines were in fact significantly helping to reduce the deviance. Next, we turned to model diagnostics to see whether there were any diagnostic features that were out of the ordinary. The plots of the residuals against each smoothed function are shown in the Appendix. The plots show that all of the residuals looked relatively reasonable except for the `cigsPerDay` variable, but the effect of `cigsPerDay` is explained in a joint smooth with smoker, which might explain why the residuals of `cigsPerDay` plotted alone look strange.

[11] See Appendix A, Figure 5

4 Conclusion

Our GLM model is specified by the formula

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = -9.13 + 0.0659 \cdot \texttt{age} + 0.0175 \cdot \texttt{sysBP} + \\ 0.561 \cdot \texttt{sexmale} + 0.0073 \cdot \texttt{glucose} + 0.0023 \cdot \texttt{totChol} + \\ 0.009613 \cdot \texttt{smoker:cigsPerDay}.\end{aligned}$$

We can interpret each coefficient as follows: Every additional year of age corresponds to a 0.0659 increase in the log-odds of being at risk for coronary heart disease. Every 1-mmHg increase in systolic blood pressure corresponds to a 0.0175 increase in the log-odds of being at risk for CHD. Being male compared to female corresponds to a 0.561 increase in the log-odds of being at risk for CHD. A 1 mg/dL increase in glucose levels corresponds to a 0.0073 increase in log-odds of risk for CHD. A 1 mg/dL increase in total cholesterol level corresponds to a 0.0023 increase in the log-odds of risk for CHD. For smokers, each additional cigarette smoked per day results in a 0.0096 increase in the log-odds of risk for CHD. As an example, our estimate for a male who is 50 years old, smokes 9 cigarettes per day, has a total cholesterol level of 200 mg/dL, has a systolic blood pressure of 132 mmHg, a BMI index of 25 kg/m^2, a heartrate of 75 bpm, a glucose level of 80 mg/dL gives

$$\begin{aligned}\text{logit}(p) &= -9.129843 + 0.065896 \cdot 50 + 0.017535 \cdot 132 + \\ &\quad 0.541446 \cdot 1 + 0.00728 \cdot 80 + 0.002272 \cdot 200 + 0.009613 \\ &\implies p = 0.1265\end{aligned}$$

so the person has a 12.65% chance of having 10-year risk for coronary heart disease.

Someone who is older, is male, has higher blood pressure, has high glucose and cholesterol levels, and smokes a lot is the most likely to have 10-year risk of CHD. People in this demographic (older and male) should keep this in mind and do their best to keep their cholesterol levels low. Smoking as little as possible would also help. If one has a medical or genetic history of high blood pressure, they may also be at higher risk. People who are aware that they are at higher risk may want to take measures to lead healthier lifestyles.

5 Evaluation

We chose to use the GLM model (from dropping missing data) as our final model because the premise of this project was to report on interpretable relationships between the predictor variables and the response variable. There are other ways to deal with missing data such as mean imputation, which we did not consider. We also did find that `BMI.na` was significant when using `na.convert.mean`, so it is possible that we lose some predictive power when dropping all missing values. Between the GLM and GAM, we discovered that our GAM final model and our GLM final model contain essentially the same predictors, with the GAM model containing extra variables `sysBP` and `diaBP`, but the GAM model is significantly less interpretable due to the polynomial terms on most of the predictors. However, we did find that the GAM did significantly reduce the deviance so if we wished to perform prediction, the GAM model may be a better choice.

The GAM model study highlights one possible limitation of our approach: we did not examine how our model performs as a predictive tool. Although our model appears to be a good fit for the data based on numerical and graphical diagnostics, we never evaluated its classification accuracy. It's possible that our model is too complex and is overfitted on the provided dataset, in which case it would perform poorly at classification for a new set of individuals. The preferred approach to address possible overfitting is to split the data into (typically 80-20 ratio) training and test sets, build our model using analysis of deviance on the training set, then evaluate its performance using classification accuracy on the test set. If we find that accuracy drops off substantially for the test set, then it is likely that our model is overfitting and should be made less complex to better balance the bias-variance tradeoff. The problem of overfitting is less severe in this case since our focus is on inference of relationships rather than on pure predictive performance, but even for inferential tasks, it is still valuable to consider whether our final model generalizes well to other datasets.

6 Appendix A: Figures

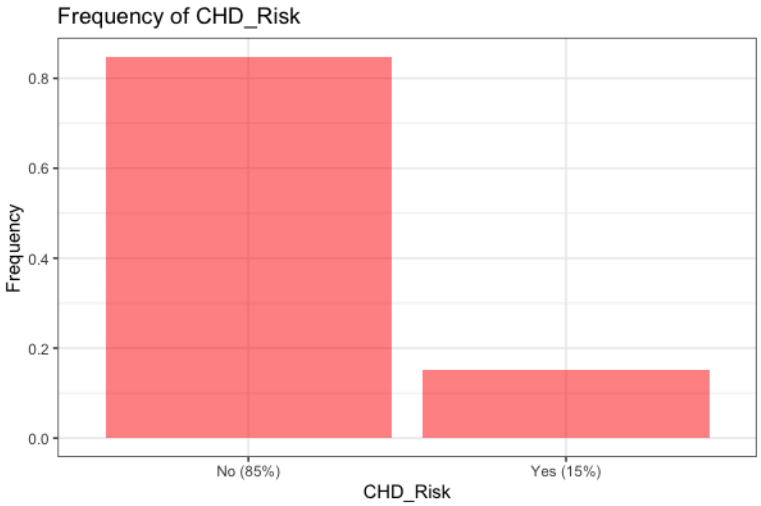

Figure 1: Frequency Histogram of the Response Variable

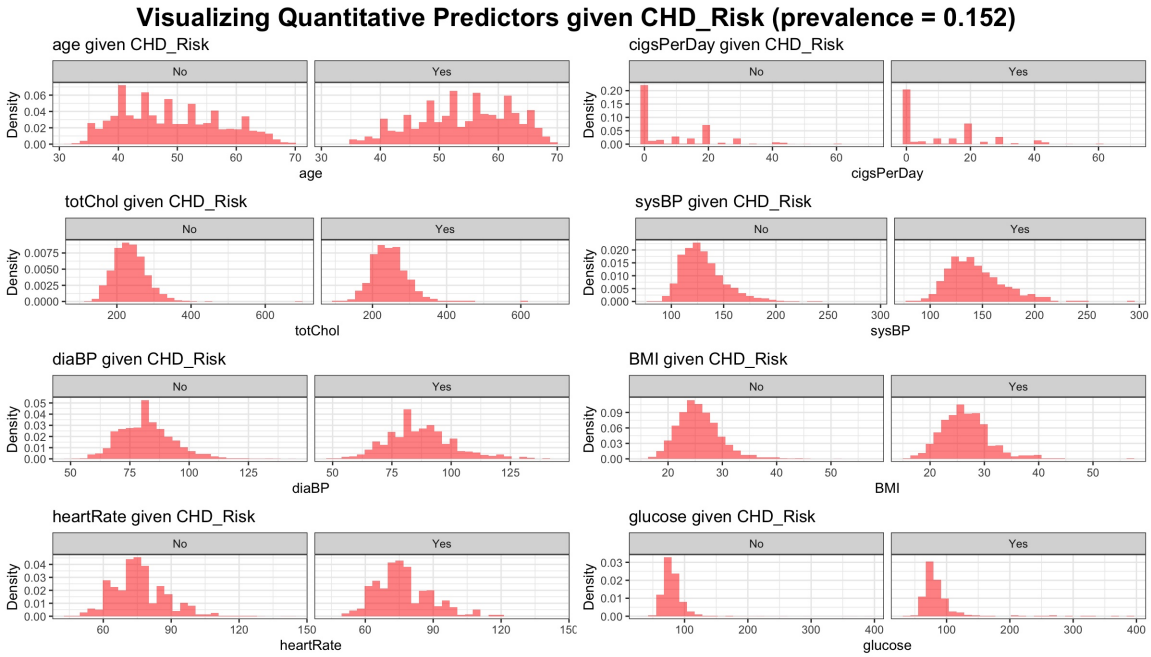

Figure 2: Quantitative Predictors Conditional on the Response

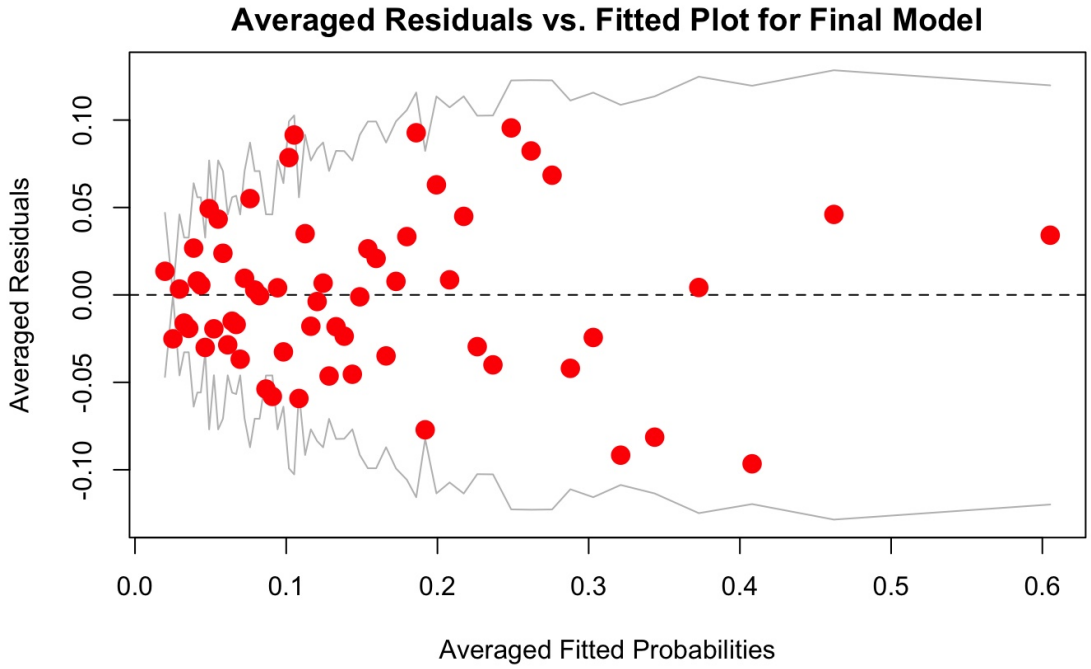

Figure 3: Residuals vs. Fitted Plot for Final GLM

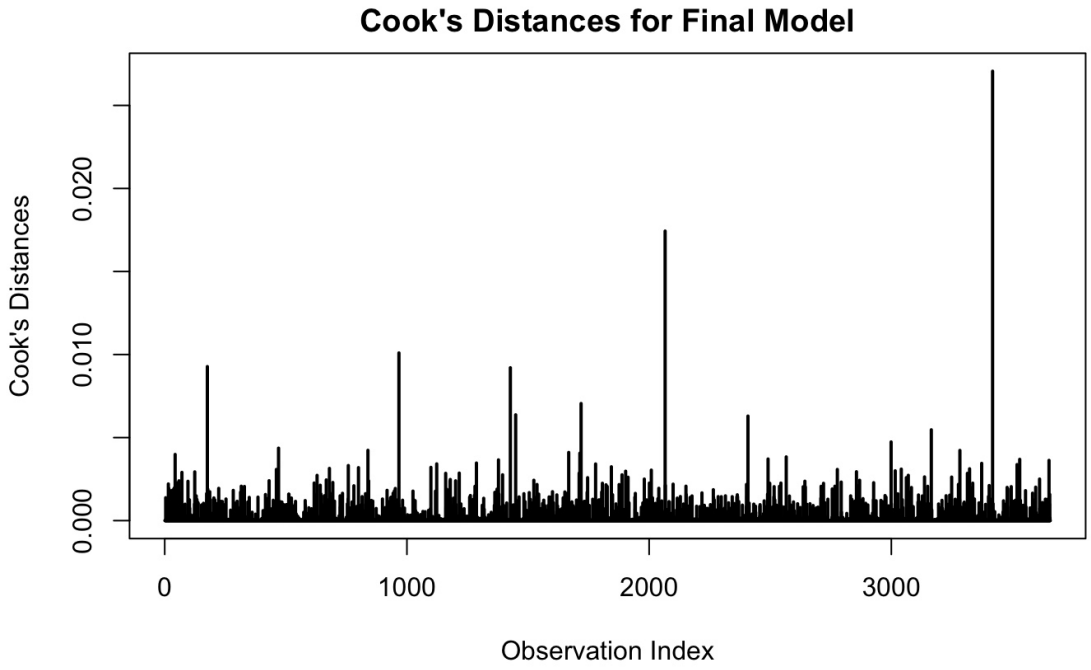

Figure 4: Cook's Distances for Final GLM

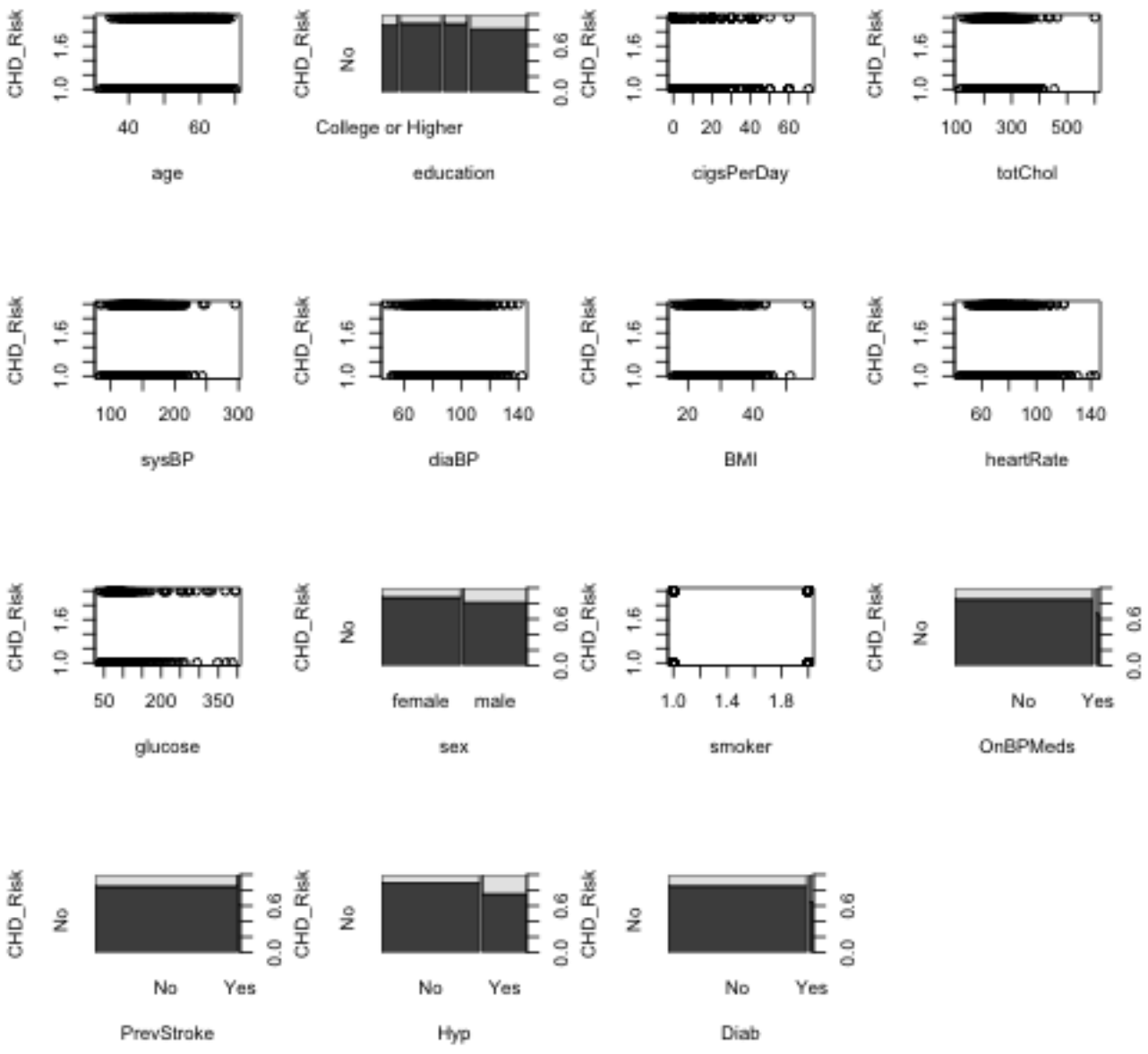

Figure 5: Response vs. Each Predictor

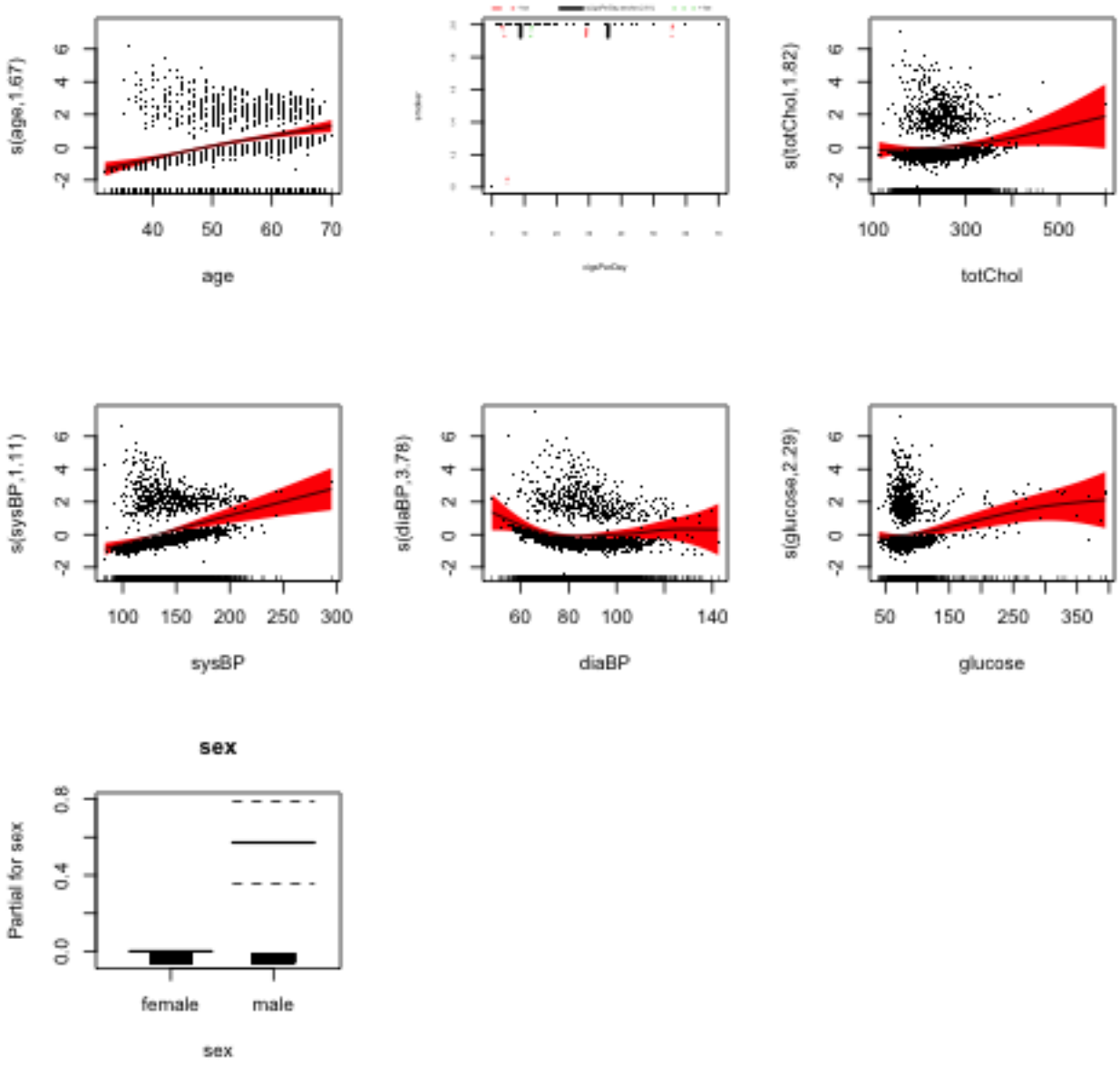

Figure 6: GAM Residual Plots

7 Appendix B: Tables

Table 1: Probabilities of Qualitative Variables Conditional on the Outcome

	CHD Risk Outcome	No	Yes
Education	College or Higher	0.12	0.11
	High School or GED	0.32	0.23
	Some College	0.17	0.14
	Some High School	0.40	0.51
Sex	female	0.59	0.47
	male	0.41	0.53
Smoker	Nonsmoker	0.51	0.48
	Yes	0.49	0.52
OnBPMeds	No	0.98	0.94
	Yes	0.02	0.07
PrevStroke	No	1.00	0.98
	Yes	0.00	0.02
Hyp	No	0.72	0.49
	Yes	0.28	0.51
Diab	No	0.98	0.94
	Yes	0.02	0.06

Table 2: Analysis of Deviance, GLM with missing values dropped

Model	Deviance	Coefficients
1	3120.5	1
1 + age	2920.3	2
1 + sysBP	2957.3	2
1 + sexmale	3090	2
1 + smoker:cigsPerDay	3110.9	2
1 + age + sysBP	2855.2	3
1 + age + sex	2883.7	3
1 + age + smoker:cigsPerDay	2881.5	3
1 + age + sysBP + sex	2805.6	4
1 + age + sysBP + smoker:cigsPerDay	2811.9	4
1 + age + sysBP + sex + smoker:cigsPerDay	2785.6	5
" + education	2782.5	6
" + totChol	2781.4	6
" + diaBP	2785.3	6
" + BMI	2784.76	6
" + heartRate	2785.6	6
" + glucose	2766.5	6
" + OnBPMeds	2784.3	6
" + PrevStroke	2783.13	6
" + Hyp	2782.85	6
" + Diab	2775.52	6
1 + age + sysBP + sex + smoker:cigsPerDay + glucose	2766.5	6
" + education	2763.4	7
" + totChol	2762.5	7
" + BMI	2766.02	7
" + OnBPMeds	2765.3	7
" + PrevStroke	2764.1	7
" + Hyp	2763.4	7
" + Diab	2766.5465	7
1 + age + sysBP + sex + smoker:cigsPerDay + glucose + totChol	2762.5	7
" + education	2758.95	8
" + OnBPMeds	2761.409	8
" + PrevStroke	2759.98.02	8
" + Hyp	2759.479	8

Table 3: Analysis of Deviance, GLM with missing values dropped, pairwise interaction terms

Model Coefficients	Deviance
1 + age + sysBP + sex + smoker:cigsPerDay + glucose + totChol	2762.5
" + age:sysBP	2761.898
" + age:sex	2762.233
" + age:ssmoker:cigsPerDay	2761.905
" + age:glucose	2762.468
" + age:totChol	2760.629
" + sysBP:sex	2760.044
" + sysBP:smoker:cigsPerDay	2762.39
" + sysBP:glucose	2762.476
" + sysBP:totChol	2762.268
" + sex:smoker:cigsPerDay	2761.693
" + sex:glucose	2762.258
" + sex:totChol	2759.904
" + smoker:cigsPerDay:glucose	2761.937
" + smoker:cigsPerDay:totChol	2762.429
" + glucose:totChol	2759.187

Model	Deviance	Coefficients
1	3611.55	1
1 + age	3396.326	2
1 + sysBP	3432.642	2
1 + sex	3578.741	2
1 + BMI.na	3597.093	2
1 + smoker:cigsPerDay	3597.093	2
1 + age + sysBP	3325.673	3
1 + age + sex	3356.181	3
1 + age + smoker:cigsPerDay	3347.071	3
1 + age + BMI.na	3383.409	3
1 + age + sysBP + sex	3273.349	4
1 + age + sysBP + smoker:cigsPerDay	3272.317	4
1 + age + sysBP + BMI.na	3313.134	4
1 + age + sysBP + smoker:cigsPerDay + sex	3245.906	5
1 + age + sysBP + smoker:cigsPerDay + BMI.na	3258.933	5
1 + age + sysBP + sex + smoker:cigsPerDay	2785.6	5
" + education	3242.524	6
" + totChol	3242.553	6
" + diaBP	3245.684	6
" + BMI	3245.329	6
" + heartRate	3245.893	6
" + glucose	3224.086	6
" + OnBPMeds	3242.964	6
" + PrevStroke	3240.315	6
" + Hyp	3242.288	6
" + Diab	3233.496	6
" + BMI.na	3231.413	6
1 + age + sysBP + sex + smoker:cigsPerDay + glucose	3224.086	6
" + education	3220.609	7
" + totChol	3220.927	7
" + BMI	3223.803	7
" + OnBPMeds	3221.249	7
" + PrevStroke	3218.611	7
" + Hyp	3219.987	7
" + Diab	3223.58	7
1 + age + sysBP + sex + smoker:cigsPerDay + glucose + BMI.na	3209.296	7
" + education	3205.643	8
" + OnBPMeds	3206.521	8
" + PrevStroke	3205.261	8
" + Hyp	3205.317	8
" + Diab	3209.279	8
" + BMI.na	3209.184	8
" + totChol	3205.576	8

Table 4: Analysis of deviance, na.convert.mean

Model Number	Deviance	Coefficients
s(age) + education + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(BMI) + s(heartRate) + s(glucose) + sex + OnBPMeds + PrevStroke + Hyp + Diab	2725.3	17
" "-s(age)	2828.8	16
" "-education	2725.3	14
" "-s(totChol)	2742.5	16
" "-s(sysBP)	2738.9	16
" "-s(diaBP)	2734.1	16
" "-s(BMI)	2732.4	16
" "-s(heartRate)	2732.7	16
" "-s(glucose)	2738.9	16
" "-sex	2748.0	16
" "-s(cigsPerDay, smoker)	2746.2	16
" "-OnBPMeds	2725.4	16
" "-PrevStroke	2727.2	16
" "-Hyp	2726.4	16
" "-Diab	2728.4	16
s(age) + education + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(BMI) + s(heartRate) + s(glucose) + sex + OnBPMeds + PrevStroke + Hyp	2728.4	16
" "-s(age)	2828.9	15
" "-education	2732.0	13
" "-s(totChol)	2742.5	15
" "-s(sysBP)	2741.6	15
" "-s(diaBP)	2737.8	15
" "-s(BMI)	2728.7	15
" "-s(heartRate)	2735.5	15
" "-s(glucose)	2744.2	15
" "-sex	2750.8	15
" "-s(cigsPerDay, smoker)	2748.1	15
" "-OnBPMeds	2728.5	15
" "-PrevStroke	2730.6	15
" "-Hyp	2729.4	15

Table 5: Analysis of deviance (GAM), dropped NA data (cont. in next table)

Model Number	Deviance	Coefficients
s(age) + education + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(BMI) + s(heartRate) + s(glucose) + sex + PrevStroke + Hyp	2728.5	15
" "-s(age)	2829.8	14
" "-education	2732.1	12
" "-s(totChol)	2742.8	14
" "-s(sysBP)	2742.2	14
" "-s(diaBP)	2737.8	14
" "-s(BMI)	2728.7	14
" "-s(heartRate)	2735.7	14
" "-s(glucose)	2739.2	14
" "-sex	2750.9	14
" "-s(cigsPerDay, smoker)	2748.2	14
" "-PrevStroke	2730.8	14
" "-Hyp	2729.4	14
s(age) + education + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(BMI) + s(heartRate) + s(glucose) + sex + PrevStroke	2729.4	14
" "-s(age)	2827.8	13
" "-education	2733.3	11
" "-s(totChol)	2743.6	13
" "-s(sysBP)	2750.5	13
" "-s(diaBP)	2740.3	13
" "-s(BMI)	2729.2	13
" "-s(heartRate)	2736.0	13
" "-s(glucose)	2741.0	13
" "-sex	2751.7	13
" "-s(cigsPerDay, smoker)	2749.2	13
" "-PrevStroke	2731.8	13

Table 6: Analysis of deviance (GAM), dropped NA data (cont. in next table)

Model Number	Deviance	Coefficients
s(age) + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(BMI) + s(heartRate) + s(glucose) + sex + PrevStroke + Hyp	2733.3	12
" "-s(age)	2841.2	11
" "-s(totChol)	2747.3	11
" "-s(sysBP)	2755.3	11
" "-s(diaBP)	2741.7	11
" "-s(BMI)	2733.8	11
" "-s(heartRate)	2740.6	11
" "-s(glucose)	2744.9	11
" "-sex	2757.6	11
" "-s(cigsPerDay, smoker)	2751.6	11
" "-PrevStroke	2733.7	11
s(age) + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(heartRate) + s(glucose) + sex + PrevStroke	2733.8	11
" "-s(age)	2841.2	10
" "-s(totChol)	2742.2	10
" "-s(sysBP)	2754.7	10
" "-s(diaBP)	2742.4	10
" "-s(heartRate)	2739.4	10
" "-s(glucose)	2745.6	10
" "-sex	2756.3	10
" "-s(cigsPerDay, smoker)	2752.4	10
" "-PrevStroke	2734.1	10

Table 7: Analysis of deviance (GAM), dropped NA data (cont. in next table)

Model Number	Deviance	Coefficients
s(age) + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(heartRate) + s(glucose) + sex + PrevStroke	2739.4	10
" "-s(age)	2848.1	9
" "-s(totChol)	2748.3	9
" "-s(sysBP)	2763.9	9
" "-s(BMI)	2751.0	9
" "-s(glucose)	2753.3	9
" "-sex	2765.0	9
" "-s(cigsPerDay, smoker)	2759.8	9
" "-PrevStroke	2741.3	9
s(age) + s(cigsPerDay, smoker) + s(totChol) + s(sysBP) + s(diaBP) + s(heartRate) + s(glucose) + sex	2741.3	9
" "-s(age)	2851.5	8
" "-s(totChol)	2749.1	8
" "-s(sysBP)	2764.6	8
" "-s(diaBP)	2753.2	8
" "-s(glucose)	2755.6	8
" "-sex	2767.2	8
" "-s(cigsPerDay, smoker)	2761.6	8

Table 8: Analysis of deviance (GAM), dropped NA data

8 Appendix C: Code

Compiled code for this project can be found at `https://github.com/sethbilliau/heartdisease`.