

STAT 331 Final Project

Krishna Prem Pasumarthy & Islam Amin

April 16, 2020

1 Summary

The objective of this report is to come up with a suitable model that best predicts the risk score for CHD based on other variates available in the Framingham Heart Study dataset. [WRITE ABOUT STATISTICAL ANALYSIS] Stepwise automated model selection is used to create a regression model for the response variate logit(chdrisk) based on other covariates and their interactions. Then after conducting some F-tests, another model is manually constructed by removing some interactions from the stepwise model. Both the models are diagnosed and the latter is selected to have better predictive as well as explanatory power.

2 Descriptive Statistics

First, take a look at summary statistics of the Framingham Heart Study dataset.

Table 1: Summary Statistics

chdrisk	sex	totchol	age	sysbp	diabp	cursmoke	cigpdny	bmi
Min. :0.0050	Female:1305	Min. :112.0	Min. :44.00	Min. : 86.0	Min. : 30.00	No :1504	Min. : 0.00	Min. :14.43
1st Qu.:0.1320	Male :1001	1st Qu.:207.0	1st Qu.:53.00	1st Qu.:122.5	1st Qu.: 73.00	Yes: 802	1st Qu.: 0.00	1st Qu.:23.22
Median :0.2240		Median :235.5	Median :60.00	Median :136.0	Median : 80.00		Median : 0.00	Median :25.40
Mean :0.2655		Mean :237.8	Mean :60.23	Mean :139.2	Mean : 81.07		Mean : 6.84	Mean :25.78
3rd Qu.:0.3448		3rd Qu.:265.0	3rd Qu.:67.00	3rd Qu.:153.0	3rd Qu.: 88.00		3rd Qu.:10.00	3rd Qu.:27.91
Max. :0.9770		Max. :625.0	Max. :81.00	Max. :246.0	Max. :130.00		Max. :80.00	Max. :46.52

diabetes	bpmeds	heartrte	glucose	prevmi	prevstrk	prevhyp	hdlc	ldlc
No :2142	No :1973	Min. : 44.00	Min. : 46.00	No :2189	No :2260	No : 957	Min. : 10.00	Min. : 20.0
Yes: 164	Yes: 333	1st Qu.: 70.00	1st Qu.: 75.00	Yes: 117	Yes: 46	Yes:1349	1st Qu.: 38.00	1st Qu.:152.0
		Median : 76.00	Median : 83.00				Median : 47.00	Median :180.0
		Mean : 77.61	Mean : 89.07				Mean : 48.89	Mean :183.1
		3rd Qu.: 85.00	3rd Qu.: 95.00				3rd Qu.: 57.00	3rd Qu.:210.0
		Max. :150.00	Max. :478.00				Max. :189.00	Max. :565.0

First observation we make from the summary is that the median and average ages are around 60, which means the survey seems to have been done on a relatively old group of people. We also have a significantly higher number of females in the study, almost 30% more than the number of males. This might affect the nature of the data to be skewed towards behaviors and physical attributes associated with females.

A further inspection of the expected coronary heart disease (CHD) risk against certain categorical variates, gives more insights.

For instance, if we take a look at expected CHD risk against whether or not an individual has hypertension, we get the following result:

```
## fhsd$prevhyp: No
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 0.005 0.077 0.140 0.176 0.216 0.944
## -----
## fhsd$prevhyp: Yes
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 0.0320 0.1980 0.2890 0.3291 0.4010 0.9770
```

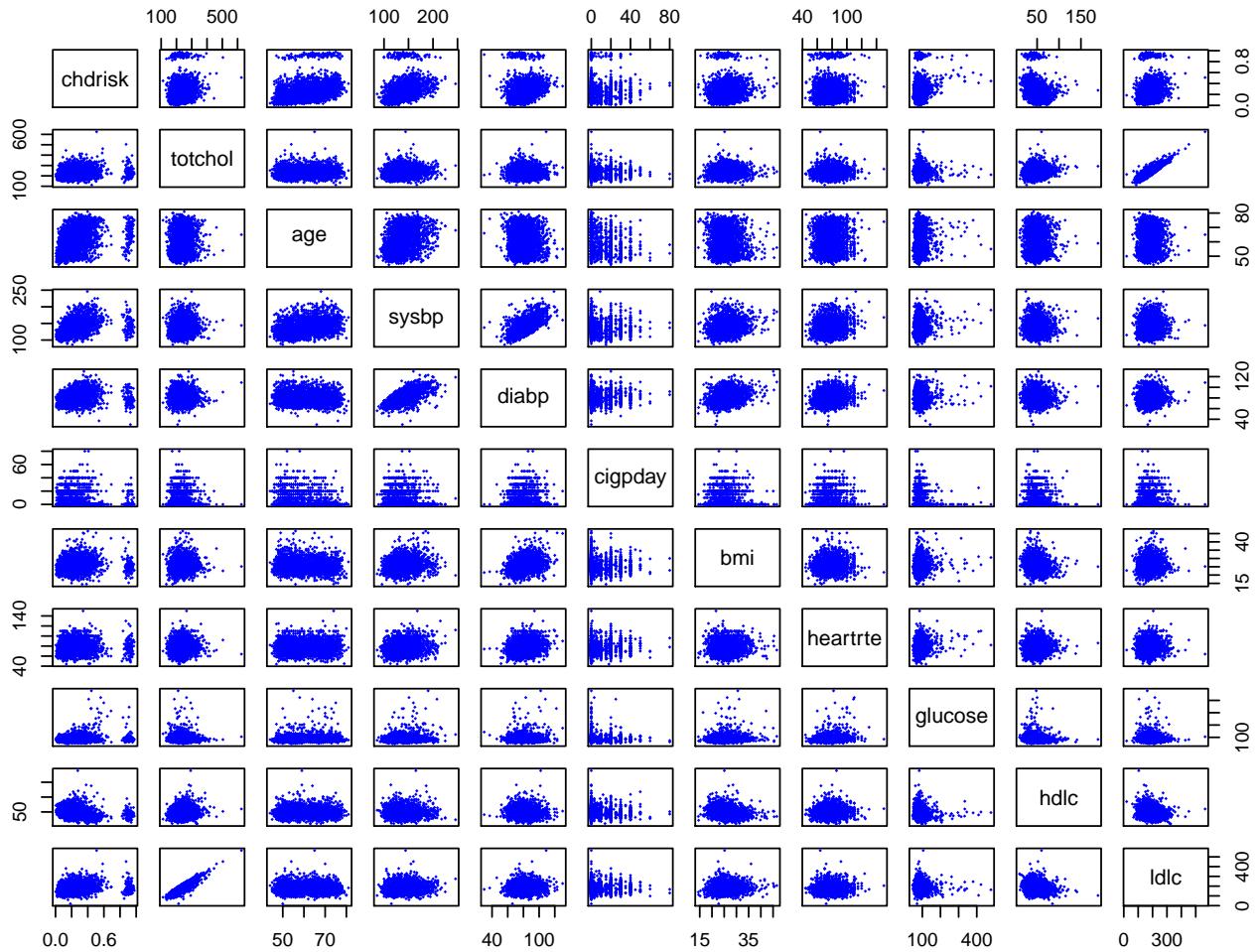
Indeed, we have that mean CHD risk given that a person has hypertension is significantly higher than the mean for people who did not have hypertension.

```
## fhsd$prevstrk: No
##   Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.0050 0.1300 0.2200 0.2611 0.3392 0.9770
##
## fhsd$prevstrk: Yes
##   Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.2020 0.3412 0.4410 0.4820 0.5060 0.9660
```

Again, we see the same results with people who had a stroke before the study, with even a higher difference between the two groups.

Now take a look at pair plots of all numeric explanatory/response variates i.e. variates excluding logical variates such as `cursmoke`.

Pair Plots of Continuous Variates



From the pair plots, we can observe a strong correlation between low density lipoprotein cholesterol and serum total cholesterol. As total cholesterol increases, low density lipoprotein cholesterol seems to increase as well. Another positive correlation can be observed between systolic and diastolic blood pressures. In fact, from this we can infer that blood pressure probably increases and decreases generally for both systolic and diastolic states.

Now take a look at the VIFs of these variates.

```

##      sexMale      totchol        age      sysbp      diabp cursmokeYes
## 1.225191    10.634882   1.489926    2.918660    2.406411    2.978609
##      cigpday       bmi diabetesYes bpmedsYes heartrte      glucose
## 2.973594    1.181865   1.286401    1.214744    1.105902    1.308923
##      prevmiYes prevstrkYes prevhypYes     hdlc      ldlc
## 1.067134    1.045746   1.823014    2.287571   10.367649

```

We observe higher VIF values higher than 10 for both serum total cholesterol and low density lipoprotein cholesterol, which means they are significantly colinear.

3 Candidate Models

3.1 Automated Model Selection

In this section we start producing a candidate model using automated model selection. Here, we choose to use a stepwise as we have observed from lectures that it usually acts as a compromise between backward and forward selection methods. This way, we avoid having relatively a lot of variates in our final model and also we capture as many necessary variates as possible.

We first try our initial and maximum models as follows,

However, we end up getting NAs in the coefficients for two interactions (Currently smoking with cigarettes per day and anti-hypertensive medication with previous hypertension).

If we investigate the relationship between these variables as shown below, we see that those who do not smoke have no cigarettes per day making the two columns cursmoke and cigpday linearly dependent.

Furthermore, if someone did not hypertension, they would not use anti-hypertensive medication causing a linear dependence between these two columns.

Table 2: cursmoke against cigpday

	0	1	2	3	4	5	6	7	8	9	10	12	14	15	16	17	18	19	20	23	25	26	27	28	30	35	40	45	50	60	80	
No	1504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Yes	0	16	18	34	11	18	24	9	18	5	76	3	3	50	6	1	8	1	279	1	14	1	1	1	119	5	64	1	10	3	2	

Table 3: cursmoke against cigpday

	No	Yes
	No	957
	Yes	333

To fix these we remove these two interactions from Mmax to retain the following model:

```

# remove the coeffs with the problem and add quadratic terms for the continuous variables
Mmax <- lm(I(logit(chdrisk)) ~ .()^2 - cursmoke:cigpday - bpmeds:prevhyp +
  I(totchol ^ 2) + I(sysbp ^ 2) + I(diabp ^ 2)
  + I(bmi ^ 2) + I(glucose ^ 2)
  + I(hdlc ^ 2) + I(ldlc ^ 2), data = fhsd)

```

We also add some quadratic terms for continuous variates in hope of having some additional predictive power.

Finally, we produce the following model using stepwise model selection:

```

## lm(formula = I(logit(chdrisk)) ~ sex + totchol + age + sysbp +
##      diabp + cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
##      glucose + prevmi + prevstrk + prevhyp + hdlc + ldlc + I(hdlc^2) +
##      I(bmi^2) + I(diabp^2) + I(sysbp^2) + sysbp:prevmi + totchol:prevhyp +

```

```

## diabetes:prevmi + prevhyp:ldlc + sysbp:prevhyp + totchol:heartrte +
## sysbp:diabetes + diabp:bmi + diabp:hdlc + prevmi:hdlc + prevmi:prevhyp +
## sex:glucose + age:ldlc + age:heartrte + cigpday:hdlc + bmi:ldlc +
## totchol:hdlc + totchol:prevmi + sysbp:heartrte + sysbp:bpmeds +
## cursmoke:hdlc + prevmi:prevstrk + diabetes:hdlc + sex:sysbp +
## cigpday:glucose + heartrte:glucose + diabp:glucose + cursmoke:ldlc +
## age:cigpday + age:hdlc + hdlc:ldlc + age:prevhyp + diabp:prevhyp +
## diabp:cursmoke + diabp:cigpday + bmi:bpmeds + bpmeds:glucose +
## age:prevmi + sex:ldlc + cigpday:heartrte + cigpday:prevmi +
## glucose:prevmi + heartrte:prevmi + bpmeds:prevstrk, data = fhsd)

```

3.2 Manual Model Selection

The following table lists terms in the stepwise model that result in insignificance when an F-test is performed by removing them solely from the model along with corresponding p-values in a sorted order.

```
## Warning: package 'stringr' was built under R version 3.5.2
```

Table 4: Variates/Interactions with insignificant p-values from F-test

cigpday:heartrte	bpmeds:prevstrk	bpmeds:glucose	diabp:cigpday	cigpday	sex:ldlc	age:prevmi	cigpday:prevmi	hdlc:ldlc
0.1506282	0.1492283	0.1189197	0.1155989	0.1151079	0.1141483	0.1097987	0.1051865	0.0923568
bmi:bpmeds	prevmi:prevstrk	heartrte:prevmi	glucose:prevmi	I(sysbp^2)	cursmoke:hdlc	age:heartrte	age:hdlc	
0.0855445	0.0699776	0.0645195	0.0588312	0.0585469	0.0566094	0.0556206	0.0510796	

Looking at the above table, removing highly insignificant continuous variate interactions `cigpday:heartrte` and `diabp:cigpday`, we have the following p-value from F-test.

```
# Remove as many insignificant continuous variate interactions as possible
anova(Mstep, update(Mstep,. ~ . - cigpday:heartrte - diabp:cigpday))$`Pr(>F)` [2]
```

```
## [1] 0.0729871
```

Assuming the insignificance threshold of 0.05, removing categorical/continuous variate interaction `bpmeds:prevstrk` results in the following p-value.

```
# Now remove insignificant interactions from categorical variates
anova(Mstep, update(Mstep,. ~ . - cigpday:heartrte - diabp:cigpday- bpmeds:prevstrk))$`Pr(>F)` [2]
```

```
## [1] 0.05655719
```

Since above p-value is just slightly greater than 0.05, removing the above interactions from stepwise model is insignificant. Therefore a reduced model can be obtained from stepwise in the following way.

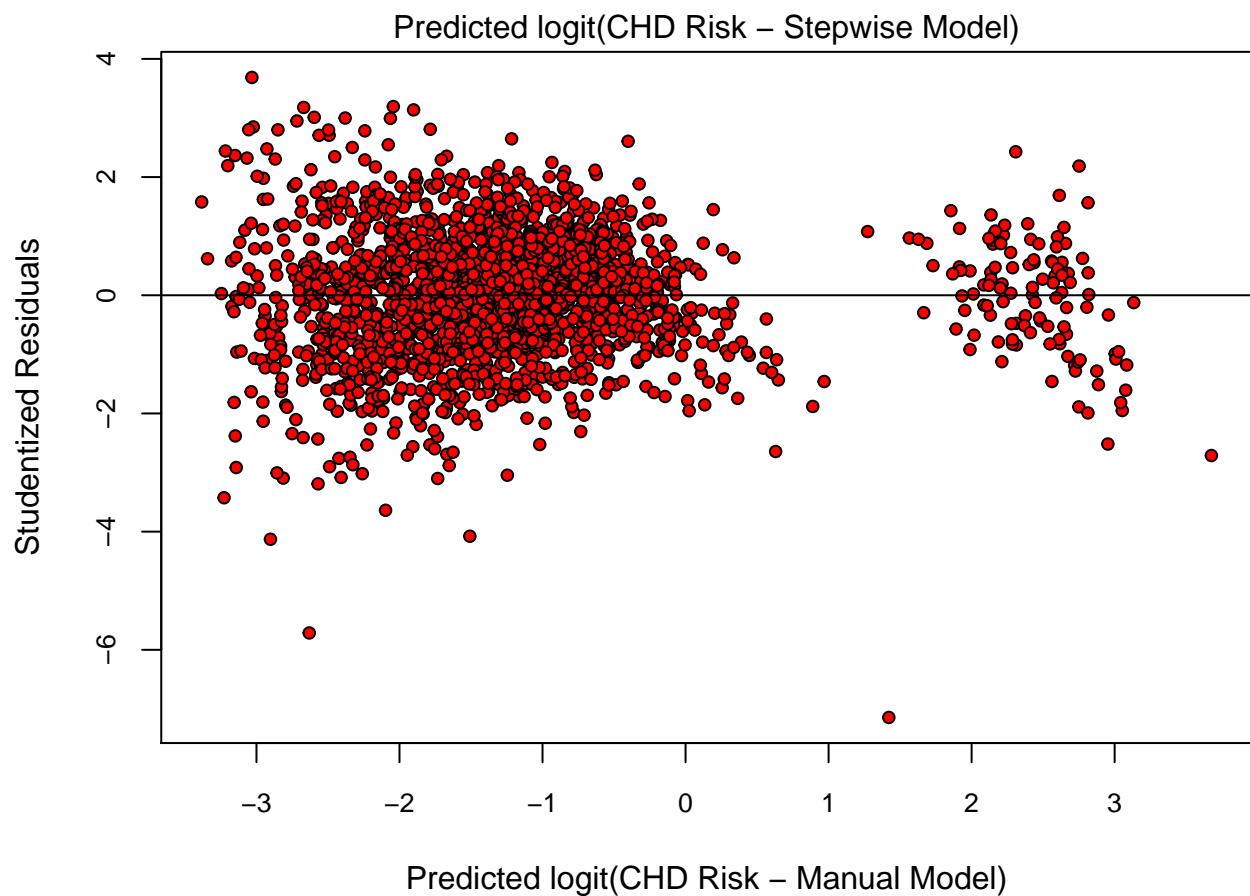
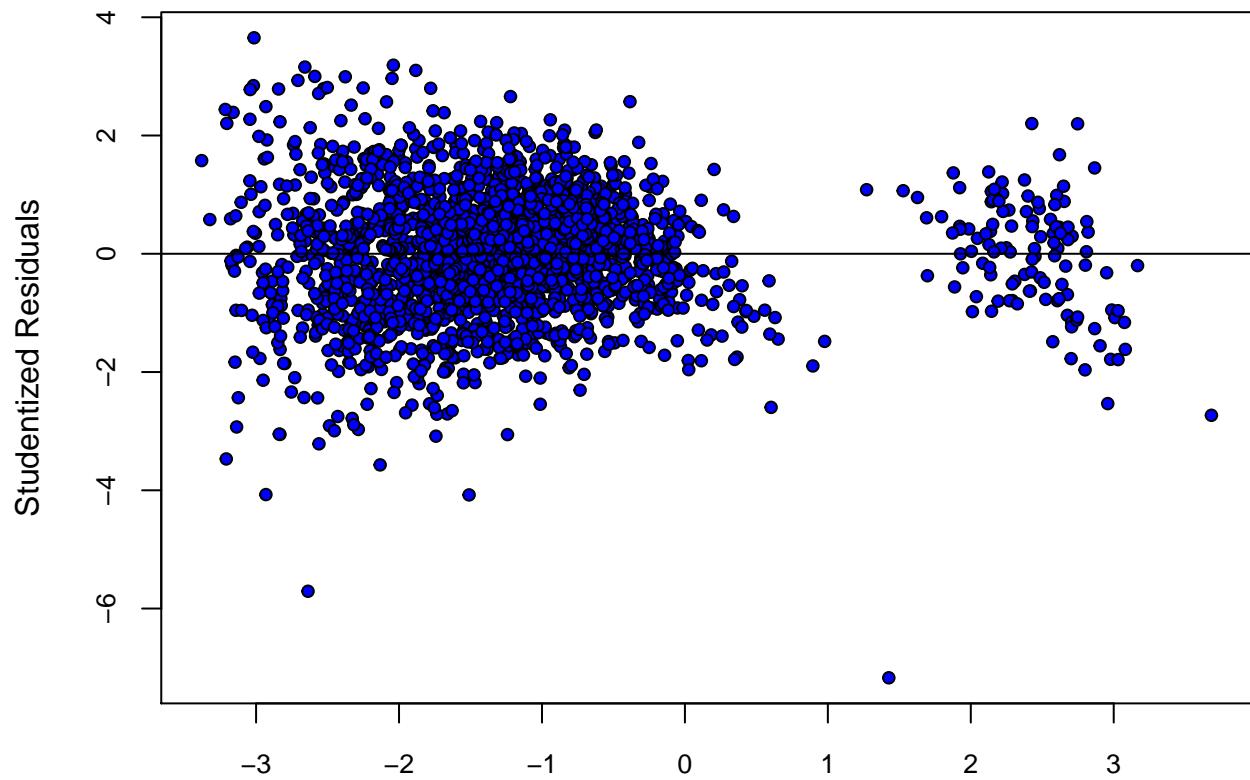
```
# Thus we have the following manually constructed model
Mmanual <- update(Mstep,. ~ . - cigpday:heartrte - diabp:cigpday - bpmeds:prevstrk)
```

4 Model Diagnostics

4.1 Residual Plots

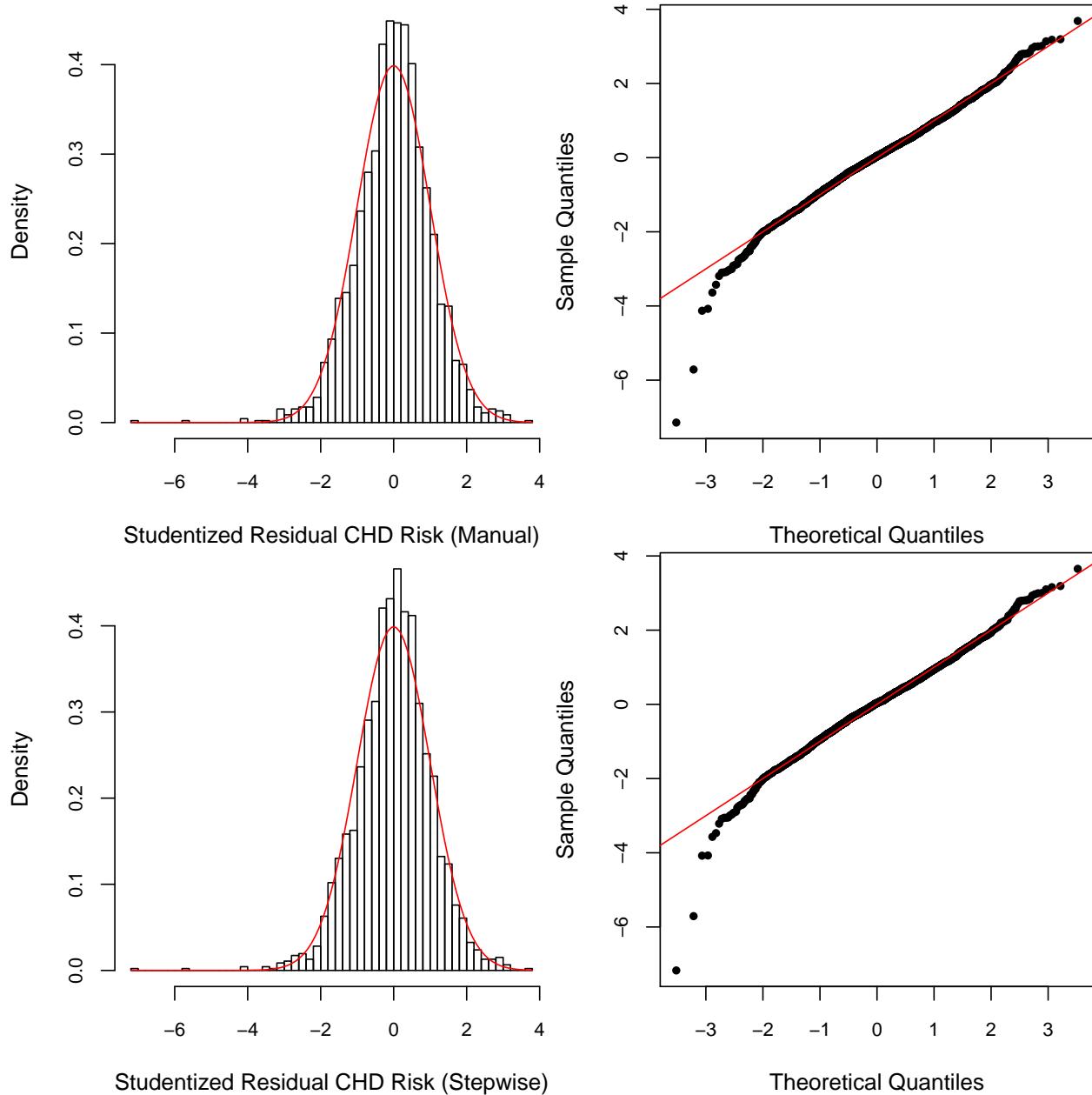
In this section we analyse the assumption that our residuals follow a normal distribution and check the homoscedasticity assumption.

First, we have that the most normal looking residuals assuming that the model is true, would be the studentized residuals on the standard deviation scale, so to check the homoscedasticity assumption we plot those values against the predicted values, as shown below:



Analysis of the graphs reveals that both models have very similar residual distributions, and for both, there seems to be a pattern of decreasing spread of residuals as the predicted logit value increases. Hence, we can conclude that both models are based on a violated homoscedasticity assumption, i.e., in light of the observed data there seems to be a change in the standard deviation of the response variate as the explanatory variables change.

Then to check our assumption of normality of residuals we plot the residuals on a QQPlot and a histogram:

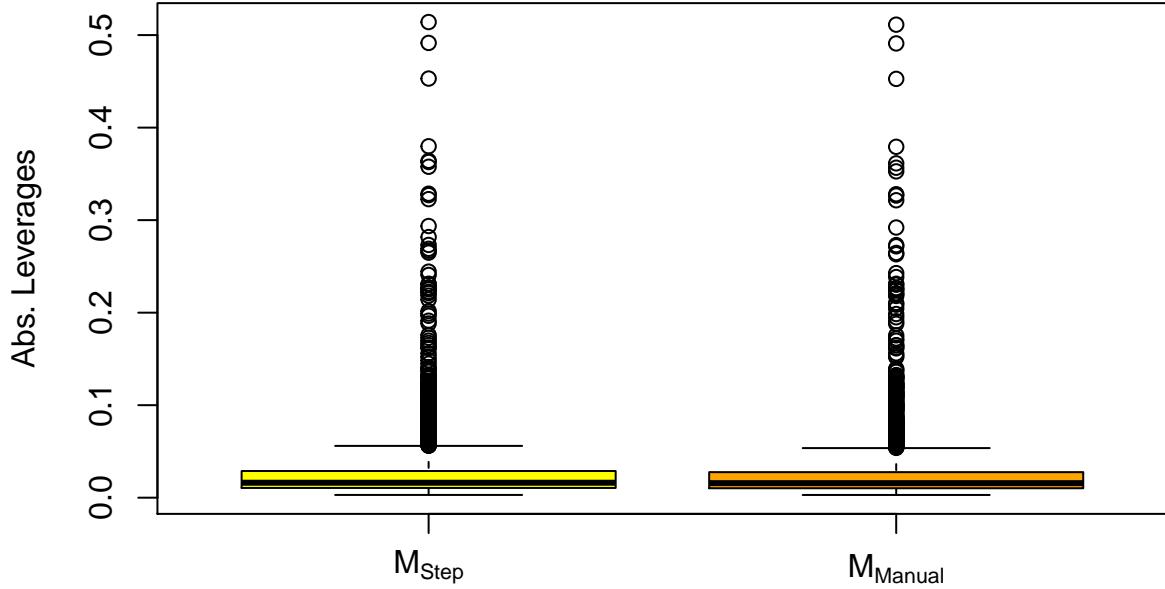


Again, from both plots we see a huge similarity between both models, and for both we seem to have a normal distribution being satisfied by the residuals. From the QQPlot, we can observe that most points lie on the theoretical line.

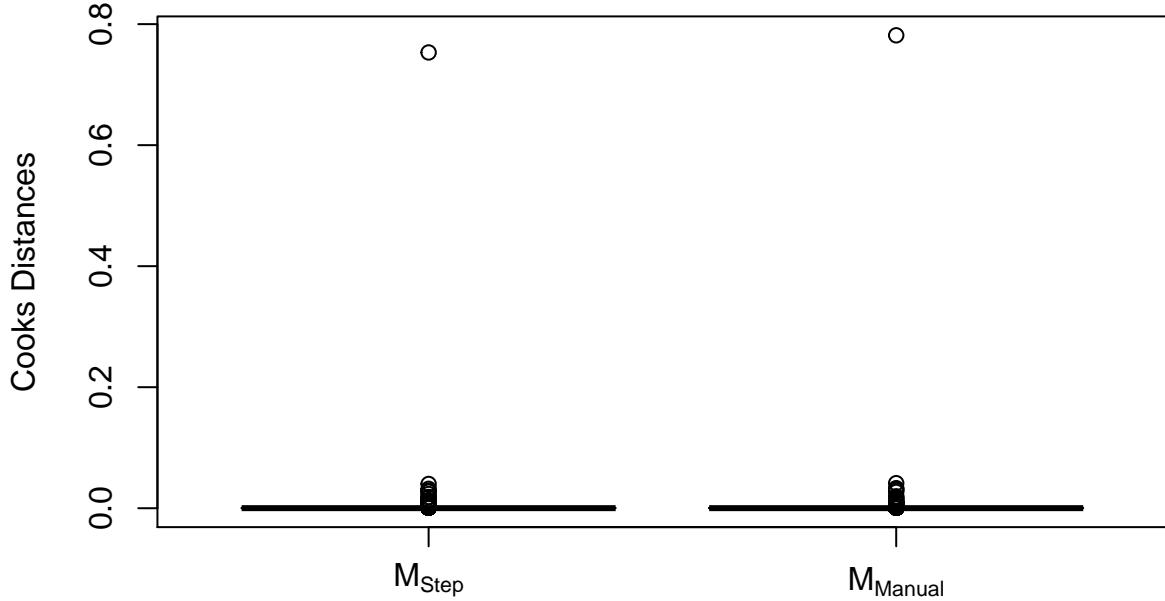
From this diagnostics there seems to not be a significant departure from our assumptions of homoscedasticity and normality of residuals.

4.2 Leverage and Influence Measures

We have the following boxplot of absolute values of leverages of both the step-wise and manual models.



Similarly, we have the following boxplot of cook's distances for both the models.



From the first plot above, we see that leverage for most observations is far from the desired value of 1 in both the models. Whereas from the second plot, cook's distance for most of the observations is close to desired value 0. And finally, both the models have very similar values for leverages and cook's distances.

5 Model Selection

5.1 Cross Validation

```
## Warning: package 'statmod' was built under R version 3.5.2
```

Before performing cross-validation analysis, function `logitnorm_mean` is created to approximate the conditional

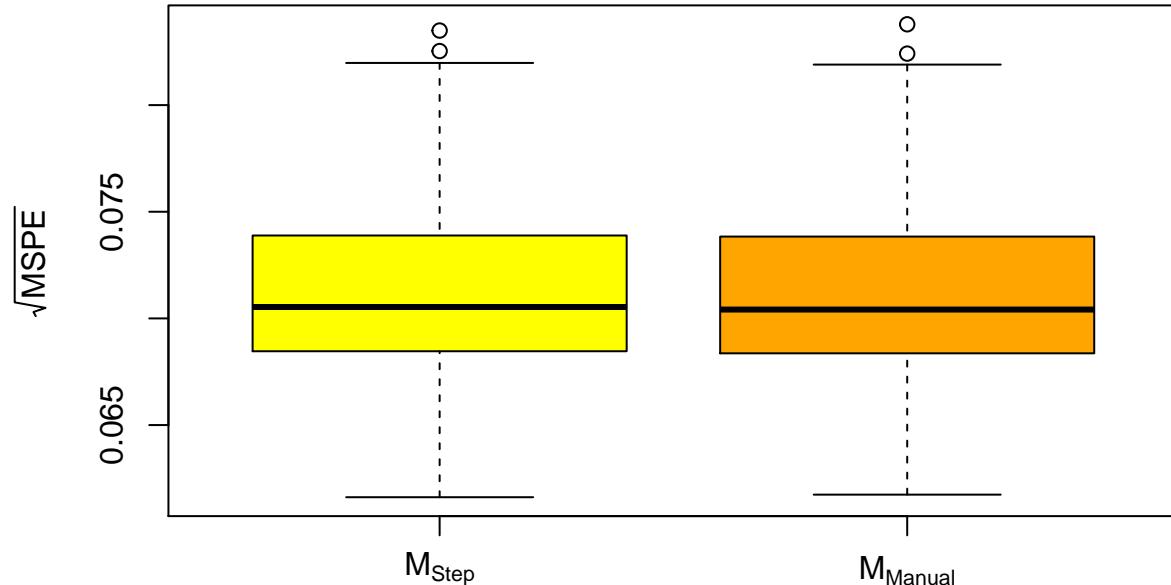
mean $E[\text{chdrisk}|\mathbf{x}]$ based on the regression model $\text{logit}(\text{chdrisk})|\mathbf{x} \sim N(\mathbf{x}'\beta, \sigma^2)$ (look into the Appendix for code). The following output is produced when tested.

```
# Test the function
mu <- c(0.7, 3.2, -1.1)
sigma <- c(0.8, 0.1, 2.3)
# Returns results expected in the project description
sapply(1:3, function(i) logitnorm_mean(mu[i], sigma[i]))
```

```
## [1] 0.6491002 0.9606606 0.3530580
```

The above function is then used to perform cross-validation analysis and the following boxplot that shows MSPE of the both the models is produced.

Root MSPE



Model chosen will be the manual one [ADD MORE] These are the parameter estimates, std.errors and p-values of the manually constructed model.

Table 5: Summary of chosen manually constructed model

	(Intercept)	sexMale	totchol	age	sysbp	diabp	cursmokeYes	cigpday	bmi
Estimate	-6.841121	0.7509308	0.0086801	0.0562327	0.0125123	-0.0426462	0.4903809	0.0463907	-0.0843445
Std. Error	1.017652	0.1622031	0.0029290	0.0101841	0.0061648	0.0106013	0.2180640	0.0087763	0.0244266
Pr(> t)	0.000000	0.0000039	0.0030739	0.0000000	0.0425109	0.0000594	0.0246223	0.0000001	0.0005647

	diabetesYes	bpmedsYes	heartrte	glucose	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc
Estimate	1.0311734	0.8595578	0.0440175	0.0023589	5.6983941	0.1404944	3.9567705	-0.0201085	-0.0063626
Std. Error	0.2918380	0.3038036	0.0078437	0.0026598	0.5685983	0.0792694	0.3583848	0.0084351	0.0035374
Pr(> t)	0.0004186	0.0047062	0.0000000	0.3752314	0.0000000	0.0764694	0.0000000	0.0172127	0.0722106

	I(hdlc^2)	I(bmi^2)	I(diabp^2)	I(sysbp^2)	sysbp:prevmiYes	totchol:prevhypYes	diabetesYes:prevmiYes	prevhypYes:ldlc	sysbp:prevhypYes
Estimate	0.0001820	0.0028956	5.97e-04	0.0000426	-0.0109765	-0.0070567	-0.6245291	0.0038926	-0.0087679
Std. Error	0.0000508	0.0004308	6.57e-05	0.0000219	0.0025360	0.0011582	0.1511829	0.0011277	0.0023122
Pr(> t)	0.0003497	0.0000000	0.00e+00	0.0520383	0.0000157	0.0000000	0.0000375	0.0005671	0.0001534

	totchol:heartrte	sysbp:diabetesYes	diabp:bmi	diabp:hdle	prevmiYes:hdle	prevmiYes:prevhypYes	sexMale:glucose	age:ldlc	age:heartrte
Estimate	-7.64e-05	-0.0061063	-0.0010390	-0.0002321	0.0149580	-0.2985780	-0.0018741	0.0000606	-0.0002377
Std. Error	1.86e-05	0.0017038	0.0002507	0.0000625	0.0038971	0.1305121	0.0007338	0.0000285	0.0000989
Pr(> t)	4.09e-05	0.0003457	0.0000353	0.0002120	0.0001274	0.0222452	0.0107144	0.0337545	0.0163262

	cigpday:hdle	bmi:ldlc	totchol:hdle	totchol:prevmiYes	sysbp:heartrte	sysbp:bpmedsYes	cursmokeYes:hdle	prevmiYes:prevstrkYes	diabetesYes:hdle
Estimate	-0.0003505	0.0001499	0.0001294	-0.0026466	-0.0001125	-0.0035185	0.0043414	-0.3881888	0.0058461
Std. Error	0.0000943	0.0000589	0.0000504	0.0009274	0.0000375	0.0015038	0.0023248	0.1951360	0.0024167
Pr(> t)	0.0002059	0.0110647	0.0102686	0.0043574	0.0027333	0.0193871	0.0619759	0.0467867	0.0156403

	sexMale:sysbp	cigpday:glucose	heartrte:glucose	diabp:glucose	cursmokeYes:ldlc	age:cigpday	age:hdle	hdle:ldlc	age:prevhypYes
Estimate	-0.0018228	-0.0000755	0.0000695	-0.0000682	-0.0010001	-0.0003358	-0.0001725	-0.0000803	-0.0103638
Std. Error	0.0009765	0.0000385	0.0000260	0.0000281	0.0004913	0.0001248	0.0000888	0.0000478	0.0029473
Pr(> t)	0.0620856	0.0500605	0.0075675	0.0154340	0.0418854	0.0071905	0.0522339	0.0927933	0.0004462

	diabp:prevhypYes	diabp:cursmokeYes	bmi:bpmedsYes	bpmedsYes:glucose	age:prevmiYes	sexMale:ldlc	cigpday:prevmiYes	glucose:prevmiYes	heartrte:prevmiYes
Estimate	-0.0110683	-0.0057752	-0.0128143	0.0014572	-0.0105250	0.0007191	-0.0087788	-0.0027211	0.0058756
Std. Error	0.0034311	0.0019148	0.0072263	0.0009960	0.0063892	0.0004619	0.0049761	0.0015063	0.0033641
Pr(> t)	0.0012740	0.0025893	0.0763179	0.1436117	0.0996373	0.1196056	0.0778372	0.0709783	0.0808493

6 Discussion

From our analysis, we make the following conclusions :

1. Any conclusions we make might be more biased towards females due to the significantly higher females than males in the study.
2. From our pair plots we can observe that high HDL cholesterol levels have people with less risk for heart disease. So it seems like higher HDL cholesterol levels are associated with less expected CHD risk. This effect can be supported by the relatively low p-value for the significance of hdlc in the presence of other variates in the model. On the contrary, people who have had previous myocardial infarctions, strokes, or hypertension seem to have a higher expected CHD risk, as shown by the relatively high positive estimated coefficients and very low p-values.
3. We can observe from the model diagnostics of the chosen model that there might be a slight departure from the assumption of homoscedasticity. In fact, it seems as though the standard deviation decreases as we get into the region of people with high risk of coronary heart disease. This might be due to the fact that we have relatively more variates that have a positive association with higher CHD risk and also less people with high CHD risk, hence we have more predictive power when presented with a person with factors of high CHD risk. However, for people with low CHD, it might be harder to predict their CHD risk due to more variation that arises from having a lot more of people with relatively low CHD risk in the study. Therefore, our conclusions regarding factors that relate to high CHD risk might be stronger than those that relate to lower CHD risks.