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. First, descriptive analysis is performed, in which collinearity is found between serum total cholestrol and low density lipoprotein cholesterol. Moreover, it is observed that there are significantly higher number of females in the study than males (around 30%), as well as relations between having previous health complications, like stokes or hypertensions, with high CHD risk. 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	cigpday	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 mean coronary heart disease (CHD) risk against certain categorical variates, gives more insights.

For instance, if we take a look at mean 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
```

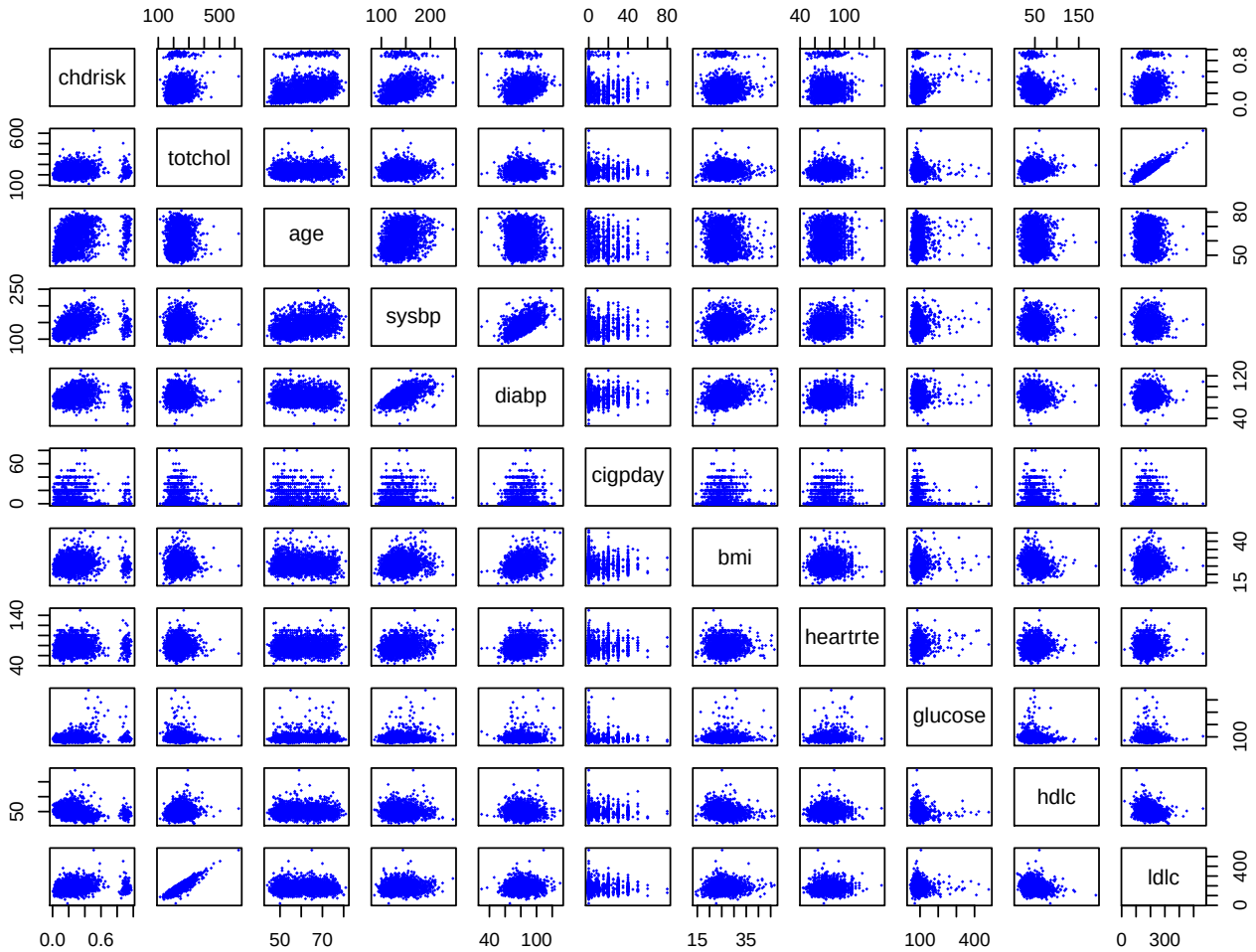
Indeed, we have that mean CHD risk given that a person has hypertension is significantly higher than the mean for people who did not have hypertension.

```
## 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 variates i.e. all variates excluding logical variates such as whether or not currently a cigarette smoker.

Pair Plots of Continuous Variates



From the pair plots, we can observe a strong linear relationship between low density lipoprotein cholesterol and serum total cholestrol. As total cholestrol increases, low density lipoprotein cholosterol 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 at the same time.

Now take a look at the VIFs of these variates.

Table 2: VIFs of Variates

sexMale	totchol	age	sysbp	diabp	cursmokeYes	cigpday	bmi	diabetesYes	bpmedsYes	heartrte	glucose	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc
1.225	10.635	1.49	2.919	2.406	2.979	2.974	1.182	1.286	1.215	1.106	1.309	1.067	1.046	1.823	2.288	10.368

We observe VIF values higher than 10 for both serum total cholestrol and low density lipoprotein cholesterol, which means they have significant colinearity with other variates.

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 a lot of variates in our final model relatively and also we capture as many necessary variates as possible.

We first try our initial and maximum models as follows.

```
## lm(formula = I(logit(chdrisk)) ~ 1, data = fhsd)
```

```
## lm(formula = I(logit(chdrisk)) ~ (.)^2, data = fhsd)
```

However, we end up getting NAs in the coefficients for two interactions namely: whether currently a cigrette smoker and number of cigarettes smoked each day, whether individual is on anti-hypertensive medication and whether the individual actually has 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 these two respective columns linearly dependent.

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

Table 3: 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
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 4: bpmeds against prevhyp

	No	Yes
No	957	1016
Yes	0	333

To fix these we remove these two interactions from the maximal model and add some quadratic terms for continuous variates in hope of having some additional predicitive power.

```
## lm(formula = 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)
```

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 insignifance when an F-test is perfomed by removing them solely from the model along with corresponding p-values in a sorted order.

Table 5: 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 between heart rate and number of cigarettes per day, between diastolic blood pressure and number of cigarettes per day 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 between whether an individual is on anti-hypertensive medication and whether an individual has had the stroke results in the following p-value.

```
# Now remove insignificant interactions from categroical 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 3 interactions from stepwise model is insiginificant. Therefore a reduced model can be obtained from stepwise in the following way.

```
## 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 +
```

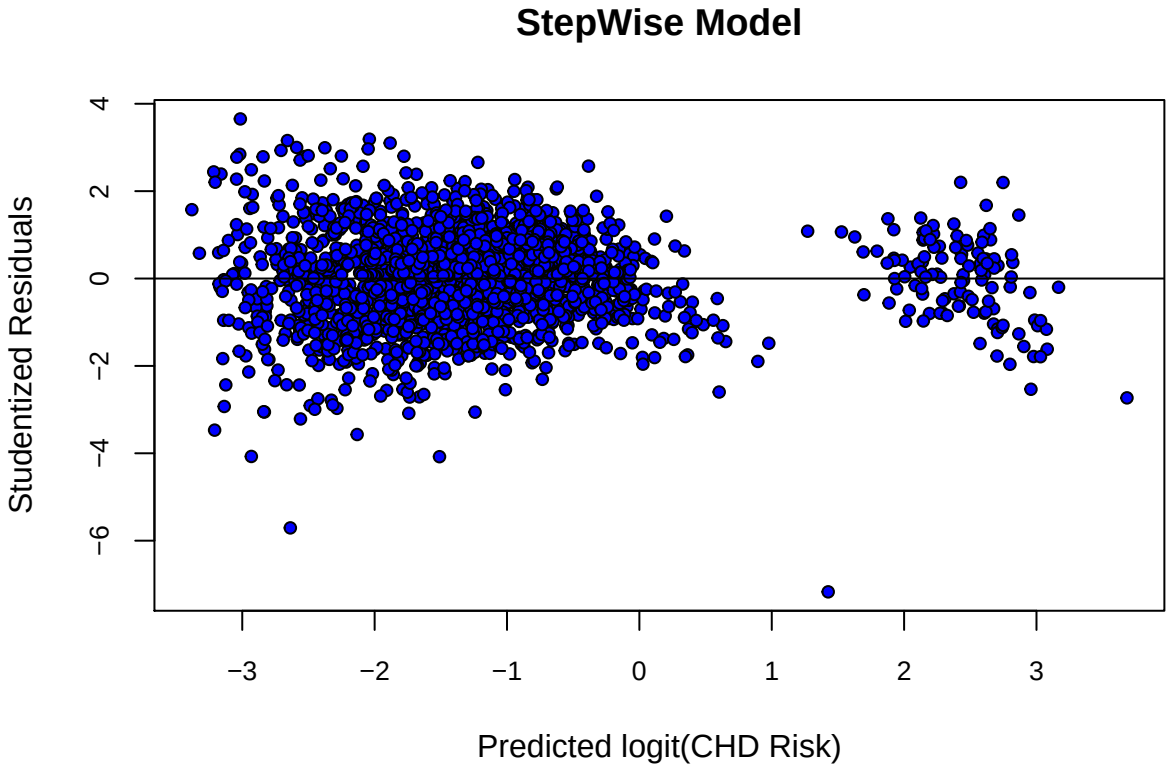
```
##     cursmoke:hdlc + prevmi:prevstrk + diabetes:hdlc + sex:sysbp +
##     cigpday:glucose + heartrte:glucose + diabp:glucose + cursmoke:ldlc +
##     age:cigpday + age:hdlc + hdlc:ldlc + age:prevhyp + diabp:prevhyp +
##     diabp:cursmoke + bmi:bpmeds + bpmeds:glucose + age:prevmi +
##     sex:ldlc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

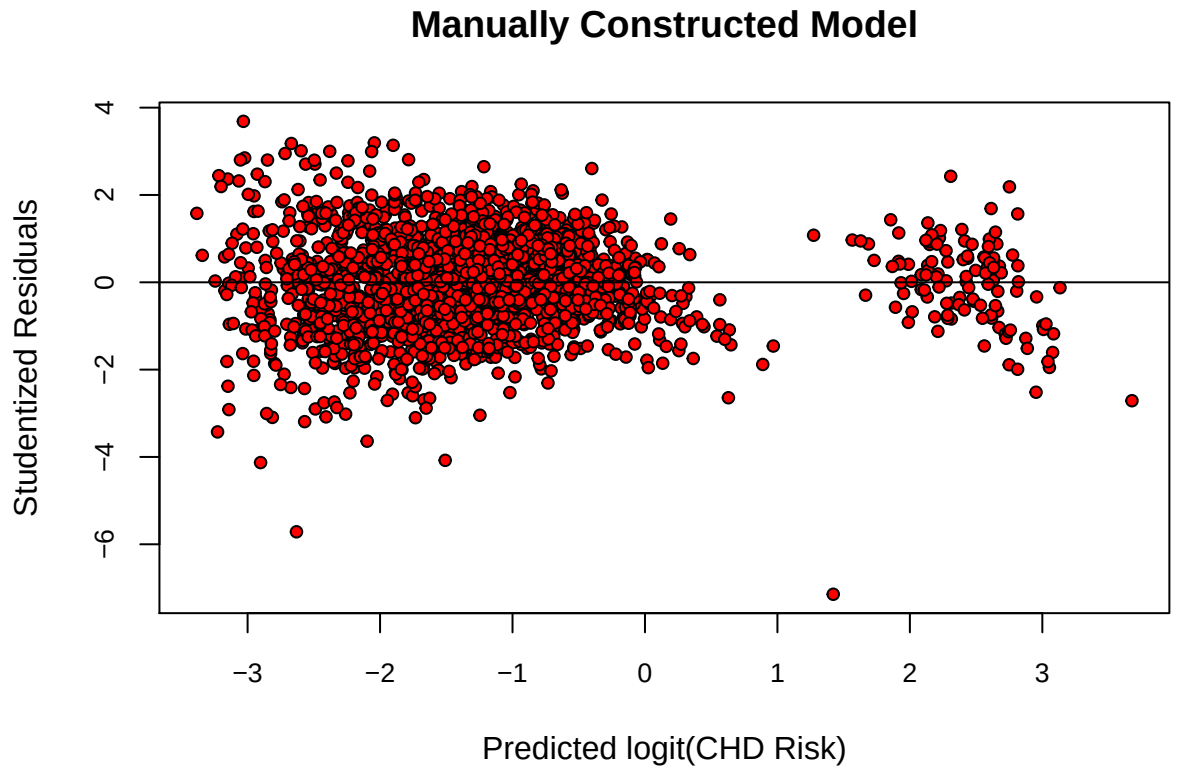
4 Model Diagnostics

4.1 Residual Plots

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

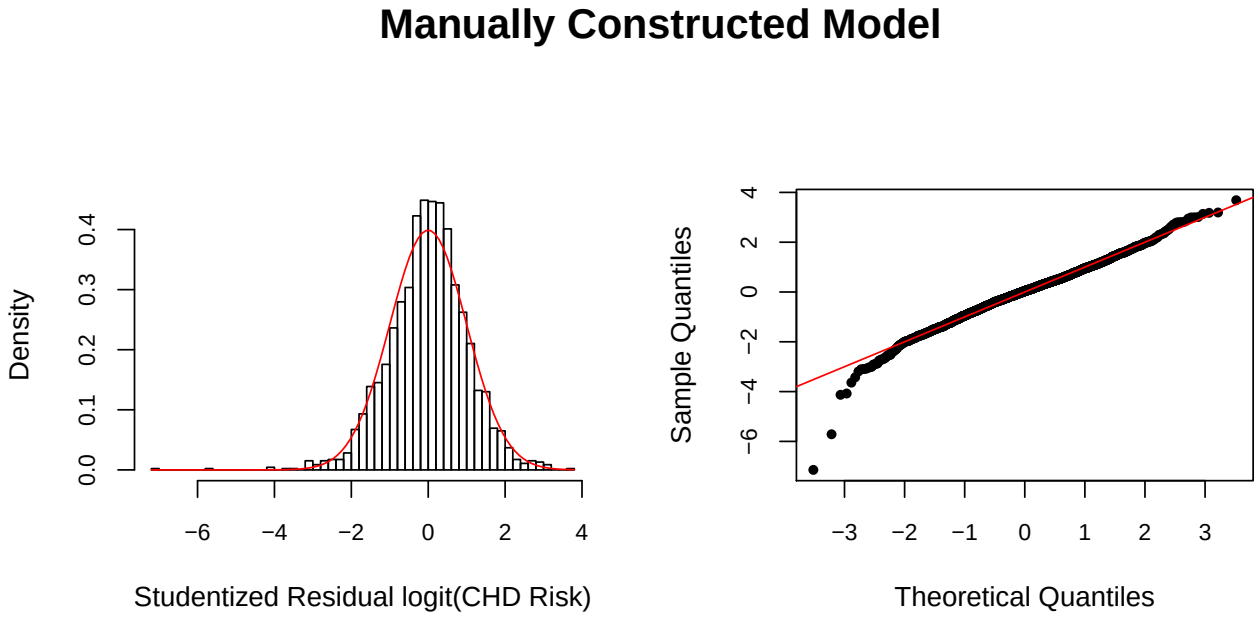
First, we have that the most normal looking residuals assuming that the model is true, would be the studentized residuals on the standard deviation scale, so to check the homoscedasticity assumption we plot those values against the predicted values, as shown below:

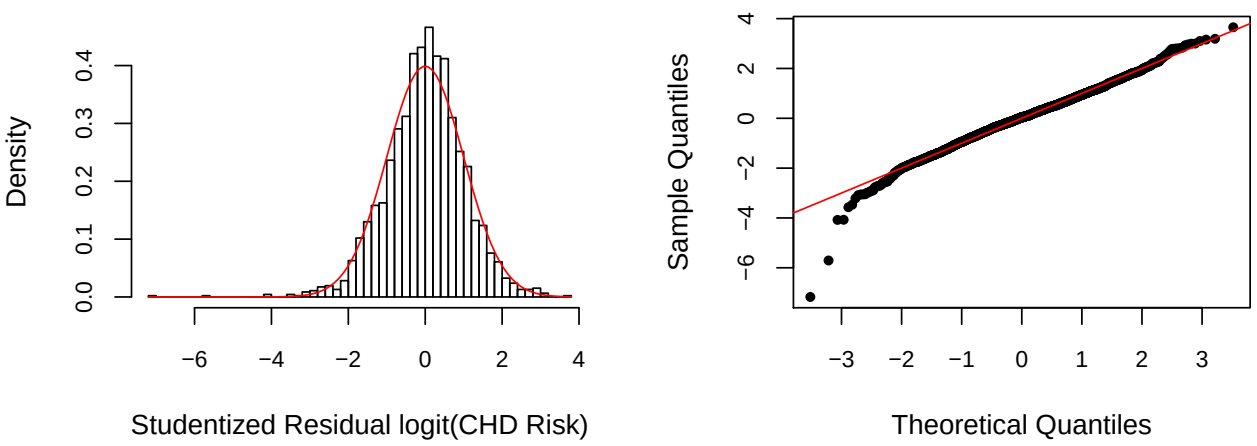




Analysis of the plots reveals that both models have very similar residual distributions, and for both, there seems to be a pattern of decreasing spread of residuals as the predicted logit value increases. Hence, we can conclude that both models are based on a violated homoscedasity assumption, i.e., 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 histogram and a QQPlot:



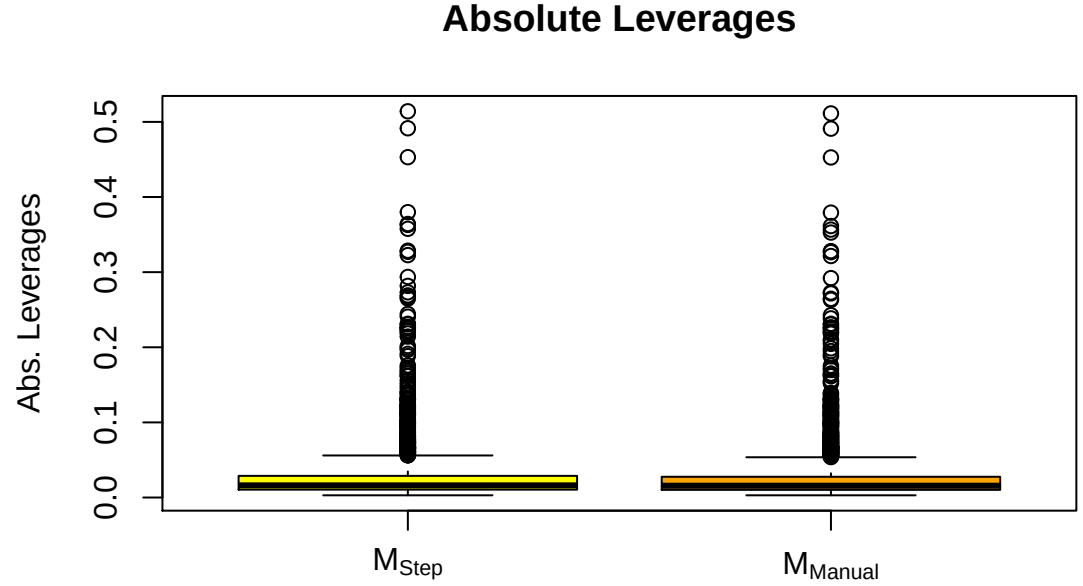


Again, from both plots we see a huge similarity between both models, and for both we seem to have an approximately normal distribution being satisfied by the residuals. From the QQPlot, we can observe that most observations lie on the theoretical line.

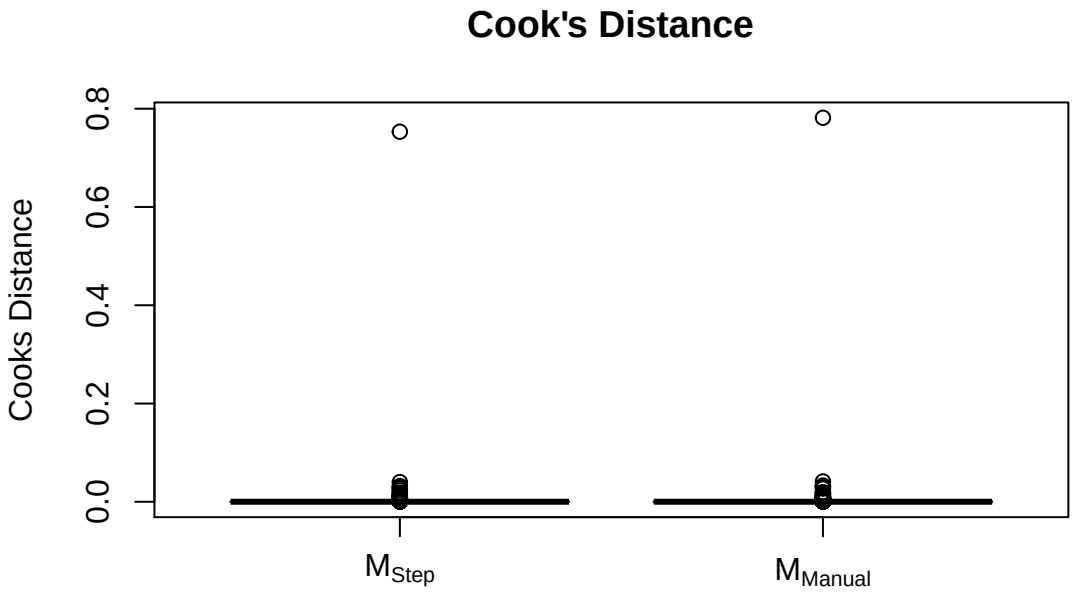
From this diagnostics there seems to not be a significant departure from our assumptions of homoscedasticity and normality of residuals.

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

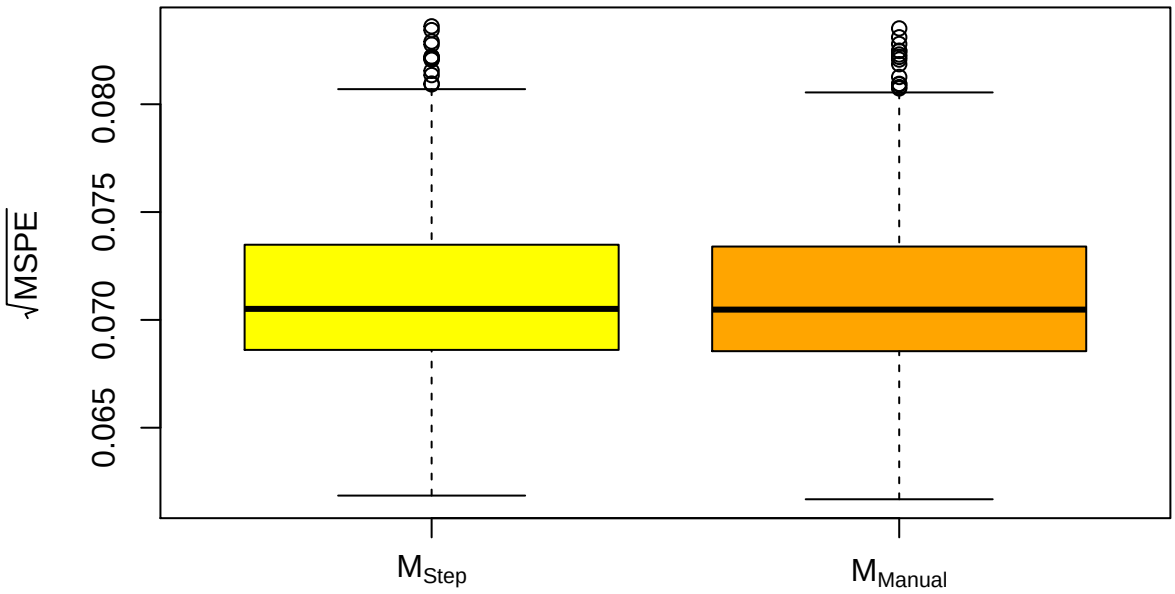
```
## Warning: package 'statmod' was built under R version 3.5.2
```

Before performing cross-valiation analysis, function `logitnorm_mean` is created to approximate the conditional mean `E[chdrisk|x]` based on the regression model logit(`chdrisk`)|x $\sim N(x'\beta, \sigma^2)$ (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 desciption
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.



From previous discussions about residuals, leverages and influence measures, both models seem to be almost identical. It can observed from the above boxplot that the manually constructed model has slightly lower MSPE values than the stepwise model and therefore has slightly better predictive power. Also, the manual model has lesser covariate interactions than the stepwise model and hence has better explanatory power as well. Therefore, we select the manually constructed model over stepwise.

These are the parameter estimates, std.errors and p-values of the manually constructed model.

Table 6: Summary of chosen manually constructed model

	(Intercept)	sexMale	totchol	age	sysbp	diabp	cursmokeYes	cigpday	bmi
Estimate	-6.8411	0.7509	0.0087	0.0562	0.0125	-0.0426	0.4904	0.0464	-0.0843
Std. Error	1.0177	0.1622	0.0029	0.0102	0.0062	0.0106	0.2181	0.0088	0.0244
Pr(>|t|)	0.0000	0.0000	0.0031	0.0000	0.0425	0.0001	0.0246	0.0000	0.0006

	diabetesYes	bpmedsYes	heartrte	glucose	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc
Estimate	1.0312	0.8596	0.0440	0.0024	5.6984	0.1405	3.9568	-0.0201	-0.0064
Std. Error	0.2918	0.3038	0.0078	0.0027	0.5686	0.0793	0.3584	0.0084	0.0035
Pr(>|t|)	0.0004	0.0047	0.0000	0.3752	0.0000	0.0765	0.0000	0.0172	0.0722

	I(hdlc^2)	I(bmi^2)	I(diabp^2)	I(sysbp^2)	sysbp:prevmiYes	totchol:prevhypYes	diabetesYes:prevmiYes	prevhypYes:ldlc	sysbp:prevhypYes
Estimate	2e-04	0.0029	6e-04	0.000	-0.0110	-0.0071	-0.6245	0.0039	-0.0088
Std. Error	1e-04	0.0004	1e-04	0.000	0.0025	0.0012	0.1512	0.0011	0.0023
Pr(>|t|)	3e-04	0.0000	0e+00	0.052	0.0000	0.0000	0.0000	0.0006	0.0002

	totchol:heartrte	sysbp:diabetesYes	diabp:bmi	diabp:hdlc	prevmiYes:hdlc	prevmiYes:prevhypYes	sexMale:glucose	age:ldlc	age:heartrte
Estimate	-1e-04	-0.0061	-1e-03	-2e-04	0.0150	-0.2986	-0.0019	0.0001	-0.0002
Std. Error	0e+00	0.0017	3e-04	1e-04	0.0039	0.1305	0.0007	0.0000	0.0001
Pr(>|t|)	0e+00	0.0003	0e+00	2e-04	0.0001	0.0222	0.0107	0.0338	0.0163

	cigpday:hdlc	bmi:ldlc	totchol:hdlc	totchol:prevmiYes	sysbp:heartrte	sysbp:bpmedsYes	cursmokeYes:hdlc	prevmiYes:prevstrkYes	diabetesYes:hdlc
Estimate	-4e-04	0.0001	0.0001	-0.0026	-0.0001	-0.0035	0.0043	-0.3882	0.0058
Std. Error	1e-04	0.0001	0.0001	0.0009	0.0000	0.0015	0.0023	0.1951	0.0024
Pr(>|t|)	2e-04	0.0111	0.0103	0.0044	0.0027	0.0194	0.0620	0.0468	0.0156

	sexMale:sysbp	cigpday:glucose	heartrte:glucose	diabp:glucose	cursmokeYes:ldlc	age:cigpday	age:hdlc	hdlc:ldlc	age:prevhypYes
Estimate	-0.0018	-0.0001	0.0001	-0.0001	-0.0010	-0.0003	-0.0002	-0.0001	-0.0104
Std. Error	0.0010	0.0000	0.0000	0.0000	0.0005	0.0001	0.0001	0.0000	0.0029
Pr(>|t|)	0.0621	0.0501	0.0076	0.0154	0.0419	0.0072	0.0522	0.0928	0.0004

	diabp:prevhypYes	diabp:cursmokeYes	bmi:bpmedsYes	bpmedsYes:glucose	age:prevmiYes	sexMale:ldlc	cigpday:prevmiYes	glucose:prevmiYes	heartrte:prevmiYes
Estimate	-0.0111	-0.0058	-0.0128	0.0015	-0.0105	0.0007	-0.0088	-0.0027	0.0059
Std. Error	0.0034	0.0019	0.0072	0.0010	0.0064	0.0005	0.0050	0.0015	0.0034
Pr(>|t|)	0.0013	0.0026	0.0763	0.1436	0.0996	0.1196	0.0778	0.0710	0.0808

6 Discussion

From our analysis, we make the following conclusions :

1. Any conclusions we make might be more biased towards females due to the signifcantly higher females than males in the study.
2. From our pair plots we can observe that people with high HDL cholestrol levels are at less risk for heart disease. So it seems like higher HDL cholestrol levels are associated with less expected CHD risk. This effect can be supported by the relatively low p-value for the significance of high density lipoprotein cholestrol in the presence of other variates in the model. On the contrary, people who have had previous mycordial infractions, 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 number of 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.
4. From the box plot of cook's distance for the manual model, it can be seen that one specific individual has an unusually high value. The following command finds that outlying observation and the corresponding cook's distance.

```
cook2[which.max(cook2)]
```

```
##     916
## 0.78151
```

Since 0.78 is an unusually high cook's distance value, the 916th individual can be excluded from the analysis.

5. This is an observational study, which means certain conclusions regarding causations can not be made. Therefore, suggesting any behavioural changes based on this report would not be justified because none of the conclusions we established are causal, e.g. we can not say that smoking less would result in a decreased risk of heart disease.
6. Finally, we do observe some variates retained in the final model with high p-values, meaning that they might be insignificant. For example, casual serum glucose has a very high p-value of 0.37. However we know that this model is derived from the stepwise model which rules out all insignificant covariates. Hence, the calculated p-value(s) for the final model are incorrect displaying post-selection inference problem as described in the course notes.