STAT331 Final Project (Winter 2020)

Jia Yi Chan (20783900), Joh Ann Lee (20807848)

15/04/2020

Summary

The objective of this report is to explore the relationship between the risk score for CHD and various explanatory variables. To determine a suitable linear model involving the main effects and interaction effects of these variables, we began by summarizing the data and building plots to visualize their relationship with one another. Then, utilizing automated and manual methods, two candidate models were constructed to perform in-depth comparison using residual plots, leverage and influence measures, and cross-validation analysis. We chose to retain the automated model (via backward elimination) as our final model mainly because it has a higher predictive power, with both models having similar explanatory power. Based on the final model, the most important factors associated with high CHD risk are age, heartrate, and whether an individual has had hypertension or a myocardial infarction. However, note that our dataset only consisted of inidividuals who are middle-aged or older. Hence, our model may not apply to younger individuals as a result of this. In addition to that, our given dataset had a very small pool of people who have had stroke before, and so our model may not accurately reflect the CHD risk of this segment of the population.

2. Descriptive Statistics

2.1 Data Summary

We wish to select two linear regression models in order to explore the relation between the risk score for CHD and some explanatory variables. Prior to model fitting, we would first like to get a basic sense of the dataset variables by analyzing its summary and its pair plots.

We will only be displaying selected variables that provide interesting perspectives to our dataset in order to save space.

	chdrisk	totchol	age	cursmoke	cigpday	prevstrk	ldlc
	Min. :0.0050	Min. :112.0	Min. :44.00	No :1504	Min. : 0.00	No :2260	Min. : 20.0
	1st Qu.:0.1320	1st Qu.:207.0	1st Qu.:53.00	Yes: 802	1st Qu.: 0.00	Yes: 46	1st Qu.:152.0
	Median :0.2240	Median :235.5	Median :60.00	NA	Median : 0.00	NA	Median :180.0
	Mean :0.2655	Mean :237.8	Mean :60.23	NA	Mean : 6.84	NA	Mean :183.1
	3rd Qu.:0.3448	3rd Qu.:265.0	3rd Qu.:67.00	NA	3rd Qu.:10.00	NA	3rd Qu.:210.0
	Max. :0.9770	Max. :625.0	Max. :81.00	NA	Max. :80.00	NA	Max. :565.0

Table 1: Selected Summary Statistics of the Framingham Heart Study Dataset

- `age`: The minimum age of all individuals who are participating in the study is 44 years old. This means our given dataset is only based on individuals who are middle-aged or older. Hence, the relation between `age` and `chdrisk` cannot be summarized for all ages (i.e. our final model can only be applied on individuals who are at least middle-aged).
- `cursmoke` and `cigpday`: Upon excluding non-smokers in the calculation for mean, our true mean for `cigpday` is now approximately 19.7. This suggests that we may not be able to fully explore the relation that depends solely on `cigpday`, as those who do smoke tend to smoke a significant amount each day.
- `prevstrk`: As our sample size for individuals who have had a stroke is very small (46), we may not be able to draw conclusions for this segment of the population.

2.2 Paired Plots

Next, let us take a look at the paired plots consisting of non-categorical data:

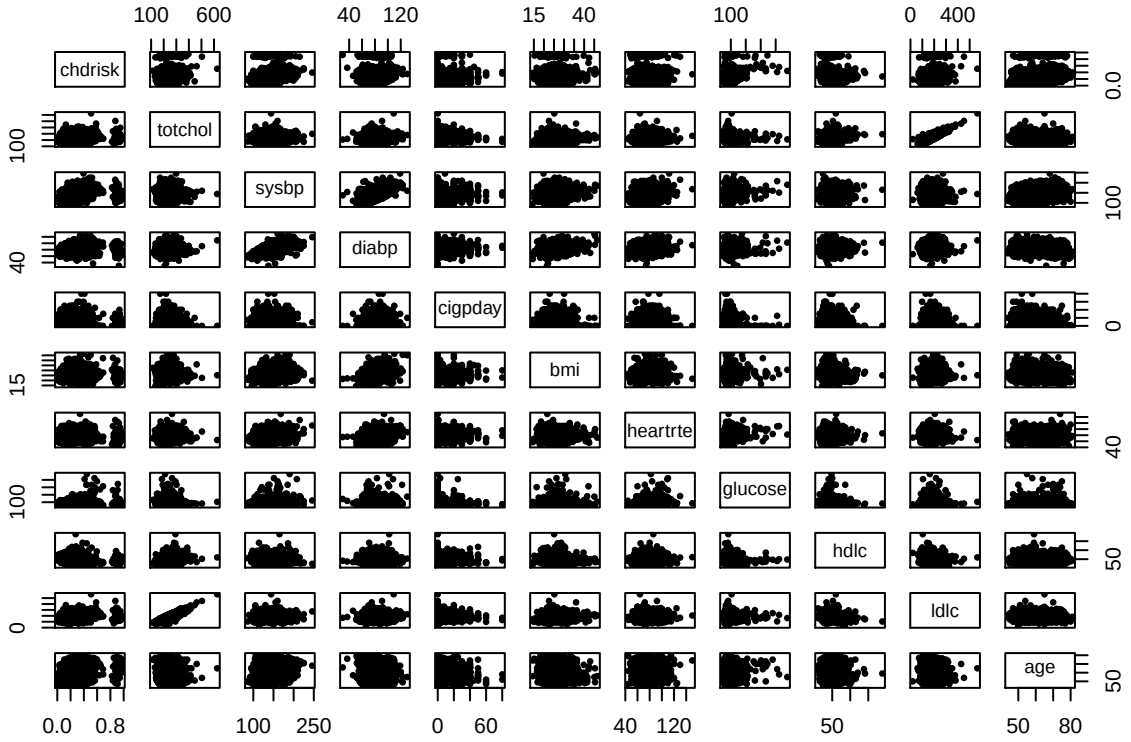


Figure 1: Paired Plots

There are a few couple of facts that are revealed by the pair plots (Fig 1):

- There are vertical lines in the paired plots involving `cigpday`. This makes sense as the number of cigarettes smoked per day is a discrete variable.
- There is a linear relationship between `totchol` and `ldlc` as shown in the paired plot between `totchol` and `ldlc`.
- There is a linear relationship between `sysbp` and `diabp` as shown in the paired plot between `sysbp` and `diabp`.
- There may be an inverse relationship between `glucose` and `hdlc` as the points in the paired plot seem to be following that of an inverse graph.
- The paired plots for `chdrisk` consist of two groups of data points that have a gap in between, suggesting that a transformation may be needed to bring the groups of data points closer to one another. This might provide a clearer picture of certain trends.

2.3 Variance Inflation Factor

In our last subsection, we will diagnose colinearity by calculating the variance inflation factor of our variables.

	totchol	age	sysbp	diabp	cigpday	bmi	heartrte	glucose	hdlc	ldlc
VIF	10.5	1.44	2.46	2.31	1.1	1.15	1.09	1.08	2.19	10.3

Table 2: VIF of the Explanatory Variables in the Dataset

Observe that the variance inflation factor for `totchol` and `ldlc` are both more than 10. This means that `totchol` and `ldlc` are both highly correlated with other explanatory variables. This is a cause for concern as the regression may have trouble figuring out if the change in the reponse variable is due to one covariate or the other.

As `totchol` variable has the highest variance inflation factor, we exclude `totchol` and rerun our computation:

	age	sysbp	diabp	cigpday	bmi	heartrte	glucose	hdlc	ldlc
VIF_rerun	1.43	2.45	2.29	1.1	1.15	1.09	1.08	1.07	1.04

Table 3: VIF of the Explanatory Variables in the Dataset (Without 'totchol')

Upon removing `totchol`, observe that the variance inflation factor for `ldlc` has now gone down to 1.039828. This suggests that `ldlc` is highly correlated with `totchol.` This makes sense as low density lipoprotein cholesterol (`ldlc`) is included in the calculation for serum total cholesterol(`totchol`).

3. Candidate Models

From the pair plot (Fig. 1) in the previous section, an inference regarding the need for a transformation in producing a model with a better fit was made. We decided on using the following transformation:

$$logit(chdrisk) \leftarrow log(chdrisk) - log(1 - chdrisk)$$

Using $logit(chdrisk)$ as the response variable, we will now create two candidate models.

3.1 Automated Model Selection

Before performing automated model selection, we will first investigate interaction terms that may result in coefficients that are NA:

```
## [1] "cursmokeYes:cigpday"  "bpmedsYes:prevhypYes"
```

- Only smokers (`cursmoke = Yes`) will smoke a non-zero number of cigarettes (`cigpday`), which is also supported by the given data. Hence, there is no interaction effect between `cigpday` and `cursmoke` since those with a non-zero value for `cigpday` are the same group of people who smoke cigarettes.
- There is no reason for individuals who do not have hypertension to take blood pressure medication. Hence, it makes sense to eliminate the possibility of an interaction effect since individuals who take blood pressure medication (`bpmeds = Yes`) are the same group of people who have hypertension (`prevhyp = Yes`).

Then, we look for nonlinear terms:

- As seen in the plot between `logit_chdrisk` and `sysbp`, there seems to be a nonlinear relationship, which is further supported by fitting a quadratic curve on the graph.

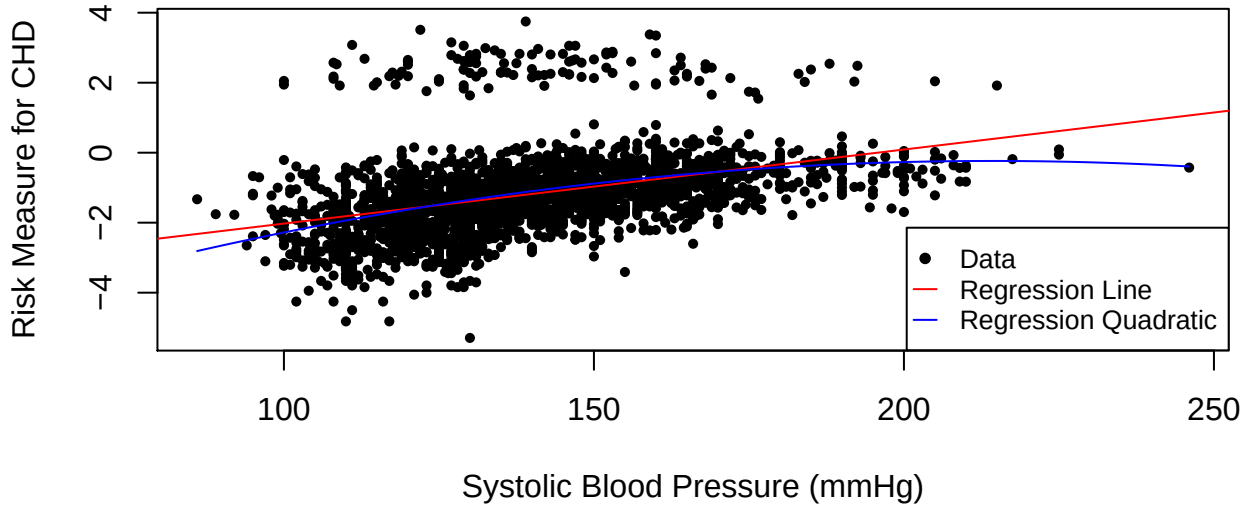


Figure 2: Non-linear Curve of CHD Risk against Systolic Blood Pressure

After eliminating these interaction terms and adding our quadratic term, we will also remove `totchol` due to its high collinearity.

We can now proceed to obtain and compare our automated models. Our choice of inputs to the automated selection procedures include:

$M0 \leftarrow lm(logit\_chdrisk \sim 1,\ data = fhs)$

$Mstart \leftarrow lm(logit\_chdrisk \sim\ . - chdrisk - totchol,\ data = fhs)$

$Mmax \leftarrow lm(logit\_chdrisk \sim (. - chdrisk)^2 - totchol - cursmoke : cigpday - bpmeds : prevhyp + I(sysbp^2), data = fhs)$

where M_0 is the minimal model that only includes intercepts, M_{start} is the initial model used in Stepwise selection (includes main effects only), M_{max} is the maximal model that includes all main effects and interactions, an additional `I(sysbp^2)` variate, but without the interaction effects between `cursmoke` and `cigpday`, along with `bpmeds` and `prevhyp`.

We ultimately decided on the Backward model by the process of elimination. Here is our first candidate model obtained via backward elimination:

$$\begin{aligned} M_{back} \leftarrow lm(formula = logit\_chdrisk \sim\ & sex + age + sysbp + diabp + cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte + \\ & glucose + prevmi + prevstrk + prevhyp + hdlc + ldlc + sex : totchol + sex : sysbp + sex : glucose + \\ & sex : prevstrk + sex : prevhyp + age : totchol + heartrte : totchol + prevhyp : totchol + hdlc : totchol + \\ & ldlc : totchol + age : diabp + age : cursmoke + hdlc : ldlc + age : heartrte + age : prevmi + age : prevhyp + \\ & age : hdlc + sysbp : diabp + sysbp : cigpday + sysbp : diabetes + sysbp : bpmeds + sysbp : heartrte + \\ & sysbp : prevmi + sysbp : prevhyp + diabp : cursmoke + diabp : cigpday + diabp : bpmeds + diabp : glucose + \\ & diabp : prevhyp + diabp : hdlc + cursmoke : bmi + cursmoke : hdlc + cigpday : bmi + cigpday : glucose + \\ & cigpday : hdlc + cigpday : ldlc + bmi : bpmeds + bmi : prevmi + bmi : ldlc + diabetes : prevmi + \\ & diabetes : hdlc + bpmeds : glucose + bpmeds : prevstrk + heartrte : glucose + prevmi : prevhyp + \\ & prevmi : hdlc + prevmi : ldlc + prevstrk : ldlc + prevhyp : ldlc, data = fhs) \quad (1) \end{aligned}$$

The following is our reasoning for selecting the Backward model:

1. The forward selection process tends to miss out on important predictors. In particular, the main effect and interaction effects of `diabp` variable have been overlooked in the Forward model, despite being significant in the Backward and Stepwise models. This suggests that the Forward model is insufficient.
2. The Backward model has the highest adjusted R-squared value. This indicates that the observed data is most well-explained by the Backward model.

3.2 Manual Model Selection

Using our Backward model as a base (full) model, we will now remove interaction effects that we deem unnecessary in terms of the interpretability of the model to obtain a reduced model. The following interaction effects will not be included in the reduced model: `bpmeds:prevstrk`, `sysbp:cigpday`, `sex:prevhyp`, `prevstrk:ldlc`, and `age:prevhyp`.

After running an F-test to determine if the removed covariates are indeed insignificant, our second candidate model is:

$$
\begin{aligned}
M_{red} \leftarrow lm(formula \ = logit\_chdrisk \sim\ & sex + age + sysbp + diabp + cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte + \\
& glucose + prevmi + prevstrk + prevhyp + prevmi : ldlc + hdlc + ldlc + sex : totchol + sex : sysbp + \\
& sex : glucose + sex : prevstrk + prevhyp : ldlc + age : totchol + heartrte : totchol + prevhyp : totchol + \\
& hdlc : totchol + ldlc : totchol + hdlc : ldlc + age : diabp + age : cursmoke + age : heartrte + age : prevmi + \\
& age : hdlc + sysbp : diabp + sysbp : diabetes + sysbp : bpmeds + sysbp : heartrte + sysbp : prevmi + \\
& sysbp : prevhyp + diabp : cursmoke + diabp : cigpday + diabp : bpmeds + prevmi : hdlc + diabp : glucose + \\
& diabp : prevhyp + diabp : hdlc + cursmoke : bmi + cursmoke : hdlc + cigpday : bmi + cigpday : glucose + \\
& cigpday : hdlc + cigpday : ldlc + bmi : bpmeds + bmi : prevmi + bmi : ldlc + diabetes : prevmi + diabetes : hdlc + \\
& bpmeds : glucose + heartrte : glucose + prevmi : prevhyp, \ data = fhs) \quad (2)
\end{aligned}
$$

4. Model Diagnostics

Now that we have our candidate models, we can perform an in-depth comparison by examining certain diagnostics in the following subsections.

4.1 Residual Plots

We will compare our candidate models by analyzing their standardized residual plots:

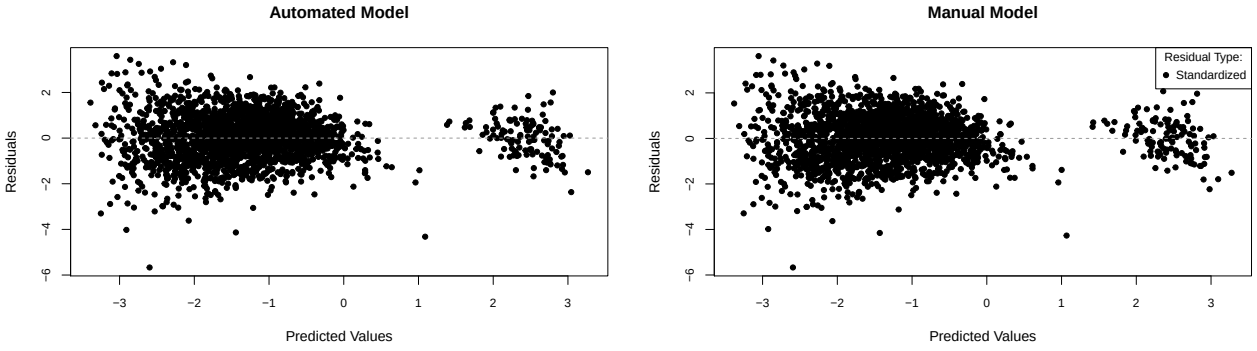


Figure 2: Standard Residuals vs. Fitted Values

In both residual plots (Fig. 2), observe that the predicted values seem to cluster into 2 distinct categories. This is not an issue as the effect of our continuous covariates is small. When this happens, the predictions will be grouped largely according to the discrete covariates.

The second pattern is that the magnitude of the residuals seems to decrease as the predicted values increase. This suggests heteroscedasticity in our linear model. We may have overlooked an important predictor, or that we still require some form of data transformation to spot a clear trend between two variables.

Overall, our residual plots (Fig. 2) for both models look similar. And so, the residual plots will not be useful in determining our final model, unfortunately.

To assess the normality of the models, we opted for histograms as the differences between the two models regarding normality are more pronounced:

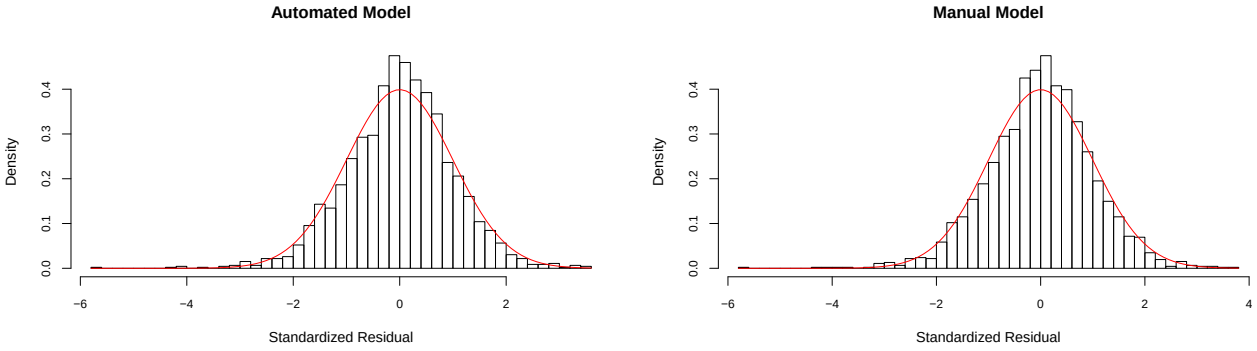


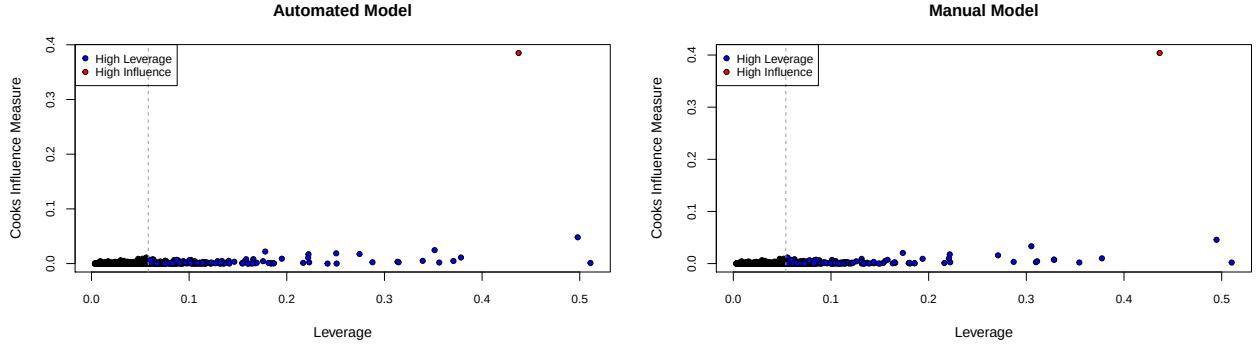
Figure 3: Histogram of Automated (Left) and Manual (Right) Model

- Observe that both histograms (Fig. 3) centre around zero. In spite of this, there are standardized residuals in both models that are way out in the tail of their normal distribution. This suggests that the errors in both model are heavy-tailed. This is also supported by our QQ-plot (Fig. 6).
- At [-2,-0.5] and [0.5,2], the manual model (the graph on the right) appears slightly more normal than the automated model. This suggests that the manual model could potentially be a better fit for our study.

4.2 Leverage and Influence Measures

Next, we wish to investigate high-leverage observations and their influences on the the fit of both candidate models.

(1) Cook's Distance



One of the observations is at least 10 times more than Cook's distance of the others (in red), whereas quite a lot of points have twice the average leverage (in blue). Based on the number of high leverage points, it is wise to assume that our model may have overlooked an important predictor, instead of finding fault with the dataset. However, even with this reasoning, the difference in the Cook's distance between the high leverage point and the rest of the points is still absurdly high. Thus, it is very likely that this observation is incorrectly measured. However, further investigation is still needed in order to fully determine if the high influence point truly affects our dataset.

5. Model Selection

Before we make our final decision on our final model, we will perform a cross-validation analysis to assess the predictive power of the two candidate models.

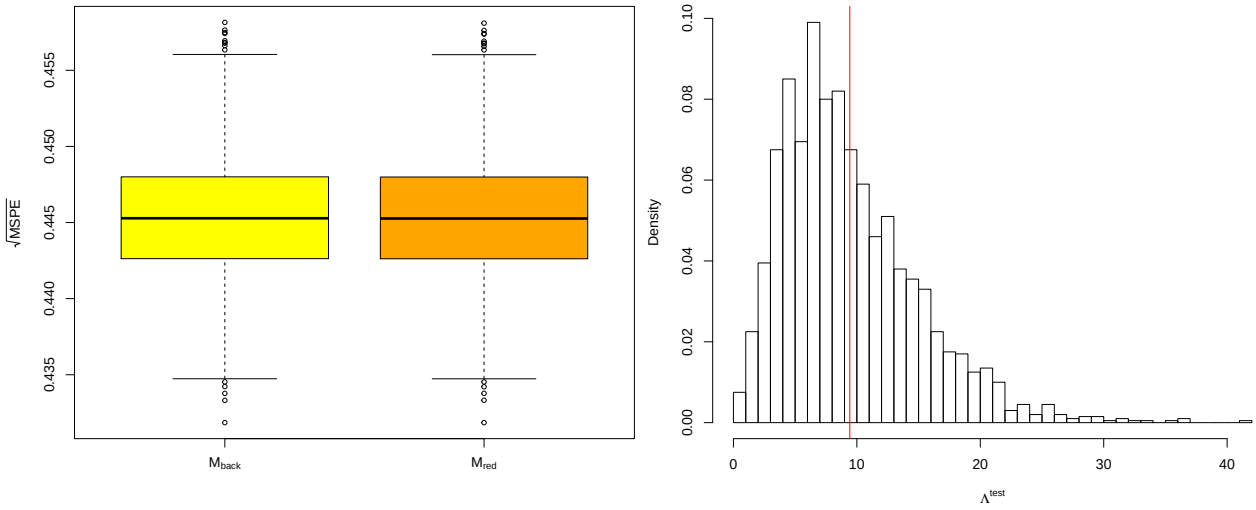


Figure 4: MSPE Boxplot (Left) and logLambda Histogram (Right)

- From the root MSPE figure (Fig. 4), the difference between the automated model (`Mback`) and the manual model(`Mred`) is minimal, and so it is impossible to determine based on this information.
- According to the mean Lambda value (Fig. 4), the cross-validation method prefers the automated model by a very large margin since the mean Lambda value is much larger than 1. In addition to that, there are way more extreme values on the positive x axis, which points to the fact that the automated model is far more accurate.

While the explanatory power for the manual model is slightly stronger than the automated model, the automated model is far more superior in terms of its predictive power, as supported by the mean Lambda value. Thus, **we will be selecting our automated model as the final model**.

The parameter estimates, standard errors, and p-values of the final model are as follows:

	(Intercept)	sexMale	age	sysbp	diabp	cursmokeYes	cigpday	bmi	diabetesYes
Estimate	-8.96	8.42e-01	0.0929	-0.00234	0.00295	0.437	0.0351	-0.0223	0.90500
Standard Error	1.12	1.96e-01	0.0135	0.00475	0.01140	0.409	0.0159	0.0113	0.29600
P-Value	0.00	1.86e-05	0.0000	0.62300	0.79600	0.285	0.0268	0.0489	0.00223

	bpmedsYes	heartrte	glucose	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc
Estimate	0.5660	0.04500	0.00465	6.240	0.870	3.540	-0.01080	0.00321
Standard Error	0.3310	0.00802	0.00269	0.614	0.358	0.405	0.00842	0.00220
P-Value	0.0874	0.00000	0.08420	0.000	0.015	0.000	0.20000	0.14400

	sexFemale:totchol	sexMale:totchol	sexMale:sysbp	sexMale:glucose	sexMale:prevstrkYes	sexMale:prevhypYes
Estimate	-0.00607	-0.00496	-0.00369	-0.002040	-0.3720	0.0872
Standard Error	0.00287	0.00281	0.00124	0.000731	0.1630	0.0562
P-Value	0.03440	0.07710	0.00295	0.005440	0.0225	0.1210

	age:totchol	heartrte:totchol	prevhypYes:totchol	hdlc:totchol	ldlc:totchol	age:diabp	age:cursmokeYes
Estimate	6.85e-05	-6.99e-05	-6.27e-03	3.03e-04	4.60e-06	-0.000484	-0.00609
Standard Error	3.00e-05	1.91e-05	1.21e-03	2.04e-05	2.50e-06	0.000132	0.00306
P-Value	2.25e-02	2.54e-04	2.00e-07	0.00e+00	6.59e-02	0.000246	0.04650

	age:heartrte	age:prevmiYes	age:prevhypYes	age:hdlc	sysbp:diabp	sysbp:cigpday	sysbp:diabetesYes
Estimate	-0.000255	-0.01210	-0.00578	-2.54e-04	3.55e-04	-1.00e-04	-0.00497
Standard Error	0.000101	0.00657	0.00338	9.11e-05	5.28e-05	6.69e-05	0.00172
P-Value	0.011300	0.06540	0.08670	5.44e-03	0.00e+00	1.35e-01	0.00382

	sysbp:bpmedsYes	sysbp:heartrte	sysbp:prevmiYes	sysbp:prevhypYes	diabp:cursmokeYes	diabp:cigpday
Estimate	-0.00549	-1.09e-04	-0.00819	-8.87e-03	-0.01050	0.000464
Standard Error	0.00177	3.84e-05	0.00258	2.05e-03	0.00341	0.000174
P-Value	0.00195	4.45e-03	0.00153	1.64e-05	0.00212	0.007660

	diabp:bpmedsYes	diabp:glucose	diabp:prevhypYes	diabp:hdlc	cursmokeYes:bmi	cursmokeYes:hdlc
Estimate	0.00615	-8.30e-05	-0.00989	-2.02e-04	0.0183	0.00758
Standard Error	0.00327	2.85e-05	0.00375	6.27e-05	0.0102	0.00243
P-Value	0.06000	3.68e-03	0.00845	1.30e-03	0.0725	0.00186

	cigpday:bmi	cigpday:glucose	cigpday:hdlc	cigpday:ldlc	bmi:bpmedsYes	bmi:prevmiYes	bmi:ldlc
Estimate	-0.000859	-6.21e-05	-4.43e-04	-3.84e-05	-0.01200	-0.0236	2.06e-04
Standard Error	0.000420	3.94e-05	1.01e-04	2.14e-05	0.00716	0.0125	5.88e-05
P-Value	0.041100	1.15e-01	1.21e-05	7.28e-02	0.09500	0.0582	4.61e-04

	diabetesYes:prevmiYes	diabetesYes:hdlc	bpmedsYes:glucose	bpmedsYes:prevstrkYes	heartrte:glucose
Estimate	-6.49e-01	0.00576	0.00185	-0.211	5.35e-05
Standard Error	1.36e-01	0.00246	0.00101	0.147	2.62e-05
P-Value	1.90e-06	0.01950	0.06810	0.149	4.16e-02

	prevmiYes:prevhypYes	prevmiYes:hdlc	prevmiYes:ldlc	prevstrkYes:ldlc	prevhypYes:ldlc	hdlc:ldlc
Estimate	-0.37500	0.01180	-0.002800	-0.00236	0.00312	-2.47e-04
Standard Error	0.13300	0.00385	0.000939	0.00149	0.00116	1.89e-05
P-Value	0.00498	0.00224	0.002860	0.11400	0.00720	0.00e+00

Table 4: Parameter Estimates, Standard Errors, and P-Values of the Final Model

6. Discussion

From our final model, we can make the following remarks with regards to the factors associated with CHD risk:

- The most important factors associated with high CHD risk are age, heartrate, and whether the individual has gotten either hypertension or a myocardial infarction. Note that our dataset only consists of middle-aged individuals or older, so this final model may not apply to younger individuals, as seen in the pair plots (Fig. 1).
- Important factors associated with low CHD risk cannot be determined from our model as our model only accounts for siginificant and insignificant factors for high CHD risk, and not the other way around.
- Based on this analysis, I would not be able to recommend behavioral changes to lower the risk of CHD. It is important to understand that correlation is not equivalent to causation. Furthermore, this is an observational study, and not an experimental one. While there may be factors associated with high CHD, we will not know for sure unless we conduct an experimental study to investigate whether these factors truly affect an individual's risk for CHD.
- There are several coefficients with high p-values retained in the final model. In particular, systolic and diastolic blood pressure (`sysbp` and `diabp`) have p-values greater than 0.5. While p-values may be used as a guideline in fitting a linear model on a given dataset, it would be unwise to follow this strictly as p-values depend on covariates that are currently present in the model, and do not reflect the actual significance of a factor. Thus, it would be best to retain coefficients with high p-values.

- There are many outliers in the given dataset, but only one of which might be approriate to remove. As mentioned earlier, it would be unwise to find fault with the dataset when there are many high leverage points, since this typically suggests a flaw in our model itself. However, based on the graph of Cook's distance against Leverage, there is a point that has a Cook's distance at least ten times that of other points'. Even if our model is flawed, the observation should not have been this different from the rest. This suggests that further investigation should take place in regards to that observation as it is entirely possible that that might have been an incorrect observation.
- Our regression assumptions of the final model do not align with the given dataset. As shown earlier in our graph of residuals against fitted values, observe that the residuals seem to decrease as our fitted values (Fig. 2) increase. Furthermore, our data does not align with the Normal distribution at the left tail within the QQ plot. Hence, the final model is deficient in the sense that it is highly likely that a predictor may have been overlooked. It is also entirely possible that our model includes all significant predictors, just that a transformation on our data is needed to better establish the relations. Thus, it is important to remember that any recommendations based on this model are not entirely accurate.

Appendix A: R Code for Descriptive Statistics

To display summary statistics:

```
fhs <- read.csv("fhs.csv") # read csv file

# display summary statistics
data <- summary(fhs[c("chdrisk", "totchol", "age",
                      "cursmoke", "cigpday", "prevstrk", "ldlc")])

# pretty formatting
kable(data, "latex")
```

To display paired plots:

```
# display paired plot
pairs(~ chdrisk + totchol + sysbp + diabp + cigpday + bmi +
        heartrte + glucose + hdlc + ldlc + age,
      pch=16, cex=.7, data=fhs)
```

To compute VIF for the first time:

```
X <- model.matrix(lm(chdrisk ~ totchol + age + sysbp + diabp + cigpday + bmi +
                       heartrte + glucose + hdlc + ldlc - 1, data = fhs))
C <- cor(X) # compute correlation matrix of X
vif <- diag(solve(C)) # take the diagonals of the inverse of C
VIF <- signif(vif, digits=3) # round to 4sf

# pretty formatting
kable(t(data.frame(VIF)), "latex")
```

To compute VIF for the second time (after removing `totchol` since `totchol` had a VIF greater than 10):

```
X2 <- model.matrix(lm(chdrisk ~ age + sysbp + diabp + cigpday + bmi + heartrte +
                        glucose + hdlc + ldlc - 1, data = fhs))
C2 <- cor(X2) # compute correlation matrix of X
vif_rerun <- diag(solve(C2)) # take the diagonals of the inverse of C
VIF_rerun <- signif(vif_rerun, digits=3) #round to 4sf


# pretty formatting
kable(t(data.frame(VIF_rerun)), "latex")
```

Appendix B: Paired Plot After Data Transformation

```
# create new response variable
logit_chdrisk <- log(fhs$chdrisk) - log(1-fhs$chdrisk)

# display paired plot
pairs(~ logit_chdrisk + totchol + sysbp + diabp + cigpday + bmi +
        heartrte + glucose + hdlc + ldlc + age, pch=16, cex=.7, data=fhs)
```

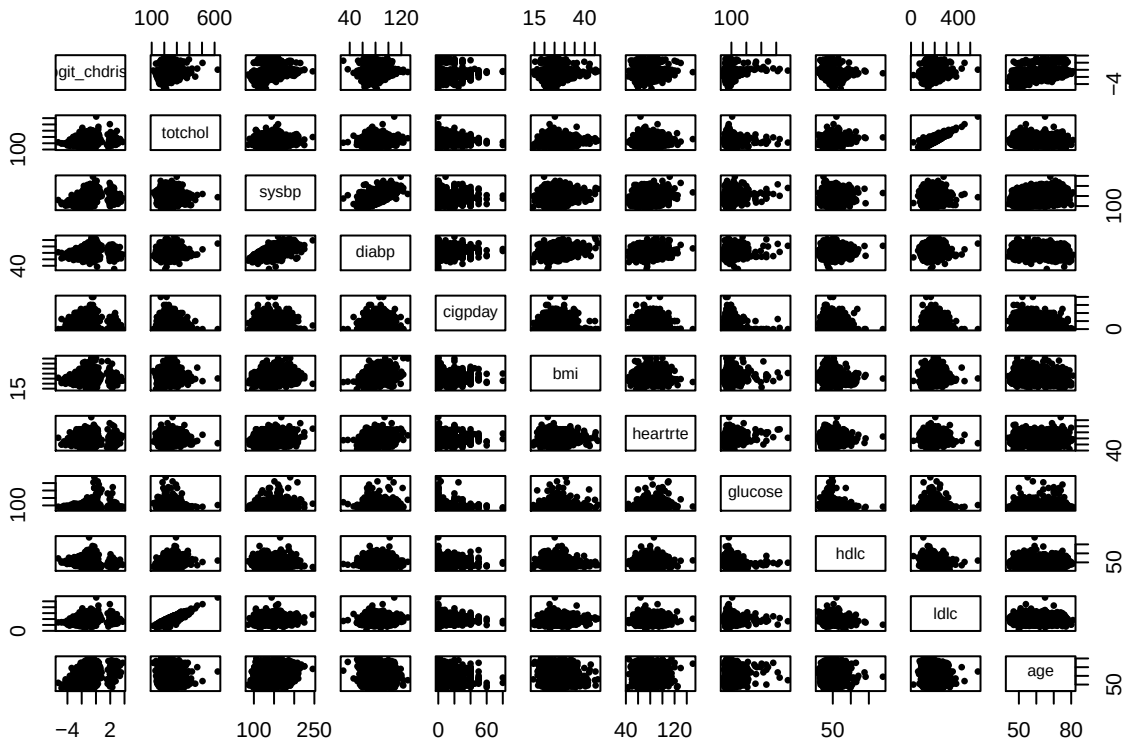


Figure 5: New Paired Plots

Observe that the points in the paired plots (Fig. 5) for `logit_chdrisk` have now been brought closer to one another, allowing us to better analyze certain trends:

As mentioned here, there seems to be a nonlinear relationship between `sysbp` and `logit_chdrisk`. This suggests that we should add an extra variable I(sysbp^2))$ within the linear model.

To fit a line and quadratic of best fit on `logit_chdrisk` against `sysbp`:

```
# initialize x and y for easy access during plotting
x <- fhs$sysbp
y <- logit_chdrisk

# plot graph
par(mai = c(.82, .82, .1, .1))
plot(x, y, xlab = "Systolic Blood Pressure (mmHg)", ylab = "Risk Measure for CHD",
     pch = 16, cex = .7)

# draw line and quadratic of best fits on same graph
```

```r
M1 <- lm(formula=y ~ x, data=fhs) # fit linear model
M2 <- lm(formula=y ~ x + I(x^2), data=fhs) # fit nonlinear model
abline(M1, col = "red") # line of best fit
curve(M2$coef[1] + M2$coef[2]*x + M2$coef[3]*x^2,
      col = "blue", add = TRUE) # quadratic of best fit
legend(x = "bottomright", legend = c("Data", "Regression Line", "Regression Quadratic"),
pch = c(16, NA, NA), pt.cex = .8, cex = .8, col = c("black", "red", "blue"),
lty = c(NA, 1, 1), seg.len = 1) # legend
```

Appendix C: R Code for Automated Model Selection

To check if we have any NA terms for the coefficients of the linear model:

```
# create new response variable
logit_chdrisk <- log(fhs$chdrisk) - log(1-fhs$chdrisk)

# create initial maximal model (includes all main and interaction effects)
Mmax_0 <- lm(logit_chdrisk ~ (. - chdrisk)^2, data=fhs)

# check for NAs in coefficients
beta.max_0 <- coef(Mmax_0)
names(beta.max_0)[is.na(beta.max_0)]
```

To obtain models using automated methods:

(1) Forward Model

```
if (!params$load_calcs_mfwd) {
  Mfwd <- step(object = M0, # starting point model
               scope = list(lower = M0, upper = Mmax), # minimal and maximal models
               direction = "forward",
               trace=FALSE) # trace prints out information
  saveRDS(Mfwd, file="forward_model.rds")
} else {
  tmp <- readRDS("forward_model.rds")
  Mfwd <- tmp
  rm(tmp)
}
Mfwd$call
```

```
## lm(formula = logit_chdrisk ~ prevmi + sysbp + sex + age + ldlc +
##     prevhyp + diabetes + hdlc + cigpday + bmi + glucose + I(sysbp^2) +
##     bpmeds + heartrte + cursmoke + prevstrk + prevmi:sysbp +
##     hdlc:totchol + ldlc:hdlc + prevhyp:totchol + ldlc:prevhyp +
##     prevmi:diabetes + sysbp:prevhyp + sysbp:diabetes + ldlc:bmi +
##     sysbp:age + prevmi:prevhyp + sex:glucose + prevmi:hdlc +
##     bmi:totchol + sysbp:hdlc + age:cigpday + prevmi:ldlc + age:ldlc +
##     heartrte:totchol + age:heartrte + hdlc:cigpday + cigpday:glucose +
##     hdlc:cursmoke + diabetes:hdlc + prevmi:bmi + sysbp:heartrte +
##     sysbp:sex + sysbp:bpmeds + prevmi:prevstrk + ldlc:cursmoke +
##     sex:totchol + prevmi:age + age:glucose + sysbp:cursmoke +
##     prevmi:cigpday + prevmi:glucose, data = fhs)
```

(2) Backward Model

```
if (!params$load_calcs_mback) {
  M0 <- lm(logit_chdrisk ~ 1, data=fhs) # minimal model
  Mmax <- lm(logit_chdrisk ~ (.- chdrisk)^2 + I(sysbp^2) - totchol - # maximal model
             cursmoke:cigpday - bpmeds:prevhyp, data=fhs)

# compute backward model
  Mback <- step(object = Mmax, # starting point model
                scope = list(lower = M0, upper = Mmax),
                direction = "backward",
                trace = FALSE)
```

```
  saveRDS(Mback, file="backward_model.rds")
} else {
  tmp <- readRDS("backward_model.rds")
  Mback <- tmp
  rm(tmp)
}
```

(3) Stepwise Model

```
if (!params$load_calcs_mstep) {
  # Starting model: All main effects without totchol (due to high VIF)
  Mstart <- lm(logit_chdrisk ~ . - chdrisk - totchol, data=fhs)
  Mstep <- step(object = Mstart, # starting point model
                scope = list(lower = M0, upper = Mmax), # minimal and maximal models
                direction = "both",
                trace = FALSE)  # trace prints out information
  saveRDS(Mstep, file="stepwise_model.rds")
} else {
  tmp <- readRDS("stepwise_model.rds")
  Mstep <- tmp
  rm(tmp)
}
Mstep$call
```

```
## lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp + cursmoke +
##     cigpday + bmi + diabetes + bpmeds + heartrte + glucose +
##     prevmi + prevstrk + prevhyp + hdlc + ldlc + sysbp:prevmi +
##     hdlc:totchol + hdlc:ldlc + heartrte:totchol + age:diabp +
##     prevhyp:totchol + diabetes:prevmi + prevhyp:ldlc + sysbp:diabetes +
##     sysbp:prevhyp + sysbp:diabp + bmi:ldlc + sysbp:heartrte +
##     sex:glucose + prevmi:prevhyp + age:cigpday + sysbp:hdlc +
##     prevmi:hdlc + age:ldlc + sex:sysbp + prevmi:ldlc + age:heartrte +
##     sysbp:bpmeds + sysbp:cursmoke + age:glucose + diabp:prevhyp +
##     age:prevhyp + diabetes:hdlc + cigpday:hdlc + cursmoke:hdlc +
##     bmi:prevmi + age:prevmi + cursmoke:ldlc + sex:totchol + cigpday:glucose +
##     prevmi:prevstrk + diabp:bpmeds + bmi:bpmeds + ldlc:totchol +
##     cursmoke:bpmeds + sex:prevhyp, data = fhs)
```

Observe that the Backward model has the most coefficients, followed by the Stepwise model, with the Forward model having the least number of coefficients.

```
# number of coefficients
data <- c(length(coef(Mfwd)), length(coef(Mback)), length(coef(Mstep)))
data <- data.frame(data)
row.names(data) <- c("Forward Model", "Backward Model", "Stepwise Model")
# pretty formatting
kable(data, format='latex', align='c', col.names=c("Number of Coefficients"))
```

	Number of Coefficients
Forward Model	53
Backward Model	67
Stepwise Model	58

Table 5: Number of Coefficients of the Automated Models

Appendix D: R Code for Manual Model Selection

```
# remove Bpemds:prevstrk
anova(Mred_test, Mback)[6]
```

```
##   Pr(>F)
## 1
## 2 0.1492
```

```
# remove sysbp:cigpday
anova(Mred_test, Mback)[6]
```

```
##   Pr(>F)
## 1
## 2 0.1169
```

```
# remove sex:prevhyp
anova(Mred_test, Mback)[6]
```

```
##   Pr(>F)
## 1
## 2  0.089 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# remove prevstrk:ldlc
anova(Mred_test, Mback)[6]
```

```
##    Pr(>F)
## 1
## 2 0.06945 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# FINAL: remove age:prevhyp
anova(Mred_test, Mback)[6]
```

```
##    Pr(>F)
## 1
## 2 0.05225 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe that our final F statistic p-value is still greater than 0.05.

Appendix E: R Code for Model Diagnostics

To plot the Standardized Residual vs Fitted Values graphs for each model:

```
## residuals against predicted values

# automated model
auto_Msum <- summary(Mback) # summary of backward model

# residuals vs predicted values
auto_y.hat <- predict(Mback) # predicted values
auto_sigma.hat <- auto_Msum$sigma
auto_res <- resid(Mback) # original residuals
auto_stan.res <- auto_res/auto_sigma.hat # standardized residuals

# compute leverages
auto_X <- model.matrix(Mback)
auto_H <- auto_X %*% solve(crossprod(auto_X), t(auto_X)) # HAT matrix
auto_h <- hatvalues(Mback) # the R way

auto_p <- length(coef(Mback))
auto_n <- nobs(Mback)
auto_hbar <- auto_p/auto_n # average leverage

# residual plot against predicted values
par(mfrow = c(1,2))
cex <- .8
plot(auto_y.hat, rep(0, length(auto_y.hat)), type = "n", # empty plot for the axis range
     ylim = range(auto_stan.res), cex.axis = cex,
     xlab = "Predicted Values", ylab = "Residuals", main="Automated Model")
# dotted line connecting each observations residuals for better visibility
segments(x0 = auto_y.hat,
         y0 = pmin(auto_stan.res),
         y1 = pmax(auto_stan.res),
         lty = 2)
points(auto_y.hat, auto_stan.res, pch = 21, bg = "black", cex = cex)
abline(h = 0, col = "grey60", lty = 2) # horizontal line


## manual model residual plot
man_Msum <- summary(Mred) # summary of reduced model

# residuals vs predicted values
man_y.hat <- predict(Mred) # predicted values
man_sigma.hat <- man_Msum$sigma
man_res <- resid(Mred) # original residuals
man_stan.res <- man_res/man_sigma.hat # standardized residuals

# compute leverages
man_X <- model.matrix(Mred)
man_H <- man_X %*% solve(crossprod(man_X), t(man_X)) # HAT matrix
man_h <- hatvalues(Mred) # the R way

man_p <- length(coef(Mred))
```

```r
man_n <- nobs(Mred)
man_hbar <- man_p/man_n # average leverage

# residual plot against predicted values
plot(man_y.hat, rep(0, length(man_y.hat)), type = "n", # empty plot to get the axis range
     ylim = range(man_stan.res), cex.axis = cex,
     xlab = "Predicted Values", ylab = "Residuals", main="Manual Model")
# dotted line connecting each observations residuals for better visibility
segments(x0 = man_y.hat,
         y0 = pmin(man_stan.res),
         y1 = pmax(man_stan.res),
         lty = 2)
points(man_y.hat, man_stan.res, pch = 21, bg = "black", cex = cex)
abline(h = 0, col = "grey60", lty = 2) # horizontal line
legend("topright", legend = c("Standardized"),
       pch = 21, pt.bg = c("black"), title = "Residual Type:",
       cex = cex, pt.cex = cex,)
```

The following is our QQ plot for both candidate models:

```r
# qq-plot
par(mfrow = c(1,2))

# automated model
qqnorm(auto_res/auto_sigma.hat, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line

# manual model
qqnorm(man_res/man_sigma.hat, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line
```

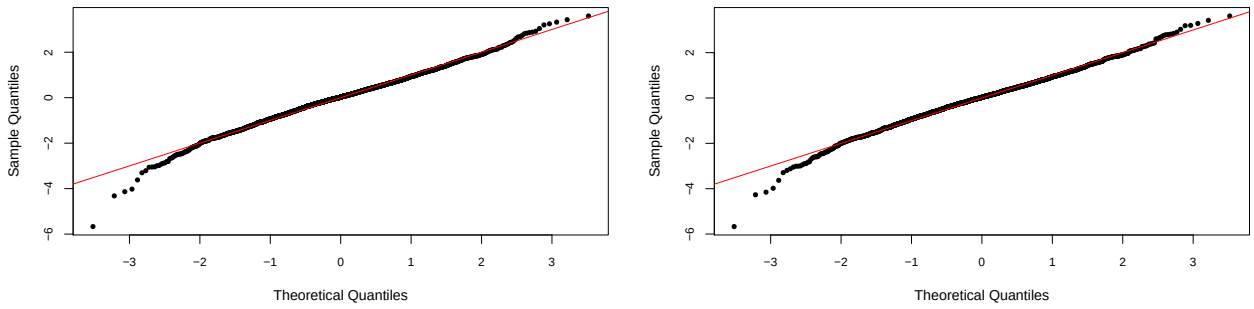


Figure 6: QQ-Plot of Automated (Left) and Manual (Right) Models

Although both tails do not seem to fall on the line at the their left tail, the manual model still has slightly higher normality due to the fact that the right tail of the manual model is closer to the 45 degree line, as opposed to the the automated model.

To display our histograms for a normality check:

```r
#### histogram ####
par(mfrow = c(1,2))
cex <- .8
```

```
## automated model ##
# plot histogram
hist(auto_res/auto_sigma.hat, breaks = 50, freq = FALSE, cex.axis = cex,
xlab = "Standardized Residual", main = "Automated Model")
curve(dnorm(x), col = "red", add = TRUE) # theoretical normal curve

## manual model ##
# plot histogram
hist(man_res/man_sigma.hat, breaks = 50, freq = FALSE, cex.axis = cex,
xlab = "Standardized Residual", main = "Manual Model")
curve(dnorm(x), col = "red", add = TRUE) # theoretical normal curve
```

To display our graphs for Cook's Distance against Leverage:

```
#### cook's distance vs. leverage ####
par(mfrow = c(1,2))

## automated model
auto_D <- cooks.distance(Mback)

# flag some of the points
auto_infl.ind <- which.max(auto_D) # top influence point
auto_lev.ind <- auto_h > 2*auto_hbar # leverage more than 2x the average
auto_clrs <- rep("black", len = auto_n)
auto_clrs[auto_lev.ind] <- "blue"
auto_clrs[auto_infl.ind] <- "red"

# plot graph
plot(auto_h, auto_D, xlab = "Leverage", ylab = "Cooks Influence Measure",
pch = 21, bg = auto_clrs, cex = cex, cex.axis = cex, main="Automated Model")
abline(v = 2*auto_hbar, col = "grey60",lty=2) # 2x average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)

##manual model
man_D <- cooks.distance(Mred)

# flag some of the points
man_infl.ind <- which.max(man_D) # top influence point
man_lev.ind <- man_h > 2*man_hbar # leverage more than 2x the average
man_clrs <- rep("black", len = man_n)
man_clrs[man_lev.ind] <- "blue"
man_clrs[man_infl.ind] <- "red"

# plot graph
plot(man_h, man_D, xlab = "Leverage", ylab = "Cooks Influence Measure",
pch = 21, bg = man_clrs, cex = cex, cex.axis = cex, main="Manual Model")
abline(v = 2*man_hbar, col = "grey60",lty=2) # 2x average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)
```

Appendix F: R Code for Model Selection

Here is our helper function to calculate the mean of the Logit-Normal distribution:

```
# Function to estimate a logit variable with a normal distribution, Y~N(mu, sigma^2)
#   We will use this function to find E[chd_risk | x]
logitnorm_mean <- function(mu, sigma) {
  # Values to be used
  v <- 1/(1+exp(-mu))
  alpha1 <- 1/(sigma^2*(1-v))
  alpha2 <- 1/(v*sigma^2)

  # Finding the nodes and weights for the sum
  gqp <- gauss.quad.prob(n=10, dist="beta", alpha=alpha1, beta=alpha2)
  xi <- gqp$nodes # x_1, ..., x_10
  wi <- gqp$weights # w_1, ..., w_10

  # Transform the nodes into their logit values
  logit_xi <- log(xi) - log(1-xi)

  # Calculating the exponent base in the sum
  g_x <- dnorm(logit_xi, mean=mu, sd=sigma, log=TRUE) - log(1-xi) -
    dbeta(xi, shape1=alpha1, shape2=alpha2, log=TRUE)

  # Estimating E[Y] using the sum of the exponents with corresponding weights
  est <- sum(wi*exp(g_x))
  return(est)
}
```

To check if our helper function works perfectly:

```
# Checking logitnorm_mean function
#   This check was supplied by the project description to ascertain the
#   accuracy of the logitnorm_mean(mu, sigma) function
mu <- c(0.7, 3.2, -1.1)
sigma <- c(.8, .1, 2.3)
sapply(1:3, function(ii) logitnorm_mean(mu[ii], sigma[ii]))
```

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

Below is our code for the boxplots and the Likelihood Ratio:

```
# Models to compare
M1 <- Mback
M2 <- Mred
Mnames <- expression(M[back], M[red]) # Cross-validation setup
nreps <- 2e3 # Number of replications
ntot <- nrow(fhs) # Total number of observations
ntrain <- 500 # Size of training set
ntest <- ntot-ntrain # Size of test set
# Allocating space for mspe, logLambda vectors
mspe1 <- rep(NA, nreps) # Sum-of-square errors for each CV replication
mspe2 <- rep(NA, nreps)
logLambda <- rep(NA, nreps) # Log-likelihod ratio statistic for each replication

for(ii in 1:nreps) {
```

```
    # randomly select training observations
    train.ind <- sample(ntot, ntrain) # training observations

    # refit the models on the subset of training data;
    M1.cv <- update(M1, subset = train.ind)
    M2.cv <- update(M2, subset = train.ind)

    # Finding mu and sigma for M1 and M2
    M1.mu <- mean(predict(M1.cv))
    M2.mu <- mean(predict(M2.cv))
    M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
    M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)

    # Out-of-sample residuals for both models:
    # Testing data - Predictions with training measures
    # Predictions are approximated using logitnorm_mean function
    M1.res <- fhs$chdrisk[-train.ind] - logitnorm_mean(M1.mu, M1.sigma)
    M2.res <- fhs$chdrisk[-train.ind] - logitnorm_mean(M2.mu, M2.sigma)

    # Mean-square prediction errors
    mspe1[ii] <- sqrt(mean(M1.res^2))
    mspe2[ii] <- sqrt(mean(M2.res^2))

    # Out-of-sample likelihood ratio
    logLambda[ii] <- sum(dnorm(M1.res, mean = 0, sd = M1.sigma, log = TRUE))
    logLambda[ii] <- logLambda[ii] -sum(dnorm(M2.res, mean = 0, sd = M2.sigma,
                                              log = TRUE))
}

# Plot rMSPE and out-of-sample log(Lambda)
par(mfrow = c(1,2))
par(mar = c(4.5, 4.5, .1, .1))
boxplot(x = list(sqrt(mspe1), sqrt(mspe2)), names = Mnames, cex = .7,
        ylab = expression(sqrt(MSPE)), col = c("yellow", "orange"))
hist(logLambda, breaks = 50, freq = FALSE,
     xlab = expression(Lambda^{test}),
     main = "", cex = .7)
abline(v = mean(logLambda), col = "red") # Average value
```

Appendix G: R Code for Final Model Display

```r
# Retaining one final model, parameter estimates, standard errors, p-values
man_sum <- summary(Mback)
Mback_est <- c(signif(Mback$coefficients, digits=3))
Mback_se <-c(signif(sqrt(diag(vcov(Mback))), digits=3))
Mback_pval <- c(signif(man_sum$coefficients[,4], digits=3))

summary <- data.frame(Mback_est, Mback_se, Mback_pval)
```