STAT 331 Final Project, Winter 2020

Hanna Nguyen Nathaniel Shek Wing Cheng

1 Summary

With this report, the goal is to try and explore the relation between the risk score for coronary heart disease and some explanatory variables. The data being analysed comes from 2306 individuals that participated in the Framingham Heart Study.

We created two candidate model, one using the automated model selection method and the other using a basic linear model with consideration of interpretability. We concluded that the former model is better, and some important factors affecting the heart disease risk score are the age, gender, number of cigarretes one smokes per day, existence of diabetes, their heart rate and previous special medical conditions.

2 Descriptive Statistics

Summary Statistics

```
##     chdrisk           sex          totchol           age            sysbp
##  Min.   :0.0050   Female:1305   Min.   :112.0   Min.   :44.00   Min.   : 86.0
##  1st Qu.:0.1320   Male  :1001   1st Qu.:207.0   1st Qu.:53.00   1st Qu.:122.5
##  Median :0.2240                 Median :235.5   Median :60.00   Median :136.0
##  Mean   :0.2655                 Mean   :237.8   Mean   :60.23   Mean   :139.2
##  3rd Qu.:0.3448                 3rd Qu.:265.0   3rd Qu.:67.00   3rd Qu.:153.0
##  Max.   :0.9770                 Max.   :625.0   Max.   :81.00   Max.   :246.0
##      diabp        cursmoke      cigpday           bmi        diabetes
##  Min.   : 30.00   No :1504   Min.   : 0.00   Min.   :14.43   No :2142
##  1st Qu.: 73.00   Yes: 802   1st Qu.: 0.00   1st Qu.:23.22   Yes: 164
##  Median : 80.00              Median : 0.00   Median :25.40
##  Mean   : 81.07              Mean   : 6.84   Mean   :25.78
##  3rd Qu.: 88.00              3rd Qu.:10.00   3rd Qu.:27.91
##  Max.   :130.00              Max.   :80.00   Max.   :46.52
##  bpmeds        heartrte         glucose       prevmi     prevstrk   prevhyp
##  No :1973   Min.   : 44.00   Min.   : 46.00   No :2189   No :2260   No : 957
##  Yes: 333   1st Qu.: 70.00   1st Qu.: 75.00   Yes: 117   Yes:  46   Yes:1349
##             Median : 76.00   Median : 83.00
##             Mean   : 77.61   Mean   : 89.07
##             3rd Qu.: 85.00   3rd Qu.: 95.00
##             Max.   :150.00   Max.   :478.00
##       hdlc             ldlc
##  Min.   : 10.00   Min.   : 20.0
##  1st Qu.: 38.00   1st Qu.:152.0
##  Median : 47.00   Median :180.0
##  Mean   : 48.89   Mean   :183.1
##  3rd Qu.: 57.00   3rd Qu.:210.0
##  Max.   :189.00   Max.   :565.0
```

Pair Plots

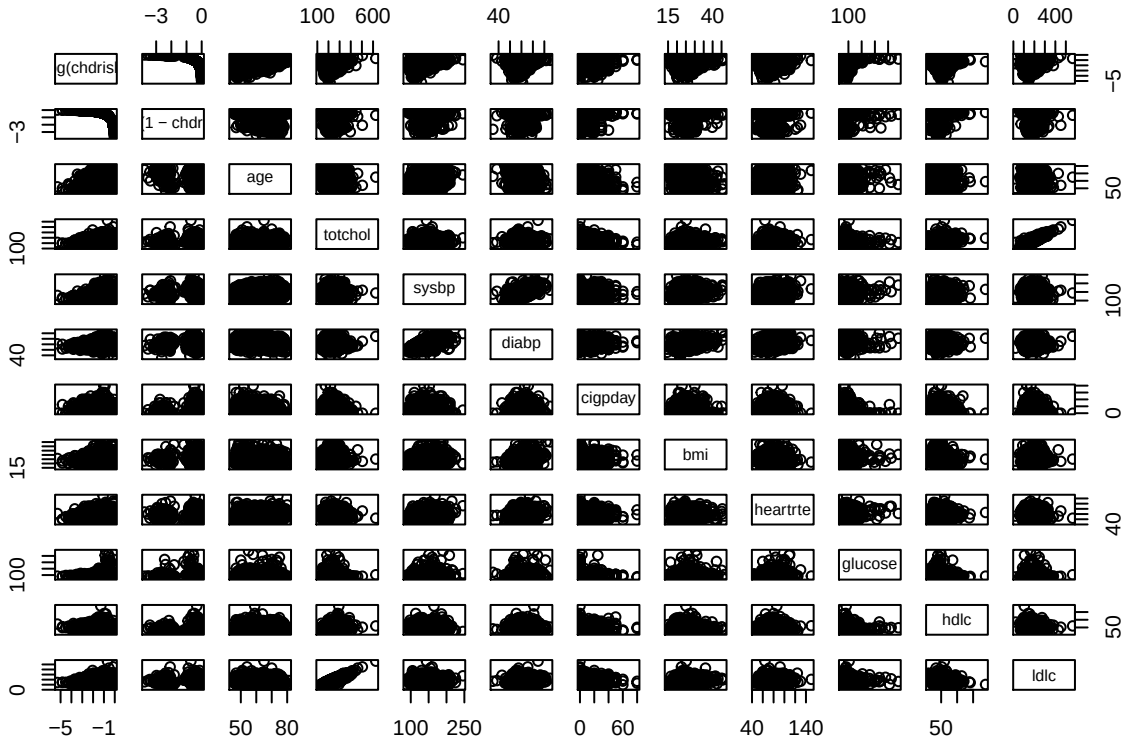

Figure 1: Pair plots for all continuous variables

Variance Inflation Factors (VIF)

```
##   totchol       age     sysbp     diabp   cigpday       bmi  heartrte   glucose
## 10.534949  1.439460  2.456700  2.305485  1.103957  1.151761  1.088789  1.076105
##      hdlc      ldlc
##  2.186297 10.288168
```

Looking at the variance inflation factors, it is clear that serum total cholesterol and low density lipoprotein cholesterol are cause for concern since their VIF results are so high. So looking at the pair plots, specifically for these two covariates, there is a linear relationship between them, which is true in real life. This will be taken into consideration when trying out different models for the dataset.

3 Candidate Models

3.1 First Candidate Model

For the first candidate model, we will consider some automatic model selection techniques. The inputs given will be one model with just the intercept, and one model every explanatory variable and all of their interactions. We will then consider forward selection, backwards selection and stepwise selection and pick one of these three.

After constructing the three models, we will look at their anova calls to decide which candidate model we will choose as our first.

F-statistics: - Forward Selection and Stepwise Selection: 9.91e-12 - Backward Selection and Forward Selection: 0.02735 - Stepwise Selection and Backward Selection: 3.089e-11

Looking at the F-statistics from the anova call, the stepwise selection is the model that will be chosen. And since the benefit from the stepwise to backwards selection is minimal, there is no need to choose the backwards selection over the stepwise selection.

3.2 Second Candidate Model

With this second candidate model, priority will be given to interpritability of the model. To start, we will with a basic model with all the explanatory variables. Then we will compare that model with one that removes serum total cholesterol and low density lipoprotein cholesterol which were the problematic VIFs that was found earlier.

Result of anova call: F-statistics: 2.2e-16

The anova test is used as a goodness of fit diagnostic between the two models. What the code is doing is testing the null hypothesis of setting all the covariates in the basic model that are not in the modified model to zero. And then what we are essentially doing is performing an F-test to identify if M2 is a model that should be chosen. As we can see from the output of anova, the simple modification of the basic linear model did much better and so for the second candidate model, we will choose the modified basic linear model that does not include serum total cholesterol and low density lipoprotein cholesterol. Since the p-value that is given from the test is so low, we will not reject our null hypothesis, and so the modified basic linear model is the one that will do better.

4 Model Diagnostics

Now that we have selected two candidate models, we will proceed to check whether they violates any assumptions of the linear model and provide an in-depth analysis regarding this.

4.1 Comparing the residual plots

We will compare the two plots that consider residuals as a function of predicted values for risk score of coronary heart disease. But first, we should choose the type of residuals whose histogram look the most normal to explore the linearity of each model.

For the first candidate model, we choose the studentized residuals because: - the plots for PRESS and DFFITS residuals do not look as normal as of studentized residuals - the plot for standardized residuals looks normal, if not the same as that of studentized residuals but the units on the y-axis for studentized residuals are not the same as that of the response variable.

When plotting this type of residuals against fitted values for coronary heart disease risk score, we see that the first candidate model produces residuals close to constant-variance with a few outliers. That is the reason why the residual scale seems a bit negatively biased as shown on both the plot and the histogram.

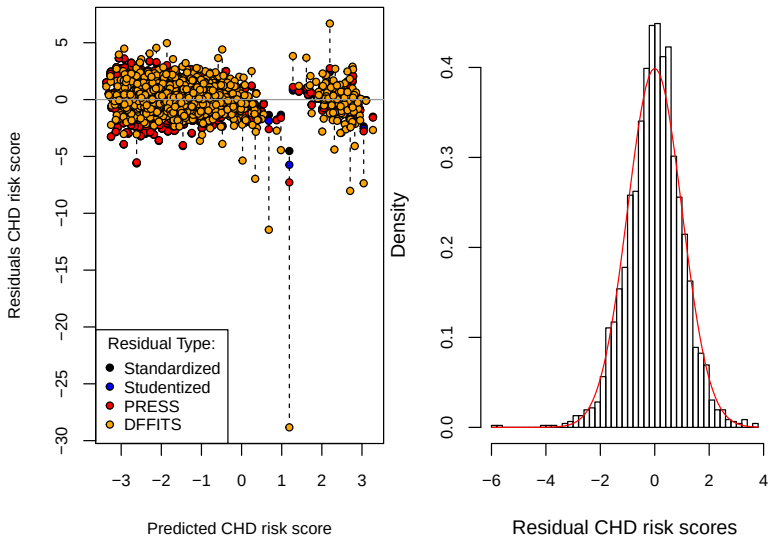

Figure 2: Residual plots for the first candidate model

Similarly, we will also use studentized residuals for the second candidate model.

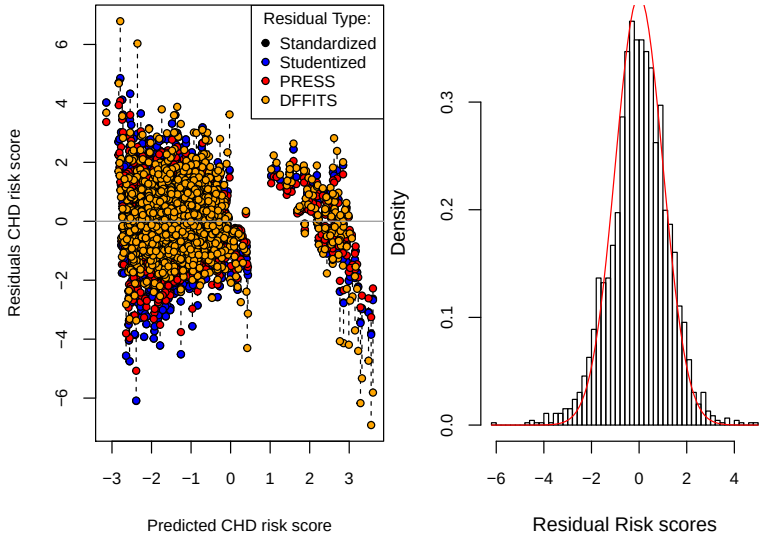

Figure 3: Residual plots for the second candidate model

Overall, the residual distribution as shown by the histogram of the second model is more normal than the first candidate model. However, when looking at the residuals vs fitted values plot, it seems that residuals are not as close to constant-variance, and there are more outliers than the first candidate model. Thus, in terms of homoscedasticity, the first candidate model is preferred.

4.2 Leverage and influence measures

Next, we will analyze any high leverage observations and highly influential measures to investigate the fit of both models on the given data.

We can see that even though there are a lot of high leverage points in both models, we will only pay attention to the points that went pass the cutoff of 0.4 leverage (since points above this cutoff is very widespread in both models) and the high influence points marked in red.

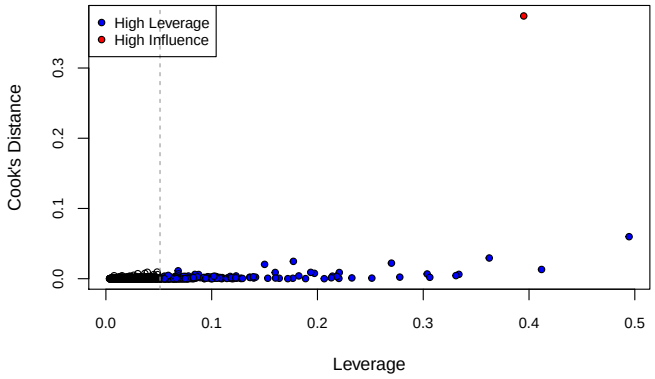

Figure 4: Leverage vs. Cook's Distance of first model

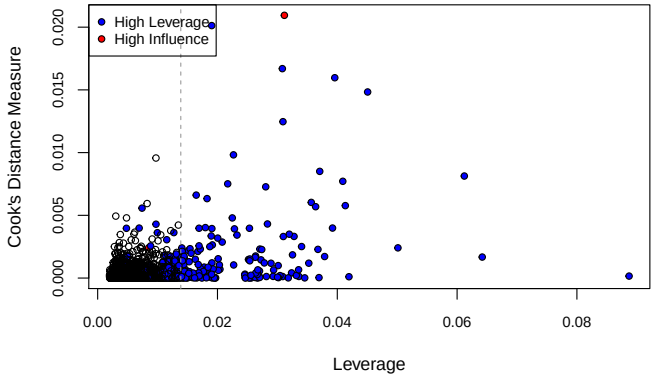

Figure 5: Leverage vs. Cook's Distance of second model

It seems that there are a lot more high leverage points in the second candidate model than the first one that went pass the cutoff. Only two points are above 0.4 in the first model, with the highest leverage being aprroximately 0.5, while the highest leverage in the second model is above 0.8. Thus, the second candidate model is highly influenced by high leverage observations.

Both models have one high influence point that is also a high leverage point. This influential point in the first candidate model went pass the cutoff of 0.4 leverage, while the point in the second candidate model is below this cutoff. Moreover, for the first model, the difference in Cook's distance measure between the high influence point and the point with the second highest Cook's distance is more than 0.25 units of measure, which suggests that this red point is an outlier and might be removed to see a better fit of the model on the data. On the other hand, for the second model, the the difference in Cook's distance measure between the high influence point and the point with the second highest Cook's distance is less than 0.5 units of measure, meaning this point is not too influential on the fit of the model. We will look at these points more in details.

```
##     chdrisk    sex totchol age sysbp diabp cursmoke cigpday   bmi diabetes
## 916   0.276 Female     293  59   165   102       No       0 25.07       No
## 276   0.509 Female     625  65   144   109       No       0 25.16       No
##     bpmeds heartrte glucose prevmi prevstrk prevhyp hdlc ldlc
```

```
## 916     No       86      83     No       No     Yes  189  104
## 276     No       64     105     No       No     Yes   60  565
```

Looking at the point at row 276 for the high leverage point of the first candidate model, we noticed that this point has total serum cholesterol and low density lipoprotein cholesterol values are too high, with both values are the maximum for each column. The total serum cholesterol and low density lipoprotein cholesterol index also have a linear relationship due to the pair plot and their high VIFs, which is a combination of bad signs for risk of heart disease. Inded, the respective coronary hear disease risk score for this data point is fairly high (the value lies in the third quartile according to the summary statistics), which implies this individual is very vulnerable to heart diseases.

```
##      chdrisk    sex totchol age sysbp diabp cursmoke cigpday   bmi diabetes
## 1141   0.851   Male     134  77   175    88       No       0 29.66       No
## 892    0.445 Female     211  55   190   102       No       0 30.58      Yes
##      bpmeds heartrte glucose prevmi prevstrk prevhyp hdlc ldlc
## 1141    Yes       84      97    Yes       Yes     Yes   39   73
## 892     Yes       82     478     No        No     Yes   48  163
```

We can also see that both the high influence point and high leverage point of the second candidate model have very high risk score for heart disease. The high influence point in row 1141 has the total serum cholesterol, systolic blood pressure and diastolic blood pressure in the third quartiles of these indexes, and the values for whether this individual has had a myocardial infarction, a stroke or hypertension are all Yes. The individual is also roughly 85.1% vulnerable to heart disease, which is even twice the risk score of the individual at the high leverage point at 44.5%. This individual also has most indexes in the third quartile, with the highest index of casual serum glucose of 478mg/dL and an inflated low density lipoprotein cholesterol of over 100mg/dl.

The second model detected more outliers than the first model, which means the first model might be more appropriate to fit with the dataset. Lastly, we will perform cross-validation for both models to select the final model to use when studying the factors involved in scoring risk of heart disease.

5 Model Selection

5.1 Cross-Validation

We will perform cross-validation with 2000 replications and 1500 observations for training.

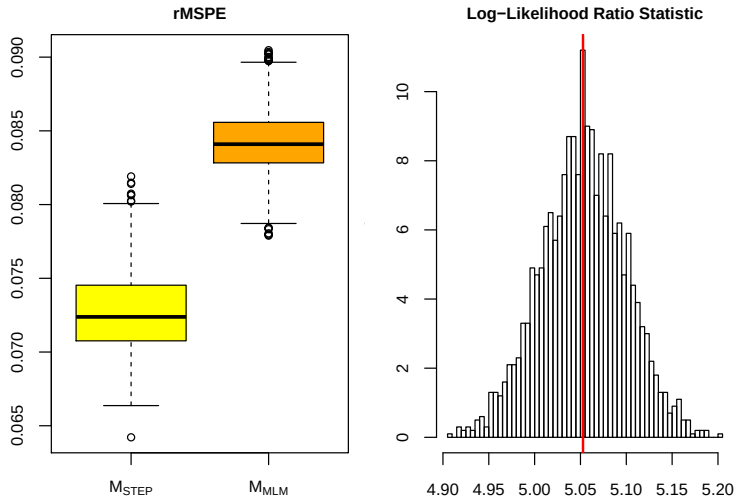

Figure 6: Cross-Validation Boxplot and Histogram

There is a difference between the models. The boxplot and the histogram using shows that the out-of-sample log-likelihood ratio statistics is very positive, leaning towards the first candidate model for fitting given data.

Therefore, we will pick the the former model, which is the Stepwise model selected through Automated Model Selection.

6 Final Model

```
##
## Call:
## lm(formula = log(chdrisk) - log(1 - chdrisk) ~ sex + totchol +
##     age + sysbp + diabp + cursmoke + cigpday + bmi + diabetes +
##     bpmeds + heartrte + glucose + prevmi + prevstrk + prevhyp +
##     hdlc + ldlc + sysbp:prevmi + age:diabp + totchol:prevhyp +
##     totchol:hdlc + hdlc:ldlc + diabetes:prevmi + prevhyp:ldlc +
##     sysbp:diabetes + totchol:heartrte + sysbp:prevhyp + sysbp:diabp +
##     bmi:ldlc + prevmi:prevhyp + sysbp:heartrte + sex:glucose +
##     age:cigpday + prevmi:hdlc + sysbp:hdlc + age:ldlc + sex:sysbp +
##     prevmi:ldlc + age:heartrte + sysbp:bpmeds + sysbp:cursmoke +
##     age:glucose + diabp:prevhyp + age:prevmi + bmi:prevmi + diabetes:hdlc +
##     cigpday:hdlc + cursmoke:hdlc + age:prevhyp + cursmoke:ldlc +
##     sex:totchol + prevmi:prevstrk + cigpday:glucose + diabp:bpmeds +
##     totchol:ldlc + bmi:bpmeds + sex:prevhyp + cursmoke:bpmeds,
##     data = fhs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67803 -0.28338  0.01316  0.28937  1.72607
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -7.716e+00  1.050e+00  -7.347 2.82e-13 ***
## sexMale                8.495e-01  1.951e-01   4.354 1.40e-05 ***
```

```
## totchol               -1.794e-03  2.211e-03  -0.812 0.417114
## age                    7.599e-02  1.255e-02   6.057 1.62e-09 ***
## sysbp                  3.588e-03  5.028e-03   0.714 0.475592
## diabp                 -1.665e-02  1.012e-02  -1.645 0.100018
## cursmokeYes            1.789e-01  2.134e-01   0.838 0.402028
## cigpday                4.030e-02  9.103e-03   4.427 1.00e-05 ***
## bmi                   -2.165e-02  1.109e-02  -1.953 0.050948 .
## diabetesYes            1.155e+00  2.853e-01   4.048 5.33e-05 ***
## bpmedsYes              7.494e-01  3.268e-01   2.293 0.021919 *
## heartrte               4.948e-02  7.831e-03   6.319 3.17e-10 ***
## glucose               -2.281e-03  2.861e-03  -0.797 0.425341
## prevmiYes              6.267e+00  6.166e-01  10.164  < 2e-16 ***
## prevstrkYes            1.829e-01  8.055e-02   2.271 0.023256 *
## prevhypYes             3.649e+00  4.043e-01   9.024  < 2e-16 ***
## hdlc                  -2.759e-02  5.804e-03  -4.753 2.13e-06 ***
## ldlc                  -5.422e-04  2.916e-03  -0.186 0.852479
## sysbp:prevmiYes       -8.676e-03  2.615e-03  -3.318 0.000920 ***
## age:diabp             -4.557e-04  1.295e-04  -3.518 0.000444 ***
## totchol:prevhypYes    -6.171e-03  1.227e-03  -5.031 5.27e-07 ***
## totchol:hdlc           3.009e-04  2.055e-05  14.642  < 2e-16 ***
## hdlc:ldlc             -2.441e-04  1.884e-05 -12.954  < 2e-16 ***
## diabetesYes:prevmiYes -6.662e-01  1.349e-01  -4.937 8.50e-07 ***
## prevhypYes:ldlc        3.005e-03  1.183e-03   2.541 0.011132 *
## sysbp:diabetesYes     -6.780e-03  1.655e-03  -4.097 4.33e-05 ***
## totchol:heartrte      -6.721e-05  1.906e-05  -3.527 0.000429 ***
## sysbp:prevhypYes      -8.953e-03  2.056e-03  -4.354 1.40e-05 ***
## sysbp:diabp            3.543e-04  5.274e-05   6.719 2.32e-11 ***
## bmi:ldlc               2.043e-04  5.910e-05   3.458 0.000555 ***
## prevmiYes:prevhypYes  -3.686e-01  1.336e-01  -2.759 0.005852 **
## sysbp:heartrte        -1.140e-04  3.804e-05  -2.996 0.002765 **
## sexMale:glucose       -2.060e-03  7.073e-04  -2.912 0.003623 **
## age:cigpday           -2.890e-04  1.294e-04  -2.234 0.025609 *
## prevmiYes:hdlc         1.117e-02  3.819e-03   2.926 0.003470 **
## sysbp:hdlc            -1.075e-04  3.442e-05  -3.123 0.001816 **
## age:ldlc               6.035e-05  2.883e-05   2.093 0.036460 *
## sexMale:sysbp         -3.680e-03  1.246e-03  -2.955 0.003164 **
## prevmiYes:ldlc        -2.863e-03  9.315e-04  -3.074 0.002139 **
## age:heartrte          -2.506e-04  1.011e-04  -2.478 0.013279 *
## sysbp:bpmedsYes       -5.752e-03  1.777e-03  -3.237 0.001227 **
## sysbp:cursmokeYes     -1.553e-03  1.099e-03  -1.413 0.157730
## age:glucose            8.069e-05  4.427e-05   1.822 0.068523 .
## diabp:prevhypYes      -1.030e-02  3.752e-03  -2.746 0.006077 **
## age:prevmiYes         -1.162e-02  6.546e-03  -1.775 0.076019 .
## bmi:prevmiYes         -2.119e-02  1.243e-02  -1.705 0.088319 .
## diabetesYes:hdlc       5.755e-03  2.447e-03   2.351 0.018790 *
## cigpday:hdlc          -3.008e-04  9.572e-05  -3.142 0.001699 **
## cursmokeYes:hdlc       5.344e-03  2.370e-03   2.254 0.024268 *
## age:prevhypYes        -6.835e-03  3.386e-03  -2.018 0.043663 *
## cursmokeYes:ldlc      -7.778e-04  4.990e-04  -1.559 0.119175
## sexMale:totchol        1.065e-03  5.156e-04   2.066 0.038934 *
## prevmiYes:prevstrkYes -3.239e-01  1.986e-01  -1.631 0.102944
## cigpday:glucose       -6.675e-05  3.793e-05  -1.760 0.078604 .
## diabp:bpmedsYes        7.124e-03  3.320e-03   2.146 0.031988 *
## totchol:ldlc           4.298e-06  2.516e-06   1.709 0.087671 .
```

```
## bmi:bpmedsYes          -1.334e-02  7.177e-03  -1.858 0.063265 .
## sexMale:prevhypYes      8.497e-02  5.633e-02   1.509 0.131553
## cursmokeYes:bpmedsYes -1.049e-01  6.975e-02  -1.504 0.132625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4764 on 2247 degrees of freedom
## Multiple R-squared:  0.8447, Adjusted R-squared:  0.8407
## F-statistic: 210.7 on 58 and 2247 DF,  p-value: < 2.2e-16
```

7 Discussions

1. Thus, through the model diagnostics, we recognized that high level of total serum cholesterol, low density lipoprotein cholesterol, blood pressure, and glucose together with previous heart-related diseases are associated with high risk score for heart disease. Individuals with low risk of getting heart diseases usually have a high level of high density lipoproterin cholesterol, with few to no history of having any special heart-related medical conditions.

2. Some behavioral changes to lower the risk for heart disease are: cutting back on fat and sugar consumption, increase exercises and have more meals with high density lipoprotein like beef

3. There are some covariates with high p-values are still retained in the final model, namely the total serum cholesterol, the systolic and diastolic blood pressure, low density lipoprotein cholesterol and some categorical covariates. These factors are important in determining the heart disease risk score, however we can only see this relationship clearly through outliers and thus are not deemed significant in the model. The model also has too many explanatory variables to begin with, so after automating the model selection process, that's why the final model still has so many covariates with high p-values.

4. Some outlying observations might be removed from the dataset as considered in the model diagnostics are the high influence point in row 276, 892 and 1141. These observations are too extreme and might affect the study of the correlation between heart disease risk and their explanatory factors. Nonetheless, it helps us in understanding many factors, especially high p-value factors, contribute the different levels of CHD risk score.

5. The final model might be overfitting the data, since the sample size for training during the cross-validation step is too high compared to the total number of replications it used. Even though the second model might be preferred in terms of predictive power, the cross-validation process still favor the first model more. The dataset contains too many covariates to begin with, so it is natural that the automated model detects many relations between explanatory variables.

8 Appendix

8.1 R Code used to generate plots

8.1.1 R code for figure 1

```r
## Pair plots
pairs(~ (log(chdrisk) - log(1-chdrisk)) + age + totchol + sysbp + diabp +
        cigpday + bmi + heartrte + glucose + hdlc + ldlc, data = fhs)
```

8.1.2 R code for figure 2

```
## Residual plots for first candidate model
par(mfrow = c(1,2), mar = c(4,4,.1,.1)) # plot frame
pch = 21 # plot character
cex = .8 # size of data point
# plot the residuals vs. fitted value for first model
plot(x = 0, type = "n", # empty plot to get the axis range
     xlim = range(y.hat),
     ylim = range(Resid), cex.lab = cex, cex.axis = cex,
     xlab = 'Predicted CHD risk score', ylab = 'Residuals CHD risk score')
# add dotted lines between residuals to enhance visibility
res.y0 <- apply(Resid, 1, min)
res.y1 <- apply(Resid, 1, max)
segments(x0 = y.hat, y0 = res.y0, y1 = res.y1, lty = 2)
# add points
for(ii in 1:4) {
  points(y.hat, Resid[,ii], pch = pch, cex = cex, bg = clrs[ii])
}
abline(h = 0, col = "grey60") # zero residual line
legend("bottomleft",
       legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.cex = cex, cex = cex,
       pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:")

# histogram for studentized residuals
hist(stud.res, breaks = 50, freq = FALSE, cex.axis = cex,
     xlab = "Studentized Residual CHD risk score", main = "")
curve(dnorm(x), col = "red", add = TRUE) # fit the normal curve
```

8.1.3 R code for figure 3

```
## Residual plots for second candidate model
par(mfrow = c(1,2), mar = c(4,4,.1,.1)) # plot frame
pch = 21 # plot character
cex = .8 # size of data point
# plot the residual vs. fitted value for first model
plot(x = 0, type = "n", # empty plot to get the axis range
     xlim = range(y.hat1),
     ylim = range(Resid1), cex.lab = cex, cex.axis = cex,
     xlab = 'Predicted CHD risk score', ylab = 'Residuals CHD risk score')
# add dotted lines between residuals to enhance visibility
res.y01 <- apply(Resid1, 1, min)
res.y11 <- apply(Resid1, 1, max)
segments(x0 = y.hat1, y0 = res.y01, y1 = res.y11, lty = 2)
# add points
for(ii in 1:4) {
  points(y.hat1, Resid1[,ii], pch = pch, cex = cex, bg = clrs[ii])
}
abline(h = 0, col = "grey60") # zero residual line
```

```
legend("bottomleft",
       legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.cex = cex, cex = cex,
       pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:")

# histogram for studentized residuals
hist(stud.res1, breaks = 50, freq = FALSE, cex.axis = cex,
     xlab = "Studentized Residual Risk scores", main = "")
curve(dnorm(x), col = "red", add = TRUE) # fit the normal curve
```

8.1.4 R code for figure 4

```
## Leverage vs. Influence plot
# Cook's Distance vs. Leverage first model
clrs[lev.ind] <- 'blue' # color for high leverage index
clrs[infl.ind] <- 'red' # color for high influence index
# plot Cook's Distance measure against leverage
plot(h, D, xlab = 'Leverage', ylab = "Cook's Distance",
     pch = pch, bg = clrs, cex = cex, cex.axis = cex)
# limit for high leverage points
abline(v = 2*hbar, col = 'grey60', lty=2)
# specifying high influence and leverage points
legend("topleft", legend = c("High Leverage", "High Influence"),
       pch = pch, pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)
```

8.1.5 R code for figure 5

```
# Cook's Distance vs. Leverage second model
clrs[lev.ind1] <- 'blue' # color for high leverage index
clrs[infl.ind1] <- 'red' # color for high influence index
# plot Cook's Distance measure against leverage
plot(h1, D1, xlab = 'Leverage', ylab = "Cook's Distance Measure",
     pch = pch, bg = clrs, cex = cex, cex.axis = cex)
# limit for high leverage points
abline(v = 2*hbar1, col = 'grey60', lty=2)
# specifying high influence and leverage points
legend("topleft", legend = c("High Leverage", "High Influence"),
       pch = pch, pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)
```

8.1.6 R code for figure 6

```
## Boxplot and histogram of out-of-sample log likelihood ratio statistics
par(mfrow = c(1,2), mar = c(5.1, 5.1, 3, 1.1)*0.6) # plot frame
cex <- .8 # size of data points

# boxplot
```

```r
boxplot(x = list(sqrt(mspe1), sqrt(mspe2)), names = Mnames,
        main = "rMSPE",
        ylab = "rMSPE",
        col = c("yellow", "orange"),
        cex = cex, cex.lab = cex, cex.axis = cex, cex.main = cex)

# out-of-sample log likelihood ratio statistics
lambda <- lambda1 - lambda2
# histogram of out-of-sample log likelihood ratio statistics
hist(log(lambda), breaks = 50, freq = FALSE,
     main = "Log-Likelihood Ratio Statistic",
     ylab = "Density",
     xlab = "log-likelihood ratio test",
     cex = cex, cex.lab = cex, cex.axis = cex, cex.main = cex)
abline(v = mean(log(lambda)), col = "red", lwd = 2) # mean of lambda
```

8.2 R Code used to generate statistics

```r
## Summary Statistics
summary(fhs)


## Variance inflation factor
X <- model.matrix(lm(log(chdrisk) - log(1 - chdrisk)~ . - 1 - sex -
                       cursmoke - diabetes - bpmeds - prevmi - prevstrk - prevhyp,
                     data = fhs)) # model matrix without catergorial data
C <- cor(X) # correlation matrix
vif <- diag(solve(C)) # diagonal lines of the model matrix
vif


## First Candidate Model
# intercept only
M0 <- lm(log(chdrisk) - log(1 - chdrisk) ~ 1, data = fhs)

# all main effects and interactions
Mmax <- lm(log(chdrisk) - log(1 - chdrisk) ~ (.)^2, data = fhs)

beta.max <- coef(Mmax)
length(beta.max) # number of coefficients
names(beta.max)[is.na(beta.max)] # coefficients that couldn't be estimated
table(fhs[c("cursmoke", "cigpday")])

Mmax <- lm(log(chdrisk) - log(1 - chdrisk) ~ (.)^2 - cursmoke:cigpday -
             bpmeds:prevhyp, data = fhs)
anyNA(coef(Mmax)) # make sure there are no coefficients with NA in the model

# starting point model: main effects only
Mstart <- lm(log(chdrisk) - log(1 - chdrisk) ~ ., data = fhs)

# forward
```

```r
system.time({
  Mfwd <- step(object = M0, # base model
               scope = list(lower = M0, upper = Mmax), # smallest and largest model
               direction = "forward",
               trace = FALSE) # trace prints out information
})

# backward
system.time({
  Mback <- step(object = Mmax, # base model
                scope = list(lower = M0, upper = Mmax),
                direction = "backward", trace = FALSE)
})

# stepwise
system.time({
  Mstep <- step(object = Mstart,
                scope = list(lower = M0, upper = Mmax),
                direction = "both", trace = FALSE)
})

# F-statistics for first model
anova(Mfwd, Mstep)
anova(Mstep, Mback)
anova(Mback, Mfwd)


## Second Candidate Model
# basic linear model
M1 <- lm(log(chdrisk) - log(1 - chdrisk) ~ ., data = fhs) #basic linear model
# modified linear model
M2 <- lm(log(chdrisk) - log(1 - chdrisk) ~ . - totchol - ldlc, data = fhs)

# F-statistics for second model
anova(M2, M1)


## Residuals for first candidate model
# predicted values
y.hat <- predict(Mstep)
sigma.hat <- sigma(Mstep)

# ordinary residuals
res <- resid(Mstep)

# standardized residuals
stan.res <- res/sigma.hat

# computing leverages for finding the studentized residuals
X <- model.matrix(Mstep)
H <- X%*%solve(crossprod(X), t(X)) # HAT matrix
h <- hatvalues(Mstep)
```

```r
# studentized residuals
stud.res <- stan.res/sqrt(1-h)

# PRESS residuals
press.res <- res/(1-h)

# DFFITS residuals
dfts.res <- dffits(Mstep)

# collect all residuals
Resid <- data.frame(stan = stan.res,
                    stud = stud.res,
                    press = press.res,
                    dffits = dfts.res)

hbar  <- ncol(model.matrix(Mstep))/nobs(Mstep) #
# standardize residuals by making them all equal at average leverage
Resid <- within(Resid, {
  stud <- stud.res * sqrt(1-hbar)
  press <- press.res * (1-hbar)/sigma.hat
  dffits <- dfts.res * (1-hbar)/sqrt(hbar)
})


## Residualss for second candidate model
# predicted values
y.hat1 <- predict(M2)
sigma.hat1 <- sigma(M2)

# ordinary residuals
res1 <- resid(M2)

# standardized residuals
stan.res1 <- res1/sigma.hat

# computing leverages for finding the studentized residuals
X1 <- model.matrix(M2)
H1 <- X1%*%solve(crossprod(X1), t(X1)) # HAT matrix
h1 <- hatvalues(M2)

# studentized residuals
stud.res1 <- stan.res1/sqrt(1-h1)

# PRESS residuals
press.res1 <- res1/(1-h1)

# DFFITS residuals
dfts.res1 <- dffits(M2)

# collect all residuals
Resid1 <- data.frame(stan = stan.res1,
                     stud = stud.res1,
                     press = press.res1,
```

```
                    dffits = dfts.res1)

hbar1  <- ncol(model.matrix(M2))/nobs(M2) # average leverage
# standardize residuals by making them all equal at average leverage
Resid1 <- within(Resid1, {
  stud <- stud.res1 * sqrt(1-hbar1)
  press <- press.res1 * (1-hbar1)/sigma.hat1
  dffits <- dfts.res1 * (1-hbar1)/sqrt(hbar1)
})


## Leverage and Influence measures for first model
clrs <- c("black", "blue", "red", "orange") # plot frame
pch = 21 # plot character
cex = .8 # size of data points
# Cook's distance for the first model
D <- cooks.distance(Mstep) # calculate Cook's Distance Measures

infl.ind <- which.max(D) # top influence point
lev.ind <- h > 2*hbar # leverage more than twice average


## Leverage and Influence measures for second model

# Cook's distance for the second model
D1 <- cooks.distance(M2) # calculate Cook's Distance Measures

infl.ind1 <- which.max(D1) # top influence point
lev.ind1 <- h1 > 2*hbar1 # leverage more than twice average



## Outliers
# Outliers from the first model
omit.ind <- c(infl.ind, # most influential
              which.max(h)) # highest leverage
# naming the row of most influential and highest leverage points
names(omit.ind) <- c("infl", "lev")
# details of most influential and highest leverage points
fhs[omit.ind,]

# Outliers from the second model
omit.ind1 <- c(infl.ind1, # most influential
               which.max(h1)) # highest leverage
# naming the row of most influential and highest leverage points
names(omit.ind1) <- c("infl", "lev")
# details of most influential and highest leverage points
fhs[omit.ind1,]

## Cross-Validation Process
require(statmod)

# logitnorm_mean function
```

```r
logitnorm_mean <- function(mu, sigma){
  v <- 1/(1+exp(-mu))
  alpha1 <- 1/((sigma^2)*(1-v))
  alpha2 <- 1/(v*(sigma^2))
  gqp <- gauss.quad.prob(n=10,dist="beta",alpha=alpha1,beta=alpha2)
  x <- gqp$nodes
  y <- gqp$weights
  g <- dnorm((log(x/(1-x))),mean=mu,sd=sigma,log = TRUE) - log(1-x) -
    dbeta(x, shape1=alpha1,shape2 = alpha2, log = TRUE)
  sum(y*(exp(g)))
}


# compare Mstep to M2
M10 <- Mstep
M20 <- M2
Mnames <- expression(M[STEP], M[MLM])

# number of cross-validation replications
nreps <- 2e3 # number of replications
ntot <- nrow(fhs) # total number of observations
ntrain <- 1500 # for fitting MLE's
ntest <- ntot-ntrain # for out-of-sample prediction

# storage space
mspe1 <- rep(NA, nreps) # mean-squared prediction errorsfor M10
mspe2 <- rep(NA, nreps) # mean-squared prediction errors for M20
lambda1 <- rep(NA, nreps) # out-of-sample log-likelihood for M1
lambda2 <- rep(NA, nreps) # out-of-sample log-likelihood for M2

# cross-validation
system.time({
  for(ii in 1:nreps) {
    if(ii%%100 == 0) message("ii = ", ii)
    train.ind <- sample(ntot, ntrain) # training observations

    # fit the models on the subset of training data
    M10.cv <- update(M10, subset = train.ind)
    M20.cv <- update(M20, subset = train.ind)

    # out-of-sample log-likelihoods
    M10.sigma <- sqrt(sum(resid(M10.cv)^2)/ntrain) # MLE of sigma
    M20.sigma <- sqrt(sum(resid(M20.cv)^2)/ntrain)

    mu_1 <- predict(M10.cv, newdata = fhs[-train.ind,])
    mu_2 <- predict(M20.cv, newdata = fhs[-train.ind,])

    # out-of-sample residuals
    M10.res <- fhs$chdrisk[-train.ind] -
      sapply(1:ntest, function(ii) logitnorm_mean(mu_1[ii],M10.sigma))
    M20.res <- fhs$chdrisk[-train.ind] -
      sapply(1:ntest, function(ii) logitnorm_mean(mu_2[ii],M20.sigma))
```

```
    # mean-square prediction errors for each model
    mspe1[ii] <- mean(M10.res^2)
    mspe2[ii] <- mean(M20.res^2)

    # since res = y - pred, dnorm(y, pred, sd) = dnorm(res, 0, sd)
    lambda1[ii] <- sum(dnorm(M10.res, sd = M10.sigma, log = TRUE))
    lambda2[ii] <- sum(dnorm(M20.res, sd = M20.sigma, log = TRUE))
  }
})



## Final model
summary(Mstep)
```