Framingham Heart Study

Yuting Tang

September 07, 2020

1 Summary

This project investigates the relation between the quantified risk of coronary heart disease (CHD) with some explanatory factors based on the data collected from 2306 individuals participating in the Framingham Heart Study. Two candidate models are developed by automated model selection and manual construction with exploration of colinear predictors, significance of covariates, and outstanding interaction effects. After the diagnostics regarding assumption violation, outlier, leverage, influential observation, and explanatory and predictive power, the automatedly selected model is chosen to be the final model. The final model consists of 56 predictors with significant ones such as myocardial infarction history (covariate `prevmi`) and hypertension (covariate `prevhyp`). Additionally, some interaction effects between continuous and categorical variables (e.g., interaction between high density lipoprotein cholesterol and presence of myocardial infarction history) are vitally significant to predict the risk of CHD.

2 Descriptive Statistics

This section provides basic information, together with exploratory diagnostics of variables in the dataset before feeding raw data into models. The diagostics are performed based on the summary statistics, pair plots, and Variance Inflation Factor (VIF) of explanatory variables.

2.1 Exploratory Diagnostics

The summary statistics provides a straightforward view of the range for continuous variables and the totals per category for categorical ones. Based on the range and categorical totals, basic analysis on whether to transform, exclude, or make additional assumptions on certain covariates can be performed. In this Framingham Heart Study, only 46 out of 2306 individuals has had a stroke before the survey was conducted. Thus, it might incur inaccurate statements regarding whether the individual has had stroke or not would impact the risk for Coronary Heart Disease (CHD). To mitigate the inaccuracy, it would be clearly stated that individuals who has had a stroke before are based on a very small sample size, so more data is required to properly support the potential relationship between the history of stroke and the risk of CHD.

```r
# summary statistics
summary(fhs)
```

```
##     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
```

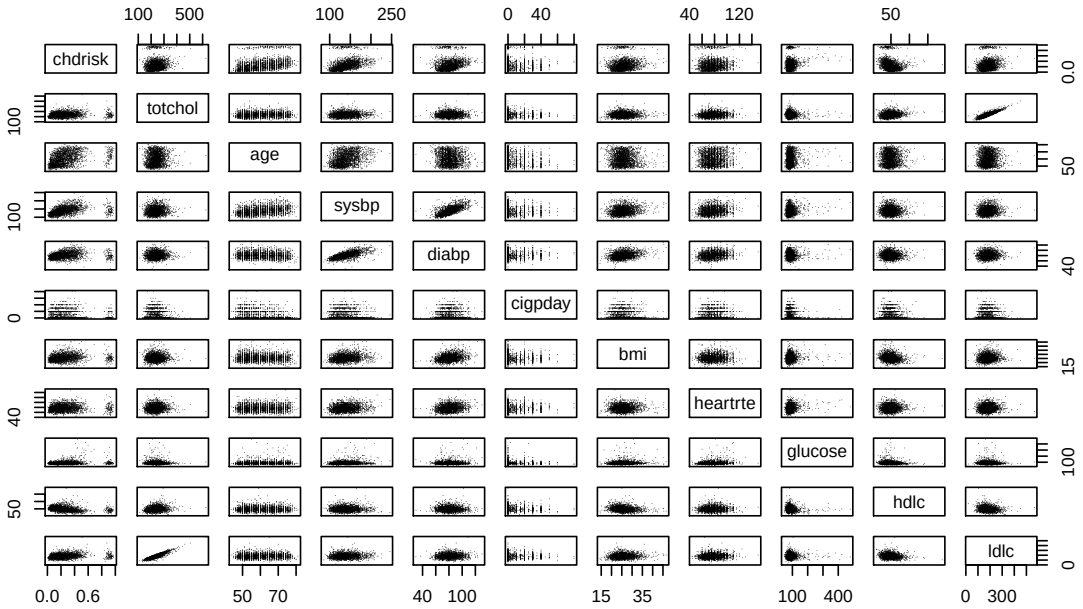

Figure 1: Pair plots for all continuous variables in the dataset.

Another preliminary diagnostics of the data is the scatter plot of each pair of continuous variables, which displays whether there exits linear relationship between the pair of variables in each plot. Categorical variables are excluded from the pair plot, since there is no linear relationship between any continuous and categorical covariates, nor between any categorical ones. The following information is revealed from the pair plots in Figure 1:

- There seems to be moderate association between `chdrisk` and any of the continuous predictors, except comparatively strong relationship between `chdrisk` and `sysbp`. This suggests that individuals with higher systolic blood pressure are exposed to higher risk for Coronary Heart Disease (CHD).
- Clearly, there is strong linear correlation between `totchol` and `ldlc`, indicating the high colinearity between these 2 predictors. It seems plausible as people with more serum total cholesterol usually have higher level of low density lipoprotein cholesterol.
- Similarly, `sysbp` and `diabp` has stronger correlation compared with other predictors, since their linear relationship is moderately strong. It is reasonable given that the majority of individuals with high systolic blood pressure also have high diastolic blood pressure.
- While `cigpday` is a continuous predictor, it only takes on 31 categories. The presence of vertical line with jitter in the `cigpday` variable typically means the study participants provided ordinal answers relating to number of cigarretes smoked each day, which is then encoded numerically as above and then averaged together.

2.2 Colinear Predictors

In addition to diagnosing colinearity by subjective assessment from Figure 1, Variance Inflaction Factor (VIF) is adopted to objectively identify the colinear predictors. As a rule of thumb, VIF's greater than 10 are causes for concern. In this study, the variance inflation factors for `totchol` and `ldlc` are 10.5 and 10.3 respectively, which indicates that `totchol` and `ldlc` are highly correlated. This agrees with the observation from Figure 1. The colinearity implies that `totchol` and `ldlc` can be written as linear combination of other predictors in the model. The observation is reasonable, as the combination of age, systolic blood pressure, diastolic blood pressure, myocardial infraction history, hypertension condition, body mass index, and high density lipoprotein cholesterol can well predict the level of low density lipoprotein cholesterol and serum total cholesterol.

When covariates are highly correlated, the regression has trouble figuring out whether the change in the response variable is due to one predictor or the other. Thus, covariates `totchol` and `ldlc` should be removed from the regression in the following model fitting section to improve the accuracy of prediction.

3 Candidate Models

This section constructs two candidate models in two different ways. The first candidate model is obtained by automated model selection which includes forward selection, backward elimination, and stepwise selection. Models generated by these methods are compared from the perspectives of running time, number of parameters, and adjusted coefficient of determination. The second candidate model is obtained by manual construction which involves F-test assessment on significance of predictors and interaction effects. Both candidate models start with the identified colinear predictors `totchol` and `ldlc` being removed.

3.1 Automated Model Selection

3.1.1 Forward Selection

The input of forward selection is the minimal model with only the intercept and no covariates. Forward selection keeps adding the most significant covariates that are significant to the response variable but

not in the current regression, until no new variables are found to be significant in the presence of the current set.

```
# start model of forward method
m_empty <- lm(logit ~ 1, data = fhs_after_vif)
```

3.1.2 Backward Elimination

Backward Elimination starts with the maximal model that includes all the main effects (i.e., linear covariates) and interaction effects (e.g., `age*sex`). However, this maximal model has 121 coefficients, 2 of which are `NA`:

```
# Coefficients with NA value
names(coef(m_all)[is.na(coef(m_all))])
```

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

To investigate the cause for the missing coefficient estimates, the number of study individuals between the pairs of covariates listed above are cross-tabulated:

Table 1: cigarettes smoked per day (horizontal) against currently smoke (vertical)

	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 2: hypertension (horizontal) against anti-hypertensive medicine (vertical)

	No	Yes
No	957	1016
Yes	0	333

Table 1 and Table 2 reveal that none of the study individuals who are not currently smoking have any cigarette smoked each day, and that none of the study individuals who are on anti-hypertensive medication has hypertension. These situations are typical in the common sense and from the medical perspective. To avoid fitting problems, a slightly modified maximum model is considered with the following changes:

- No interaction between `cursmoke` and `cigpday`.
- No interaction between `bpmeds` and `prevhyp`.
- Add quadratic terms for the continuous covariates `age`, `sysbp`, `diabp`, `cigpday`, `bmi`, `heartrte`, `glucose`, and `hdlc`.

The updated maximal model is coded in R as:

```
# new full model as input to backward elimination
m_all <- lm(logit ~ (.)^2 - cursmoke:cigpday - bpmeds:prevhyp + I(age^2)
              + I(sysbp^2) + I(diabp^2) + I(cigpday^2) + I(bmi^2) + I(heartrte^2)
              + I(glucose^2) + I(hdlc^2), data = fhs_after_vif)
```

The input of backward elimination is the updated maximal model. Backward Elimination keeps dropping the least significant covariates that are in the current regression but not significant to the response variable, until all remaining covariates are significant.

3.1.3 Stepwise Selection

The starting point model of stepwise selection is the model somewhere between the maximal model used in backward elimination and the minimal model used in forward selection. Here the model with main effects only (i.e., with all the linear covariates) is adopted as the input to stepwise selection. Stepwise selection is a compromise between the above two approaches which can add the most significant or drop the least significant variable at any stage, until no further covariates can be added nor removed.

```
# start model of stepwise selection
m_regular <- lm(logit ~ ., data = fhs_after_vif)
```

3.1.4 Comparison of Automated Selections

The three models generated by three automated selection methods are compared from the perspectives of running efficiency, number of parameter, and coefficient of determination.

From the perspective of running efficiency, forward selection M_{fwd} is the fastest calculation, while backward elimination M_{bwd} is the slowest and stepwise selection M_{step} is somewhere between the two. Regarding the number of parameters, M_{bwd} and M_{step} have the same number (56) of coefficiencts, while M_{fwd} has only 1 fewer parameter (55) than the other two.

Before jumping into the coefficient of determination, whether models are nested within each other can be checked using `F`-test. Model M_1 is nested in M_2, if the covariates in M_1 are subset of those in M_2. However, the predictor `sysbp:age` is in M_{fwd} but not in M_{bwd} nor M_step. Similarly, `sex:sysbp` is in the M_{bwd} but not in the M_step, and conversly, `age:diabp` is in the M_step but not in the M_{bwd}. Since any of these models are not nested within each other, `F`-test cannot be applied to compare them.

The last criteria is the adjusted coefficient of determination R^2_{adj}, which measures the explanatory power (explained in Section 5). M_{bwd} has the largest R^2_{adj} value (0.8207), followed by M_{step} (0.8205), and M_{fwd} has the smallest R^2_{adj} (0.8196). Note that the R^2_{adj} difference between M_{bwd} and M_{fwd} is only 0.0002, which is quite marginal.

In general, backward elimination tends to perform better than forward selection, as forward selection is greedier in including marginally significant covariate. For example, suppose there are 3 covariates x_1, x_2, and x_3. The best regression is $y \quad 1 + x_1 + x_2$, but the most significant covariate in the presense of only the intercept is x_3. Then x_3 would be added to the forward selection model with no hope of achieving the regression. Also, that M_{bwd} is better than the M_{fwd} is supported by the larger R^2_{adj}.

The remaining is to compare the backward elimination and stepwise selection model. In terms of the coefficient of determination R^2_{adj}, M_{bwd} has marginally larger (0.0002) R^2_{adj} value. However, from the perspective of running efficiency, M_{bwd} runs more than 4 times longer than M_{fwd} does. Thus, in conclusion, the stepwise selection model M_{fwd} is chosen to be the candidate model constructed by the automated model selection.

3.2 Manual Model Selection

The second candidate model is manually constructed. The starting point of manual construction is the model including only main effects (i.e., all linear covariates without interaction effects) without colinear predictors `totchol` and `ldlc` identified in Section 2.2.

Insignificant predictors are identified by subjective assessment from Figure 1 and objective verification of significance test. The insignificance of the following predictors is derived from subjective justification, confirmed by significance test in Appendix 7, and removed from model in such order:

- Whether the study individuals currently smoke cigarettes (`cursmoke`) seems not to impact the risk of coronary heart disease (CHD) too much. Instead, what really matters should be how many cigarettes the study individuals smoke per day, which is represented by covariate `cigpday` in the presence of the current model.

- Figure 1 indicates that there is no apparent pattern in the pair plot of CHD risk and the diastolic blood pressure (`diabp`).
- Only individuals with stroke history and quite fragile immunity have small possibility of getting caught by CHD. Thus, whether individuals with stroke history or not would have marginal impact on the risk of CHD.

In conclusion, a reduced model is generated with the covariates `cursmoke`, `diabp`, and `prevstrk` dropped from the starting point model. Since the reduced model is nested in the starting point model, F-test is applied to verify whether the reduced version gives a better fitted model. The following code tests the null hypothesis

$$H_0 : \beta_{cursmoke} = \beta_{diabp} = \beta_{prevstrk} = 0$$

with significance level $\alpha = 0.05$.

```
# start point of manual constructed model
m_all_man <- lm(logit ~ ., data = fhs_after_vif)
# after removing cursmoke and diabp, remove prevstrk according to its largest p-value
m_red_man <- lm(logit ~ . - cursmoke - diabp - prevstrk, data = fhs_after_vif)
# get p-value of F-test with (manually constructed) full model vs. reduced model
c("p-value of F-test" = round(anova(m_red_man, m_all_man)[2, 6], 5))
```

```
## p-value of F-test
##           0.29097
```

Since p-value $=>\alpha = 0.05$, we do not reject the null hypothesis, which means the diastolic blood pressure, cigarette smoking, and stroke history are insignificant to the risk of CHD. Thus, the reduced model provides better regression.

Note that the current reduced model contains only main effects. Outstanding interaction effects from the stepwise selected model should considered to add to the manually reduced model. The promising interactions, as verified by F-test in Appendix 7, are added to the model in the following order:

- Since frequent cigarette smoking combined with low level of high density lipoprotein cholesterol polynomially increases the risk of CHD, interaction `cigpday:hdlc` is added to the model.
- Interaction between systolic blood pressure and whether the individual has had myocardial infarction is added, because the myocardial infarction history combined with low systolic blood pressure strongly increases the exposure to CHD (Paolo et al, 2014).
- Similarly, interaction is between systolic blood pressure and heart rate is supported by the fact that high heart rate and low systolic blood pressure are associated with an increased risk of CHD.
- The last promising interaction is between systolic blood pressure and whether the individual is on anti-hypertensive medication. The reason behind this consideration is people who are on anti-hypertensive medication and have low systolic blood pressure usually encounter high risk of CHD (Chirag et al, 2016).

Another interaction between whether the person currently smokes and high density lipoprotein cholesterol is considered, as cigarette smoker with low level of high density lipoprotein cholesterol has high risk of CHD. However, its significance is rejected by F-test performed in Appendix 7.

```
m_red_man$call
```

```
## lm(formula = logit ~ . - cursmoke - diabp - prevstrk + cigpday:hdlc +
##     sysbp:prevmi + sysbp:heartrte + sysbp:bpmeds, data = fhs_after_vif)
```

In conclusion, after all insignificant main effects are removed from the starting point model, the above 4 promising interaction effects derived from the automated model selection are added to the manually constructed model. Now, the manually constructed model consists of 17 coefficients with all p-value $< \alpha = 0.05$, which means all covariates in the current manually constructed model are significant to the risk of CHD.

4 Model Diagnostic

In the previous section, 2 candidate models are developed by stepwise selection and manual construction. This section performs model diagnostics on the candidate models by checking the violation of assumptions and ploting leverage and influence measures.

4.1 Residual Plots

Given the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, where\ \epsilon_i \quad i.i.d.\ N(0, \sigma^2),\ \ for\ i = 1, \ldots, n$$

, the following model assumptions are verified by residual plots:

- The conditional mean of y_i is linear to x_i: $E(y_i|x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.
- The conditional variance of y_i is constant: $Var(y_i|x_{i1}, \ldots, x_{ip}) = \sigma^2$.
- The errors $\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}$ are i.i.d. Normal distributed: $\epsilon_i \quad i.i.d.\ N(0, \sigma^2)$.

Both original and studentized residuals are adopted to verify the above assumptions. The original residuals are used to verify the assumptions on linear conditional mean and constant conditional variance, while the studentized residuals are utilized to verify the assumption on normality, given that studentized residuals are most likely to be approximately normal compared to the standardized ones if the model is correct.

Figure 2 plots the residuals against the fitted values for both stepwise selected and manually constructed models. According to Figure 2, observations in both graphs distribute randomly around the horizontal line $e_i = y_i - \hat{y}_i = 0$, which indicates the assumption of linear conditinal mean is satisfied by both models. In terms of constant variance, since the observations in both graphs (in Figure 2) remain in the constant range without variation, the assumption of constant conditional variance is met by both models.

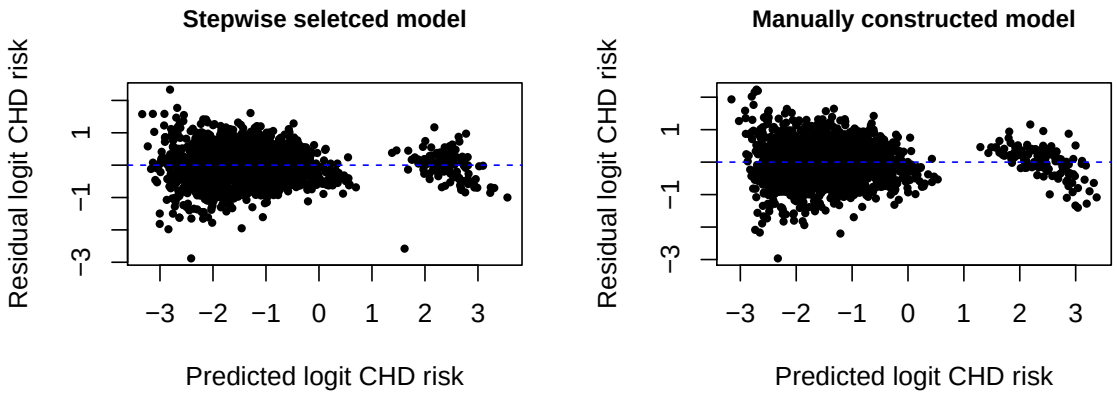

Figure 2: Residuals against predicted values.

Figure 3 and Figure 4 plot the histogram and QQ-plot of studentized residuals with theoretical normal distribution. From the histograms in both Figure 3 and Figure 4, it is illustrated that the distribution of residuals $e_i = y_i - \hat{y}_i$ in both stepwise selected and manually constructed models are approximately symmetric and roughly fit the blue normal distrubution curve. Additionally, QQ-plots in both Figure 3 and Figure 4 suggest that studentized residuals of both models distribute around the straight line. Thus, the assumption of normality is satisfied by both models. However, according to the histogram

in Figure 4, the manually constructed model might slighly violate the normality assumption, as the residuals seem somewhat negatively biased.

In conclusion, both stepwise selected and manually generated models do not violate the assumptions of linear conditional mean and constant conditional variance. However, while the stepwise selected model does not violate normality, the manually constructed model marginally violate the normality by being somewhat negatively biased.

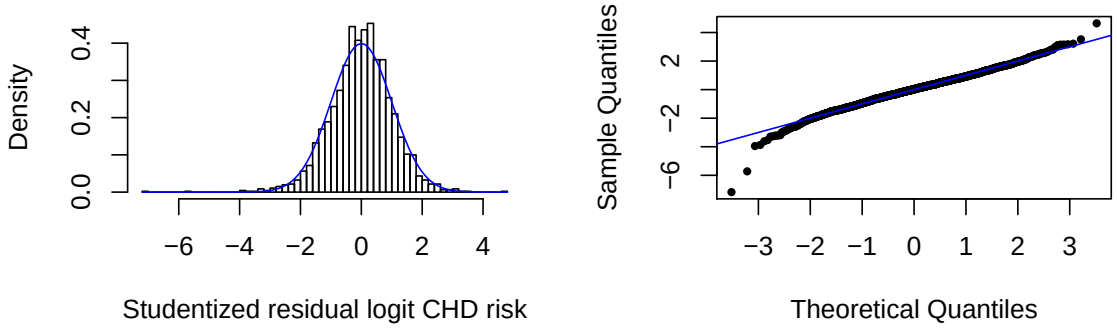

Figure 3: Stepwise selected model: Histogram and QQ-plot of studentized residuals with theoretical normal distribution.

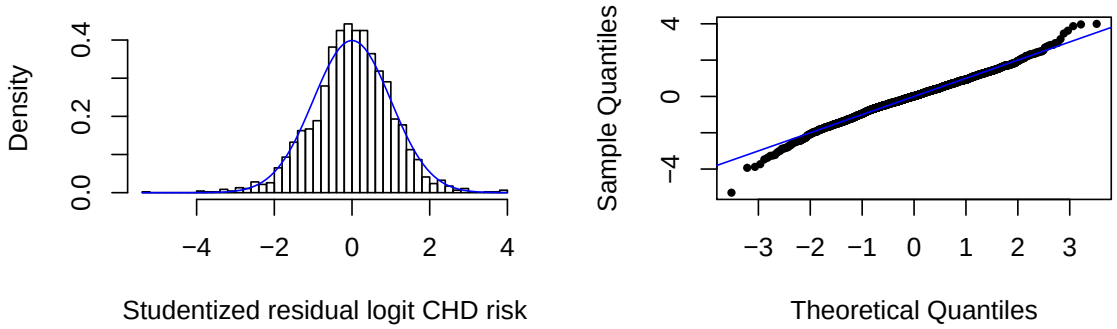

Figure 4: Manually constructed model: Histogram and QQ-plot of studentized residuals with theoretical normal distribution.

4.2 Leverage and Influence Measures

Outliers, which are observations with unusually large residuals compared to others, are typically the result of important covariates having been ommited from the model. Thus, the fewer the outliers, the better the model fitting. Revealed by the residual plots in Figure 2, the stepwise selected model generates approximately 2 outliers which are at the very bottom of the left graph, while the manually constructed mmodel derives roughly 2 outliers that are at the bottom left corner of the right graph. Since both models have approximately 2 outliers, both models somewhat omit marginal number of significant predictors.

While outliers are based on subjective assessment from residual plots, leverage is another criteria that provides objective judgement on the model fitting. An observation is a leverage point if it has unusually

large or small explanatory varibales with extremely large leverage. Thus, the fewer the leverage points, the better the regression fitting. Leverage values which are twice the average value are considered to be high. Figure 5 plots cook's distance against leverage, where the leverage points with leverage values more than twice the average value are marked as green. Derived from Figure 5, there are 232 leverage points generated by the stepwise selected model, whereas there are 30 fewer (i.e., 202) leverage points discovered in the manually constructed model. Therefore, the manually constructed model provides better fitness than the stepwise selected one in terms of leverage.

In addition to the outlier and leverage, influential observation is also utilized to compare the candidate models. An observation is an influential observation if removing it would markedly change the statistical analysis. In other words, the influential observation represents the impact of excluding a particular data point. Such observations are highlighted by red in Figure 5. It can be observed in left graph of Figure 5 that one observation (in red) has almost 8 times the Cook's distance of the others; whereas in right part of Figure 5, the red observation has only about 2 times the Cook's distance of the others. Thus, although both models leave 1 influential observation, the one left by the stepwise selected model would generate larger impact if being excluded.

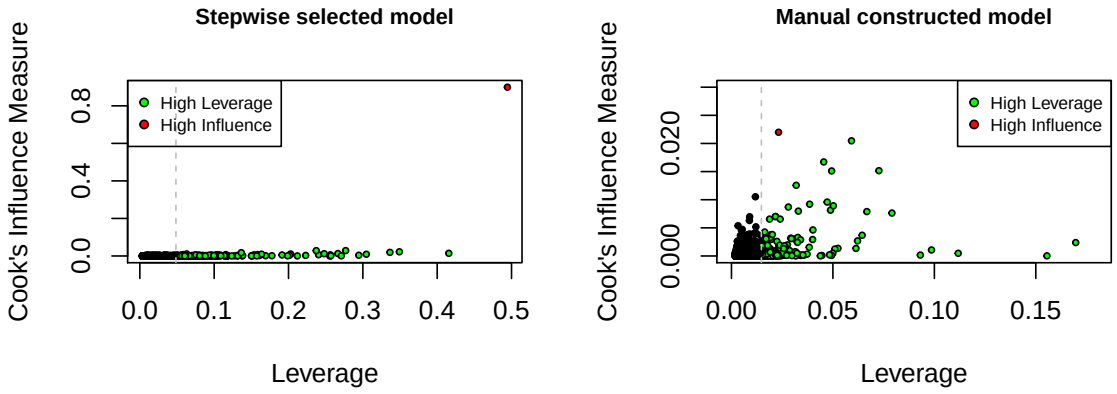

Figure 5: Cook's distance against leverage.

In conclusion, both stepwise selected and manually generated models are at the same level of fittness in terms of outlier. However, regarding leverage and influential observation, the manually generated model would be preferred over the stepwise selected one with fewer leverage points and smaller impact if the influential observation is removed.

5 Model Selection

This section judges the two candidate models from the perspective of predictive and explanatory power. Combined with diagnostics performed in the previous section regarding assumption validation, outlier, leverage, and influential observation, one of the two candidate models would be chosen to regress the sample data in this Framingham Heart Study.

5.1 Explanatory Power

Here the adjusted coefficient of determination R^2adj is utilized to assess the explanatory power of the two candidate models. R^2adj measures the proportion of total variation in the response variable that is explained by the regression model (the explanatory variables). Additionally, adding a new significant covariate to the model would increase R^2_{adj} value, while adding a new insignificant covariate to the model would decrease R^2_{adj} value. Thus, the larger the R^2_{adj} value, the better the model. Demonstrated by the following calculation, the stepwise selected model has larger R^2adj value (0.82)

than the manually constructed model does (0.78). Thus, the stepwise selected model overweighs the manually constructed one in terms of the explanatory power.

5.2 Predictive Power

Cross-validation analysis is performed to assess the predictive power of the candidate models. Cross-validation is an excellent approach to model selection, as it avoids favoring the model that overfits data by diving the n observation into disjoint training and testing subsets. However, it does have some caveats such as expensive computation and being entirely based on predictive as opposed to explanatory power. Figure 6 displays the square root of the Mean Square Prediction Error (rMSPE) statistic for a cross-validation comparison between the stepwise selected model and manually constructed model. The difference between the medians in Figure 6 is small (only around 0.01), but there is preference for the richer model which is generated by stepwise model selection for its smaller median of rMSPE.

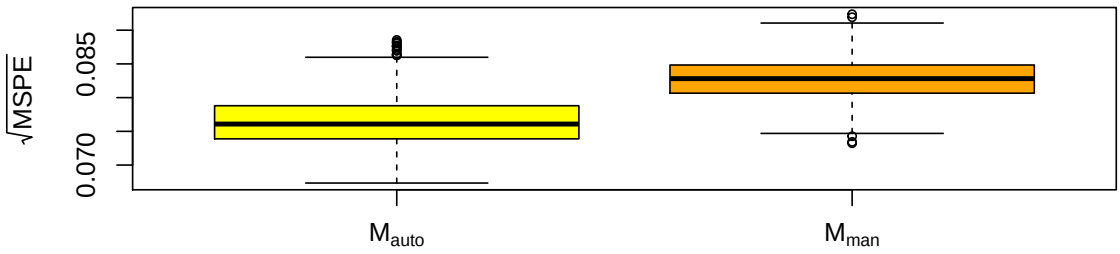

Figure 6: Cross-validation model comparison results: Root mean square prediction error (rMSPE) of stepwise selected model.

5.3 Final Model

The final model is selected from the perspectives of assumption validation, outlier, leverage, influential observation, explainatory power, and predictor power. From Section 4, it can be concluded that stepwise selected model is preferred over the manually constructed one regarding assumption validation, as the manually constructed one slightly violates the normality assumption. In terms of outlier, leverage, and influential observation, the manually constructed model is better for fewer leverage points and smaller impact if the influential observation is removed. However, assumption validation should be given priority over outlier, leverage, and influential observation. If the model violates any assumption, it does not even interpret data in the correct manner. Thus, the stepwise selected model is prefered. From the perspective of explanatory and predictive power (in Section 5), it is still the stepwise selected model that has stronger powers.

To conclude, the stepwise selected model is chosen to be the final model with all assumptions validated and relatively strong explanatory and predictive power. The summary of the final model about its parameter estimates, standard errors, and p-value (significance of predictors) is displayed in Table 3.

6 Discussion

This Framingham Heart Study explores the relation between the quantified risk for coronary heart disease (CHD) and some explanatory variables. This study indicates that smoking, being diabetic, having hypertension, and presence of myocardial infarction history are strongly associated with high

Table 3: Summary on parameter estimates, standard errors, and p-values of final model.

	Estimate	Std. Error	P-value		Estimate	Std. Error	P-value
(Intercept)	-7.79569	1.55204	0.00000	diabp:hdlc	-0.00024	0.00007	0.00135
sexMale	0.95986	0.17613	0.00000	sysbp:diabetesYes	-0.00634	0.00184	0.00057
age	0.11138	0.02888	0.00012	prevmiYes:hdlc	0.01527	0.00406	0.00017
sysbp	0.01141	0.00672	0.08969	age:heartrte	-0.00020	0.00011	0.05700
diabp	-0.02645	0.01827	0.14793	diabp:bmi	-0.00092	0.00029	0.00148
cursmokeYes	1.06933	0.30030	0.00038	sexMale:glucose	-0.00206	0.00074	0.00571
cigpday	0.00608	0.00675	0.36760	prevhypYes:hdlc	-0.00573	0.00188	0.00240
bmi	-0.07754	0.03966	0.05071	age:prevhypYes	-0.01053	0.00387	0.00651
diabetesYes	1.09256	0.31467	0.00053	cigpday:hdlc	-0.00034	0.00010	0.00100
bpmedsYes	0.62343	0.27365	0.02281	bmi:prevhypYes	-0.02035	0.00771	0.00833
heartrte	0.01407	0.00979	0.15089	age:cursmokeYes	-0.00919	0.00313	0.00331
glucose	0.00504	0.00227	0.02645	diabp:glucose	-0.00007	0.00003	0.01054
prevmiYes	5.17146	0.47148	0.00000	prevmiYes:prevhypYes	-0.28433	0.13907	0.04102
prevstrkYes	0.22511	0.08528	0.00836	prevmiYes:prevstrkYes	-0.42638	0.20950	0.04195
prevhypYes	3.87041	0.48215	0.00000	diabp:prevhypYes	-0.00657	0.00403	0.10331
hdlc	-0.00108	0.00969	0.91123	diabp:cursmokeYes	-0.00657	0.00217	0.00256
I(hdlc^2)	0.00030	0.00002	0.00000	age:bmi	0.00064	0.00038	0.08827
I(diabp^2)	0.00056	0.00008	0.00000	sexMale:age	-0.00658	0.00277	0.01778
I(cigpday^2)	0.00027	0.00008	0.00061	sysbp:bpmedsYes	-0.00407	0.00162	0.01201
I(bmi^2)	0.00269	0.00047	0.00000	diabetesYes:hdlc	0.00525	0.00262	0.04516
I(heartrte^2)	0.00017	0.00005	0.00035	age:hdlc	-0.00018	0.00010	0.06584
I(glucose^2)	0.00001	0.00000	0.00429	bpmedsYes:hdlc	0.00309	0.00208	0.13869
I(sysbp^2)	0.00006	0.00002	0.00950	sysbp:heartrte	-0.00015	0.00004	0.00031
I(age^2)	-0.00031	0.00017	0.07646	bmi:prevmiYes	-0.02125	0.01320	0.10758
sysbp:prevmiYes	-0.01155	0.00270	0.00002	heartrte:hdlc	-0.00010	0.00006	0.09375
age:diabp	-0.00024	0.00016	0.13647	cursmokeYes:bpmedsYes	-0.24368	0.11771	0.03855
diabetesYes:prevmiYes	-0.73835	0.14299	0.00000	cigpday:bpmedsYes	0.00776	0.00505	0.12475
sysbp:prevhypYes	-0.01208	0.00251	0.00000	cursmokeYes:hdlc	0.00370	0.00251	0.14093

CHD risk, given that the covariates `cursmoke`, `diabetes`, `prevhyp` and `prevmi` have positive and comparatively large (greater than 1) coefficient estimates in the final model as shown in Table 3. On the other hand, individuals with low level of high density lipoprotein cholesterol, low diastolic blood pressure, low casual serum glucose, and fewer cigarettes per day are exposed to low CHD risk, which is supported by the positive but relatively small (smaller than 0.001) coefficient estimates of the predictors `hdlc^2`, `diabp^2`, `glucose^2`, and `cigpday^2` in the final model as shown in Table 3.

In addition to the rich knowledge gained from this study, some behavioral changes can be recommended to lower the risk of CHD. For example, a male individual who currently smokes and suffer from either diabetes, hypertension, or myocardial infarction history would be recommended to quit smoking or lower the number of cigarettes smoked per day, intake food with low fat but high protein, reduce the intake of salt and sugar, and exercise regularly. Such behavioral change is recommended based on the positive and strong relation between smoking (`cursmoke` and `cigpday`), diabetes (`diabetes`), hypertension(`prevhyp`), myocardial infarction history (`prevmi`) and CHD risk. The strong and positive relations are demonstrated by the positive and relatively large parameter estimates in Table 3. Lowering the number of cigarettes smoked per day or quitting smoking would lower the value of `cidpday`. Intaking food with low fat but high protein would ease diabetes situation. Similarly, reducing the intake of salt and sugar would avoid exaggeration of hypertension, and exercising regularly would mitigae both daibetes and hypertension. All these behavioral changes would finally reduce CHD risk.

Though the final model does not violate any regression assumption, there are a few (approximately 3) coefficients with high `p`-values (i.e. `p`-value $> 0.05 =$ significance level α). It is possible as the covariate itself is insignificant, but its quardratic form or interaction with other predictors are vitally important. For example, as shown in the Table 3, high density lipoprotein cholesterol is insignificant, given the covariate `hdlc`'s `p`-value $= 0.9 > \alpha = 0.05$, However, its quardratic form `hdlc^2` and its interaction with multiple categorical variables such as `prevmi:hdlc` and `prevhyp:hdlc` are all significant to CHD risk, as all corresponding `p`-values are significantly smaller than 0.05. This also indicates that high density lipoprotein cholesterol itself is not important to estimate CHD risk. However, individuals who have high level of high density lipoprotein cholesterol and hypertension or myocardial infarction history should be cautious for preventing CHD.

7 Appendix

All R code for Section 2:

```
# Appendix for Sec 2

fhs <- read.csv("fhs.csv", header = TRUE)
# summary statistics
summary(fhs)

# pair plot of all continuous variables
pairs(~ chdrisk + totchol + age + sysbp + diabp + cigpday + bmi + heartrte + glucose
      + hdlc + ldlc, pch = 16, cex = 0.1, data = fhs)
```

```
# Get all continuous (non-categorical) variable
fhs_noc_index <- unlist(lapply(fhs, is.numeric))
fhs_noc <- fhs[ , fhs_noc_index]

# Design matrix excluding intercept and all categorical variables
X <- model.matrix(lm(chdrisk ~ . - 1, data = fhs_noc))

# VIF of continuous variable: calculate VIF using R^2
# @param x: continuous predictors for which to calculate R^2
# @return VIF for the predictor x
vif_R2 <- function(x) {
  # formula to regress x on other continuous predictors (other than chdrisk)
  regress <- formula(paste0(x, "~ . - chdrisk"))
  M <- lm(regress, data = fhs_noc)
  R2 <- summary(M)$r.square # extract R^2
  1/(1-R2)
}

# Apply the function vif to every continuous predictor
vif <- sapply(colnames(X), vif_R2)
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
```

```
# method 2 to verify correctness of VIF
#C <- cor(X)
#vif2 <- diag(solve(C))
#vif - vif2

# drop highest in this case totchol, recalculate vif
fhs_drop1 <- subset(fhs, select = -c(totchol))
X_d1 <- model.matrix(lm(chdrisk ~ . - 1 - totchol, data = fhs_noc))
vifd1 <- sapply(colnames(X_d1), vif_R2)
vifd1
```

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

```r
# drop highest in this case ldlc, recalculate vif
fhs_drop2 <- subset(fhs, select = -c(totchol, ldlc))
X_d2 <- model.matrix(lm(chdrisk ~ . - 1 - totchol - ldlc, data = fhs_noc))
vifd2 <- sapply(colnames(X_d2), vif_R2)
vifd2
```

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

```r
c(any_colinear = any(vifd2[vifd2 > 10]))
```

```
## any_colinear
##        FALSE
```

All R code for Section 3:

```r
# Appendix for Sec 3

fhs_after_vif <- fhs

# modify the fhs data to have logit column value given by log(chdrisk) - log(1-chdrisk).
fhs_after_vif$logit <- log(fhs_after_vif$chdrisk) - log(1-fhs_after_vif$chdrisk)

# modify the fhs data to delete columns identified by VIF above
fhs_after_vif <- subset(fhs_after_vif, select = -c(totchol,ldlc,chdrisk) )

# start model of forward method
m_empty <- lm(logit ~ 1, data = fhs_after_vif)
```

```r
# start model of backward method
m_all <- lm(logit ~ (.)^2, data = fhs_after_vif)

length(coef(m_all))
```

```
## [1] 121
```

```r
# checking if there exist NA coef in m_all
anyNA(coef(m_all))
```

```
## [1] TRUE
```

```r
# Coefficients with NA value
names(coef(m_all)[is.na(coef(m_all))])
```

```r
# cross-tabulation of cigarettes smoked per day against currently smoke
kable(table(fhs_after_vif[c("cursmoke", "cigpday")]), booktabs = T, caption =
        "\\label{tab:cross1} cigarettes smoked per day (horizontal) against
        currently smoke (vertical)") %>%
  kable_styling(latex_options = c("striped", "scale_down", "hold_position"))

# cross-tabulation of hypertension against anti-hypertensive medicine
kable(table(fhs_after_vif[c("bpmeds","prevhyp")]), booktabs = T, caption =
        "\\label{tab:cross2} hypertension (horizontal) against
        anti-hypertensive medicine (vertical)") %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = "center", font_size = 7, full_width = T)
```

```r
# new full model as input to backward elimination
m_all <- lm(logit ~ (.)^2 - cursmoke:cigpday - bpmeds:prevhyp + I(age^2)
              + I(sysbp^2) + I(diabp^2) + I(cigpday^2) + I(bmi^2) + I(heartrte^2)
              + I(glucose^2) + I(hdlc^2), data = fhs_after_vif)

# check if there is still any NA
anyNA(coef(m_all))
```

```
## [1] FALSE
```

```r
# start model of stepwise selection
m_regular <- lm(logit ~ ., data = fhs_after_vif)
```

```r
if(!params$load_calcs) {
# forward selection
  system.time({
    m_fwd <- step(object = m_empty, # starting point empty model
                  scope = list(lower = m_empty, upper = m_all), # empty and full model
                  direction = "forward",
                  trace = FALSE)
  })
  # backward elimination
  system.time({
    m_bwd <- step(object = m_all, # starting point full (maximum) model
                  scope = list(lower = m_empty, upper = m_all), # empty and full model
                  direction = "backward",
                  trace = FALSE)
  })
  # stepwise selection
  system.time({
    m_step <- step(object = m_regular,
                      # starting point regular model (include all main effects)
                  scope = list(lower = m_empty, upper = m_all), # empty and full model
                  direction = "both",
                  trace = FALSE)
  })
  # if first time compile, same the models, since calculation takes a long time
  saveRDS(list(mf = m_fwd, mb = m_bwd, ms = m_step), file = "AutoModels.rds")
} else {
  # simply load values back
  tmp <- readRDS("AutoModels.rds")
  m_fwd <- tmp$mf
  m_bwd <- tmp$mb
  m_step <- tmp$ms
  rm(tmp) # optionally remove tmp from workspace
}
```

```r
# display the number of covariates in each model
c(fwd = length(coef(m_fwd)), bwd = length(coef(m_bwd)), step = length(coef(m_step)))
```

```
##  fwd  bwd step
##   55   56   56
```

```r
# display all covariates in forward selected model
col_fwd <- names(coef(m_fwd))
col_fwd
```

```
##  [1] "(Intercept)"           "prevmiYes"             "sysbp"
##  [4] "sexMale"               "age"                   "prevhypYes"
##  [7] "hdlc"                  "I(hdlc^2)"             "diabetesYes"
## [10] "I(bmi^2)"              "I(cigpday^2)"          "I(heartrte^2)"
## [13] "I(sysbp^2)"            "bmi"                   "bpmedsYes"
## [16] "I(glucose^2)"          "prevstrkYes"           "cursmokeYes"
## [19] "heartrte"              "cigpday"               "I(diabp^2)"
## [22] "diabp"                 "prevmiYes:sysbp"       "sysbp:age"
## [25] "prevmiYes:diabetesYes" "prevhypYes:hdlc"       "sysbp:prevhypYes"
## [28] "prevmiYes:hdlc"        "sysbp:diabetesYes"     "prevhypYes:bmi"
## [31] "prevmiYes:prevhypYes"  "sysbp:hdlc"            "sexMale:age"
## [34] "sysbp:bpmedsYes"       "prevmiYes:bmi"         "sysbp:cursmokeYes"
## [37] "age:cursmokeYes"       "sexMale:diabetesYes"   "sysbp:heartrte"
## [40] "age:heartrte"          "hdlc:cigpday"          "age:bmi"
## [43] "hdlc:diabetesYes"      "sysbp:bmi"             "prevmiYes:bpmedsYes"
## [46] "heartrte:diabp"        "diabetesYes:diabp"     "age:prevhypYes"
## [49] "prevhypYes:diabp"      "hdlc:diabp"            "cursmokeYes:diabp"
## [52] "bmi:diabp"             "age:hdlc"              "hdlc:prevstrkYes"
## [55] "hdlc:heartrte"
```

```
# display all covariates in backward eliminated model
col_bwd <- names(coef(m_bwd))
col_bwd
```

```
##  [1] "(Intercept)"           "sexMale"               "age"
##  [4] "sysbp"                 "diabp"                 "cursmokeYes"
##  [7] "cigpday"               "bmi"                   "diabetesYes"
## [10] "bpmedsYes"             "heartrte"              "glucose"
## [13] "prevmiYes"             "prevstrkYes"           "prevhypYes"
## [16] "hdlc"                  "I(sysbp^2)"            "I(diabp^2)"
## [19] "I(cigpday^2)"          "I(bmi^2)"              "I(heartrte^2)"
## [22] "I(glucose^2)"          "I(hdlc^2)"             "sexMale:age"
## [25] "sexMale:sysbp"         "sexMale:glucose"       "sexMale:prevhypYes"
## [28] "age:sysbp"             "age:cursmokeYes"       "age:bmi"
## [31] "age:heartrte"          "age:prevhypYes"        "age:hdlc"
## [34] "sysbp:diabetesYes"     "sysbp:bpmedsYes"       "sysbp:heartrte"
## [37] "sysbp:prevmiYes"       "sysbp:prevhypYes"      "diabp:cursmokeYes"
## [40] "diabp:bmi"             "diabp:glucose"         "diabp:prevhypYes"
## [43] "diabp:hdlc"            "cursmokeYes:bpmedsYes" "cursmokeYes:hdlc"
## [46] "cigpday:bpmedsYes"     "cigpday:hdlc"          "bmi:prevmiYes"
## [49] "bmi:prevhypYes"        "diabetesYes:prevmiYes" "diabetesYes:hdlc"
## [52] "heartrte:hdlc"         "prevmiYes:prevstrkYes" "prevmiYes:prevhypYes"
## [55] "prevmiYes:hdlc"        "prevhypYes:hdlc"
```

```
# display all covariates in stepwise selected model
col_step <- names(coef(m_step))
col_step
```

```
##  [1] "(Intercept)"           "sexMale"               "age"
##  [4] "sysbp"                 "diabp"                 "cursmokeYes"
##  [7] "cigpday"               "bmi"                   "diabetesYes"
## [10] "bpmedsYes"             "heartrte"              "glucose"
## [13] "prevmiYes"             "prevstrkYes"           "prevhypYes"
## [16] "hdlc"                  "I(hdlc^2)"             "I(diabp^2)"
## [19] "I(cigpday^2)"          "I(bmi^2)"              "I(heartrte^2)"
## [22] "I(glucose^2)"          "I(sysbp^2)"            "I(age^2)"
```

```
## [25] "sysbp:prevmiYes"       "age:diabp"             "diabetesYes:prevmiYes"
## [28] "sysbp:prevhypYes"      "diabp:hdlc"            "sysbp:diabetesYes"
## [31] "prevmiYes:hdlc"        "age:heartrte"          "diabp:bmi"
## [34] "sexMale:glucose"       "prevhypYes:hdlc"       "age:prevhypYes"
## [37] "cigpday:hdlc"          "bmi:prevhypYes"        "age:cursmokeYes"
## [40] "diabp:glucose"         "prevmiYes:prevhypYes"  "prevmiYes:prevstrkYes"
## [43] "diabp:prevhypYes"      "diabp:cursmokeYes"     "age:bmi"
## [46] "sexMale:age"           "sysbp:bpmedsYes"       "diabetesYes:hdlc"
## [49] "age:hdlc"              "bpmedsYes:hdlc"        "sysbp:heartrte"
## [52] "bmi:prevmiYes"         "heartrte:hdlc"         "cursmokeYes:bpmedsYes"
## [55] "cigpday:bpmedsYes"     "cursmokeYes:hdlc"
```

```
# check if any of the models are nested within each other
col_fwd[!(col_fwd %in% col_bwd)]
col_fwd[!(col_fwd %in% col_step)]
col_bwd[!(col_bwd %in% col_step)]
col_step[!(col_step %in% col_bwd)]
```

```
# compare these models using adjusted R^2
c("fwd_adj_R2" = round(summary(m_fwd)$adj.r.squared, 4),
  "bwd_adj_R2" = round(summary(m_bwd)$adj.r.squared, 4),
  "step_adj_R2" = round(summary(m_step)$adj.r.squared, 4))
```

```
##  fwd_adj_R2  bwd_adj_R2 step_adj_R2
##      0.8196      0.8207      0.8205
```

Appendix for Sec 3.2

```
m_all_man <- lm(logit ~ ., data = fhs_after_vif)

# find all insignificant x by p value
p_val <- summary(m_all_man)$coefficients[,4]
names(p_val)[!(p_val < 0.05)]
```

```
## [1] "diabp"       "cursmokeYes" "prevstrkYes"
```

```
# delete insignificant x
# first, remove cursmoke since cursmoke has the most significant p_value
#   which is > 0.05
m_red_man <- lm(logit ~ . - cursmoke, data = fhs_after_vif)
p_val <- summary(m_red_man)$coefficients[,4]
names(p_val)[!(p_val < 0.05)]
```

```
## [1] "diabp"       "prevstrkYes"
```

```
# after removing cursmoke, remove diabp since diabp has the
#    most significant p_value which is > 0.05
m_red_man <- lm(logit ~ . - cursmoke - diabp, data = fhs_after_vif)
p_val <- summary(m_red_man)$coefficients[,4]
names(p_val)[!(p_val < 0.05)]
```

```
## [1] "prevstrkYes"
```

```r
# after removing cursmoke and diabp, remove prevstrk since prevstrk has
#    the most significant p_value which is > 0.05
m_red_man <- lm(logit ~ . - cursmoke - diabp - prevstrk, data = fhs_after_vif)
p_val <- summary(m_red_man)$coefficients[,4]
c(insignificant_x = any(!(p_val < 0.05)))
```

```
## insignificant_x
##           FALSE
```

```r
names(coef(m_red_man))
```

```
##  [1] "(Intercept)" "sexMale"     "age"         "sysbp"       "cigpday"
##  [6] "bmi"         "diabetesYes" "bpmedsYes"   "heartrte"    "glucose"
## [11] "prevmiYes"   "prevhypYes"  "hdlc"
```

```r
# get p-value of F-test with (manually constructed) full model vs. reduced model
c("reduced man vs. full man" = round(anova(m_red_man, m_all_man)[2, 6], 5))
```

```
## reduced man vs. full man
##                  0.29097
```

```r
# this part is for checking our assumptions on
#  interaction effect that is manually added to our initial model

# covariates in stepwise selected model
name_step <- names(coef(m_step))
# covariates currently in manual constructed model
name_manul <- names(coef(m_red_man))

count <- 1
# F test of manual constructed model with and with out promising
#  interation effect
Fs <- matrix(nrow = 2, ncol = length(coef(m_bwd)))
for (i in 1:length(coef(m_step))) {
  name_needed <- name_step[i]
  # if the covariate is in manual model already
  if (!(name_needed %in% name_manul)) {
    # if it is an interaction effect
    if (length(grep(":+", name_needed)) == 1) {
      a <- strsplit(name_needed, ":")[[1]]
      a1 <- a[1]
      a2 <- a[2]

      # eliminate the Yes/Male in the name of interaction effect
      # modify the name so that it could fit into model as x
      if (length(grep("Y+", a1)) == 1) {
        a1 <- substr(a1, 1, nchar(a1) - 3)
      }
      if (length(grep("Y+", a2)) == 1) {
        a2 <- substr(a2, 1, nchar(a2) - 3)
      }
      if (length(grep("M+", a1)) == 1) {
        a1 <- substr(a1, 1, nchar(a1) - 4)
      }
      if (length(grep("M+", a2)) == 1) {
```

```
        a2 <- substr(a2, 1, nchar(a2) - 4)
      }
      name_needed <- paste(a1, ":", a2)
    } else if (length(grep("Y+", name_needed)) == 1) {
      # modify the name so that it could fit into model as x
      name_needed <- substr(name_needed, 1, nchar(name_needed) - 3)
    }

    # setup regression for themanual constructed model with promising
    #  interation effect
    regress <- formula(paste0("logit ~ sex + age+ sysbp +
                              cigpday + bmi + diabetes +
                              bpmeds + heartrte + glucose +
                              prevmi + prevhyp + hdlc + ", name_needed))
    M <- lm(regress, data = fhs_after_vif)

    # save the p-value of F-test
    Fs[1, count] <- name_needed
    Fs[2, count] <- anova(M, m_red_man)[2,6]
    count <- count + 1
  }
}
Fs <- Fs[,!(is.na(Fs[2, ]))]

# models that have p_value of F test less than 0.05
Fs[1, as.numeric(Fs[2, ]) < 0.05]
```

```
##  [1] "I(hdlc^2)"         "I(diabp^2)"        "I(bmi^2)"
##  [4] "I(sysbp^2)"        "sysbp : prevmi"    "diabetes : prevmi"
##  [7] "sysbp : prevhyp"   "sysbp : diabetes"  "age : heartrte"
## [10] "sex : glucose"     "prevhyp : hdlc"    "age : prevhyp"
## [13] "cigpday : hdlc"    "prevmi : prevhyp"  "prevmi : prevstrk"
## [16] "age : bmi"         "sysbp : bpmeds"    "sysbp : heartrte"
## [19] "bmi : prevmi"
```

```
# Add promising interaction effect from stepwise selection: cigpday:hdlc
m_with_interact <- lm(logit ~ . - cursmoke - diabp -
                        prevstrk + cigpday:hdlc,
                      data = fhs_after_vif)
round(anova(m_red_man, m_with_interact)[2, 6], 5)
```

```
## [1] 0.00399
```

```
m_red_man <- m_with_interact

# Add promising interaction effect: sysbp:prevmi
m_with_interact <- lm(logit ~ . - cursmoke - diabp -
                        prevstrk + cigpday:hdlc +
                        sysbp:prevmi,
                      data = fhs_after_vif)
anova(m_red_man, m_with_interact)[2, 6]
```

```
## [1] 3.206591e-12
```

```r
m_red_man <- m_with_interact

# Add promising interaction effect: cursmoke:hdlc
m_with_interact <- lm(logit ~ . - cursmoke - diabp -
                        prevstrk + cigpday:hdlc +
                        sysbp:prevmi + cursmoke:hdlc,
                      data = fhs_after_vif)
round(anova(m_red_man, m_with_interact)[2, 6], 5)
```

```
## [1] 0.38757
```

```r
# Add promising interaction effect: sysbp:heartrte
m_with_interact <- lm(logit ~ . - cursmoke - diabp -
                        prevstrk + cigpday:hdlc +
                        sysbp:prevmi + sysbp:heartrte,
                      data = fhs_after_vif)
round(anova(m_red_man, m_with_interact)[2, 6], 5)
```

```
## [1] 1e-05
```

```r
m_red_man <- m_with_interact

# Add the last promising interaction effect: sysbp:bpmeds
m_with_interact <- lm(logit ~ . - cursmoke - diabp -
                        prevstrk + cigpday:hdlc +
                        sysbp:prevmi + sysbp:heartrte +
                        sysbp:bpmeds,
                      data = fhs_after_vif)
round(anova(m_red_man, m_with_interact)[2, 6], 5)
```

```
## [1] 0.00644
```

```r
m_red_man <- m_with_interact

# check all of the coefficients in updated manual model are significant
m_red_man$call
```

```
## lm(formula = logit ~ . - cursmoke - diabp - prevstrk + cigpday:hdlc +
##     sysbp:prevmi + sysbp:heartrte + sysbp:bpmeds, data = fhs_after_vif)
```

```r
round(summary(m_red_man)$coefficients[,4], 5)
```

```
##     (Intercept)         sexMale             age           sysbp         cigpday
##         0.00000         0.00000         0.00000         0.00000         0.00001
##             bmi     diabetesYes       bpmedsYes        heartrte         glucose
##         0.00000         0.00000         0.00178         0.00000         0.00094
##       prevmiYes      prevhypYes            hdlc    cigpday:hdlc sysbp:prevmiYes
##         0.00000         0.00000         0.00000         0.00191         0.00000
##  sysbp:heartrte sysbp:bpmedsYes
##         0.00002         0.00644
```

All R code for Section 4:

```r
par(mfrow = c(1, 2))
# Check assumptions for m_step
# mean0 and constant var
res_auto <- residuals(m_step)
plot(predict(m_step), res_auto, pch = 16, cex = 0.7,
     xlab = "Predicted logit CHD risk", ylab = "Residual logit CHD risk",
     main = "Stepwise seletced model", cex.main = 0.9)
abline(h = 0, lty = 2, col = "blue")

# Check assumptions for m_red_man
# mean0 and constant var
res_man <- residuals(m_red_man)
plot(predict(m_red_man), res_man, pch = 16, cex = 0.7,
     xlab = "Predicted logit CHD risk", ylab = "Residual logit CHD risk",
     main = "Manually constructed model", cex.main = 0.9)
abline(h = 0, lty = 2, col = "blue")

# Check normality with studentized residual of automated model
sigma_hat_auto <- sigma(m_step)
X_auto <- model.matrix(m_step) # design matrix
H_auto <- X_auto %*% solve(crossprod(X_auto), t(X_auto)) # hat matrix
h_auto <- diag(H_auto)
res_stu_auto <- resid(m_step) / (sqrt(1 - h_auto) * sigma_hat_auto)
par(mfrow = c(1, 2)) # formatting
# histogram
hist(res_stu_auto, breaks = 50, freq = FALSE, cex = 0.7,
     xlab = "Studentized residual logit CHD risk", main = "")
curve(dnorm(x), col = "blue", add = TRUE) # theoritecal normal curve
# qq-plot
qqnorm(res_stu_auto, pch = 16, cex = 0.7, main = "")
abline(a = 0,b = 1, col = "blue") # add 45 degree line

# Check normality with studentized residual of manual model
sigma_hat_man <- sigma(m_red_man)
X_man <- model.matrix(m_red_man) # design matrix
H_man <- X_man %*% solve(crossprod(X_man), t(X_man)) # hat matrix
h_man <- diag(H_man)
res_stu_man <- resid(m_red_man) / (sqrt(1 - h_man) * sigma_hat_man)
par(mfrow = c(1, 2)) # formatting
# histogram
hist(res_stu_man, breaks = 50, freq = FALSE, cex = 0.7,
     xlab = "Studentized residual logit CHD risk", main = "")
curve(dnorm(x), col = "blue", add = TRUE) # theoritecal normal curve
# qq-plot
qqnorm(res_stu_man, pch = 16, cex = 0.7, main = "")
abline(a = 0,b = 1, col = "blue") # add 45 degree line
```

```r
par(mfrow = c(1, 2))

# leverage & influence measure
p_auto <- length(coef(m_step))
n <- nobs(m_step)
# h_auto <- hatvalues(m_step)
hbar_auto <- p_auto/n # average leverage
# 2*mean(h_auto)
D_auto <- cooks.distance(m_step) # cook's distance
influence_index_auto <- which.max(D_auto) # top influence point
```

```r
leverage_index_auto <- (h_auto > 2*hbar_auto) # leverage more than 2x the average

# set colours in the plot
colours_auto <- rep("black", length = n)
colours_auto[leverage_index_auto] <- "green"
colours_auto[influence_index_auto] <- "red"

# count the number of high leverage
num_leverage_auto <- c(num_leverages = table(colours_auto)[[2]])
num_leverage_auto

plot(h_auto, D_auto, pch = 21, cex = 0.6,
     xlab = "Leverage", ylab = "Cook's Influence Measure", bg = colours_auto,
     main = "Stepwise selected model", cex.main = 0.8)
abline(v = 2 * hbar_auto, col = "grey", lty = 2) # 2 * average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21, cex = 0.7,
       pt.bg = c("green", "red"))


# leverage & influence measure
p_man <- length(coef(m_red_man))
n <- nobs(m_red_man)
hbar_man <- p_man/n # average leverage
D_man <- cooks.distance(m_red_man) # cook's distance
influence_index_man <- which.max(D_man) # top influence point
leverage_index_man <- (h_man > (2*hbar_man)) # leverage more than 2x the average

# set colours in the plot
colours_man <- rep("black", length = n)
colours_man[leverage_index_man] <- "green"
colours_man[influence_index_man] <- "red"

# count the number of high leverage
num_leverage_man <- c(num_leverages = table(colours_man)[[2]])
num_leverage_man

plot(h_man, D_man, pch = 21, cex = 0.6, xlab = "Leverage",
     ylab = "Cook's Influence Measure", bg = colours_man,
     xlim=c(0,0.18), ylim=c(0, 0.03), main = "Manual constructed model",
     cex.main = 0.8)
abline(v = 2 * hbar_man, col = "grey", lty = 2) # 2 * average leverage
legend("topright", legend = c("High Leverage", "High Influence"), pch = 21,
       cex = 0.7, pt.bg = c("green", "red"))
```

All R code for section 5:

```r
# assess explanatory power using R^2_adj
c(R2adj_auto = round(summary(m_step)$adj.r.squared, 5),
  R2adj_man = round(summary(m_red_man)$adj.r.squared, 5))
```

```
## R2adj_auto  R2adj_man
##    0.82051    0.77953
```

```r
if(!params$load_calcs) {
  # assess predictive power by cross validation
  require(statmod)
```

```r
fhs_mspe <- fhs
# modify the fhs data to have logit column value given by
#   log(chdrisk) - log(1-chdrisk).
fhs_mspe$logit <- log(fhs_mspe$chdrisk) - log(1 - fhs_mspe$chdrisk)
# modify the fhs data to delete columns identified by VIF
fhs_mspe <- subset(fhs_mspe, select = -c(totchol,ldlc))
# setup auto and manual model
m_auto <- m_step
m_man <- m_red_man

# Cross-validation setup
nreplicate <- 1e3 # number of replicates
ntotal <- nrow(fhs_after_vif) # number of total data
ntrain <- 1906 # number of training data
ntest <- ntotal - ntrain # number of testing data
mspe_auto <- rep(NA, nreplicate)
mspe_man <- rep(NA, nreplicate)

# logitnorm_mean: calculate approximation for E(Gamma) where Gamma ~ N(mu, sigma^2)
# @param mu: mean of the normal destribution
# @param sigma: sd of the normal destribution
# @return expected value of Gamma where logit(Gamma) follows normal distribution
#         defined by mu and sigma
logitnorm_mean <- function(mu, sigma) {
  v <- 1/(1 + exp(-mu))
  a1 <- 1/(sigma^2 * (1-v))
  a2 <- 1/(v * sigma^2)
  gqp <- gauss.quad.prob(n = ntest, dist = "beta", alpha = a1, beta = a2)
         # note that may need to change 10 to actual Stest
  x <- gqp$nodes
  w <- gqp$weights
  g <- dnorm(log(x) - log(1 - x), mean = mu, sd = sigma, log = TRUE)
  g <- g - log(1 - x) - dbeta(x, shape1 = a1, shape2 = a2, log = TRUE)
  sum(w*exp(g))
}

# Check if logitmean is correct using 10 as number of tests data
#mu <- c(0.7, 3.2, -1.1)
#sigma <- c(.8, .1, 2.3)
# logitnorm_mean only accepts for one (mu, sigma) pair at a time
#sapply(1:3, function(ii) logitnorm_mean(mu[ii], sigma[ii]))

system.time({
  for (i in 1:nreplicate) {
    # randomly select ntrain data as training data
    train_i <- sample(ntotal, ntrain)
    # refit both model with training data only, and generate new fitted model
    ma_train <- update(m_auto, subset = train_i)
    mm_train <- update(m_man, subset = train_i)

    # generate testing data
    test_data <- fhs_mspe[-train_i,]

    # calculate mu for logit(chdrisk)
    ma_mus <- predict(ma_train, newdata = test_data)
    mm_mus <- predict(mm_train, newdata = test_data)
```

```
      # calculate sigma for logit(chadrisk)
      ma_sigma <- sqrt(sum((ma_train$residual)^2) / ntrain)
      mm_sigma <- sqrt(sum((mm_train$residual)^2) / ntrain)

      # calculate approximation for E(y|x=xi, beta_hat_train, sigma_hat_train)
      E_gamma_a <- sapply(1:ntest, function(i) logitnorm_mean(ma_mus[i], ma_sigma))
      E_gamma_m <- sapply(1:ntest, function(i) logitnorm_mean(mm_mus[i], mm_sigma))

      # calculate mean_square prediction error mspe
      mspe_auto[i] <- mean((test_data$chdrisk - E_gamma_a)^2)
      mspe_man[i] <- mean((test_data$chdrisk - E_gamma_m)^2)
    }
  })
  # names of the corvariates in both model
  Mnames <- expression(M[auto], M[man])
  saveRDS(list(ma = mspe_auto, mm = mspe_man, Mn = Mnames), file = "mspe_calc.rds")
} else {
  # just load the calculations
  tmp <- readRDS("mspe_calc.rds")
  mspe_auto <- tmp$ma
  mspe_man <- tmp$mm
  Mnames <- tmp$Mn
  rm(tmp) # optionally remove tmp from workspace
}
```

```
boxplot(x = list(sqrt(mspe_auto), sqrt(mspe_man)), names = Mnames, cex = .7,
        ylab = expression(sqrt(MSPE)), col = c("yellow", "orange"))
```

```
# Produce a compact table on the summary of the final model
ctable <- round(summary(m_step)[["coefficients"]]
                [, c("Estimate", "Std. Error", "Pr(>|t|)")], 5)
colnames(ctable)[3] <- "P-value"
ctable_left <- ctable[1:28, ]
ctable_right <- ctable[29:56, ]
kable(list(ctable_left, ctable_right), caption = "\\label{tab:final}
      Summary on parameter estimates, standard errors, and p-values of final model.",
        booktabs = T) %>%
      kable_styling(latex_option = "striped", font_size = 6.5, position = "center")
```

8 References

Paolo V., Gianpaolo R., Fabio A., Bruno T., Giuseppe M., Janice P., Peggy G., Peter S., Koon T., and Salim Y. (2014). Systolic and Diastolic Blood Pressure Changes in Relation With Myocardial Infarction and Stroke in Patients With Coronary Artery Disease. https://www.ahajournals.org/doi/10.1161/hypertensionaha.114.04310

Chirag B., Sangita G., Franz H. (2016). Isolated Systolic Hypertension: An Update After SPRINT. DOI:https://doi.org/10.1016/j.amjmed.2016.08.032