1. Summary

The purpose of this project is to use the knowlege we learnt in STAT 331: Applied Linear Models to explore the relation between the risk score for coronary heart disease (CHD) and some explanatary variables. Explanatary variables include continuous random variables such as blood pressure, body mass index, various cholesterol levels, as well as categorical variables such as age, sex, if individual is currently a smoker and etc.

In order to perform statistical analysis, we need to construct a model. Among the three automatic model selection methods: Forward Selection, Backward Elimination, and Stepwise Selection, we selected Stepwise Selection due to its high adjusted R square, which indicates valid contribution of each variable to the model. In addtion, we constructed another model manually and used several methods to compare with stepwise selection, including: residual plots, influence and leverage measures and predictive and explanatry power comparison. Stepwise Selection Model was selected as final model because of the comparison results.

To help lower the risk of CHD, it is recommended that individuals maintain healthy lifestyles and schedule routined body examinations. This is because many diseases and body malfunctions could trigger CHD as the analysis of model shows.

2. Descriptive Statistics

Of the 2306 individuals who participated in the Framingham Heart Study, their risk measures for coronary heart disease (CHD) vary from 0.005 to 0.977. 18 variables have been evaluated in terms of their relation to risk score for CHD, and data have been collected. From the summary of statistics, it is obvious that extremem values are present for all the continuous variables. Judging from the pair plots (1) of all the continuous variables, there exists a strong colinear relationship betwwen variables `totchol`, Serum total cholesterol level, and `ldlc`, low density lipoprotein cholesterol level. These two variables also have variance inflation factor (VIF) larger than 10, which is a cause for concern, and would be removed from models in the further study of this project.

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
```

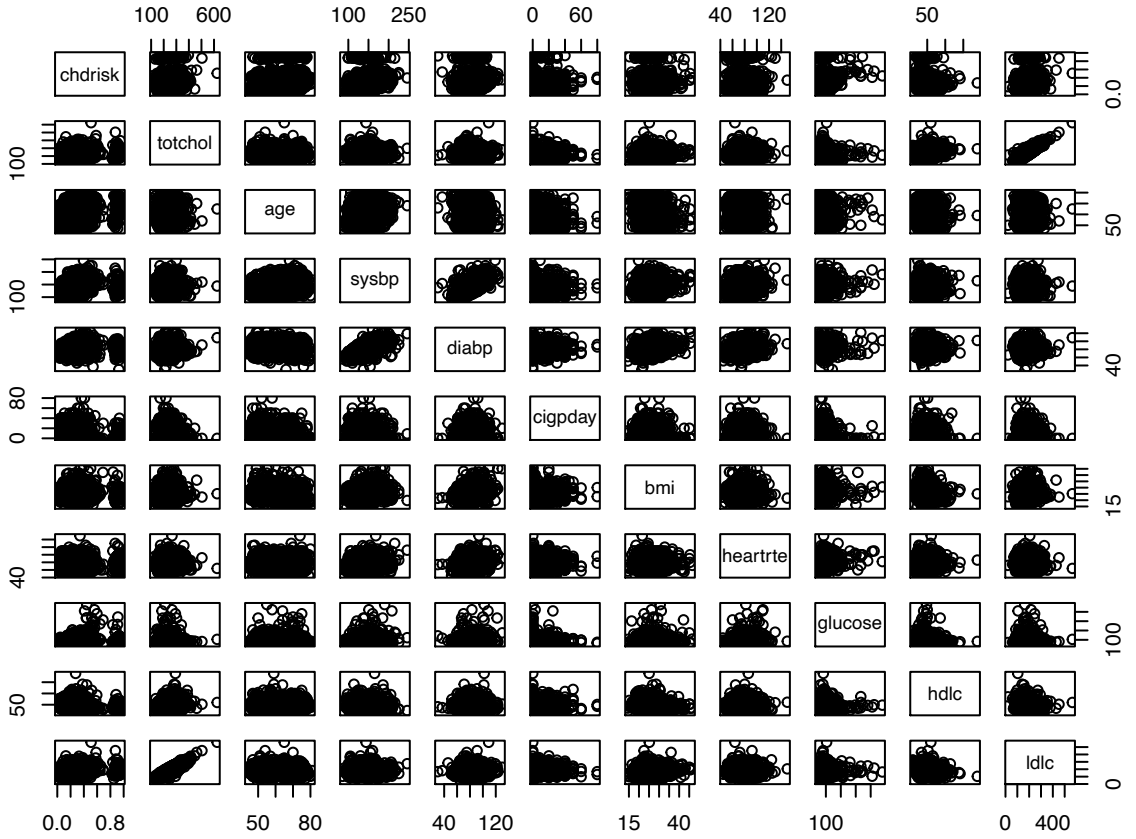

Figure 1: Pair Plot for Continious Variables

3. Candidate Model

3.1 Automatic Model Selection

Three automated model selection methods have been used: Forward Selection, Backward Elimination, and Stepwise Selection. To conduct the automated model selection process, a few parameters need to be constructed.

Response Variable is the variable whose variation depends on other variables. In this report, `logit(chdrisk)` was defined as the response variable, with definition: `logit(chdrisk)=log(chdrisk)-log(1-chdrisk)`. `chdrisk` refers to the risk measure for coronary heart disease (CHD).

Maximal, minimal and starting models need to be constructed to set an upper bound, lower bound, and starting position for the stepping function in R. These models are defined as below:

```
# Construct Maximal Model, Mininal Model, and Starting Model
Mmax <- lm(logit_chdrisk ~ (.-chdrisk)^2, data = fhs) # All main effects and interactions
M0 <- lm(logit_chdrisk ~ 1, data = fhs) # Mininal Model contains only intercept
# Starting Model for Stepwise selection, containing main effects only
Mstart <- lm(logit_chdrisk ~ .-chdrisk, data = fhs)
```

An efficient R built-in function, `step`, was implemented to conduct the selection processes. Due to the large amount of variables involved, stepping results are stored in three separate `rds` files for easy access.

System run time, parameter numbers, and adjusted R square are extracted from the three methods, to help select a better suited model:

- Forward seleciton: 0.6 seconds, 53 parameters, $R^2_{adj} = 0.8367$
- Backward elimination: 10.25 seconds, 66 parameters, $R^2_{adj} = 0.8413$
- Stepwise selection: 2.36 seconds, 58 parameters, $R^2_{adj} = 0.8407$.

Both backward and stepwise selection methods tend to provide better simulation because they start from a full model that contains all main effects and interactions. This is why forward selection has been removed from the scope, even though it has lowest runtime. Furthermore, since the increase in adjusted R square indicates the amount of improvement to the model each new parameter brings, the model generated by stepwise selection, `Mstep` is selected as first candidate model.

3.2 Manually Constructed Model

The manually constructed model, `Mmanual`, is a model constructed based on backward elimination model presented in section 3.1. According to what was discussed in the previous sections, backward elimination performs well in constructing the model. Even though the runtime is much longer than the other two, it also contains most variables. Thus if a new model is constructed based on `Mback`, there is a smaller possibility of missing any significant variables. In addition, removing insignificant main effects and interactions will shorten the runtime of model.

To start with, according to the pair plots (1) in section 2 and the calculation of VIF, the explanatary variable `totchol` and `ldlc` appear to have very strong linear relationship and both of their VIF are greater then 10. Thus, any explanatary variables related to `totchol` and `ldlc` in this step have been removed, and a new model `Mred` has been produced.

Next, interactions in `Mred` have been analyzed. In `Mback`, there are many interations. These interactions can be divided into two parts. The first part contains interations between a random variable and a categorical variable, while the other part contains interations between a categorical variable and another categorical variable. In comparison, the second part is less meaningful because interaction effects between two categorical variables could be either 0 or 1, wich would not contribute very meaningful information to the model itself. Thus we removed these parts and a new model `Mred1` has been generated.

Now, by looking at the summary of `Mred1`, it is clear that some interactions between two random variables have significantly large p-values. So in this step, interactions with large p-values, which are less significant, would be removed. After this, `Mred2` model has been constructed, and an F-test has been performed between `Mred2` and `Mred1`. The p-value of F: $0.1451 > 0.05$ proves that `Mred2` is better.

Above step was repeated three more times, until `Mred5` was generated. The summary of this final model shows that all explanatory variables are significant now due to small enough p-values. Therefore, `Mred5` was selected as the manual constructed model, with details displayed below:

```
Mred5 <- lm(formula = logit_chdrisk ~ sex + age +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevhyp +
    sysbp:diabp + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:glucose + diabp:hdlc + cigpday:hdlc +
    bmi:prevmi, data = fhs)
```

4. Model Diagnostic

4.1 Residual Plots

The Studentized residuals are used for model residual comparison because when the model assumptions are correct, the corresponding histogram would be closest to normal distribution, which is easy to identify through graphs.

In Residuals against Predicted Values for logitnorm Figure 2, residuals against predicted values for both `Mstep` and `Mmanual` are plotted, using both the usual residuals and corresponding-units version of the studentized residuals. `Mmanual` is more scattered than `Mstep`, with distinct difference between studentized residuals and the usual residuals. There seems to be some departures from the homoscedasticity assumption, because both residuals are slightly biased towards negative values. To further compare the models, histograms and QQ-plots of studentized residuals vs. theoretical normal distribution are plotted.

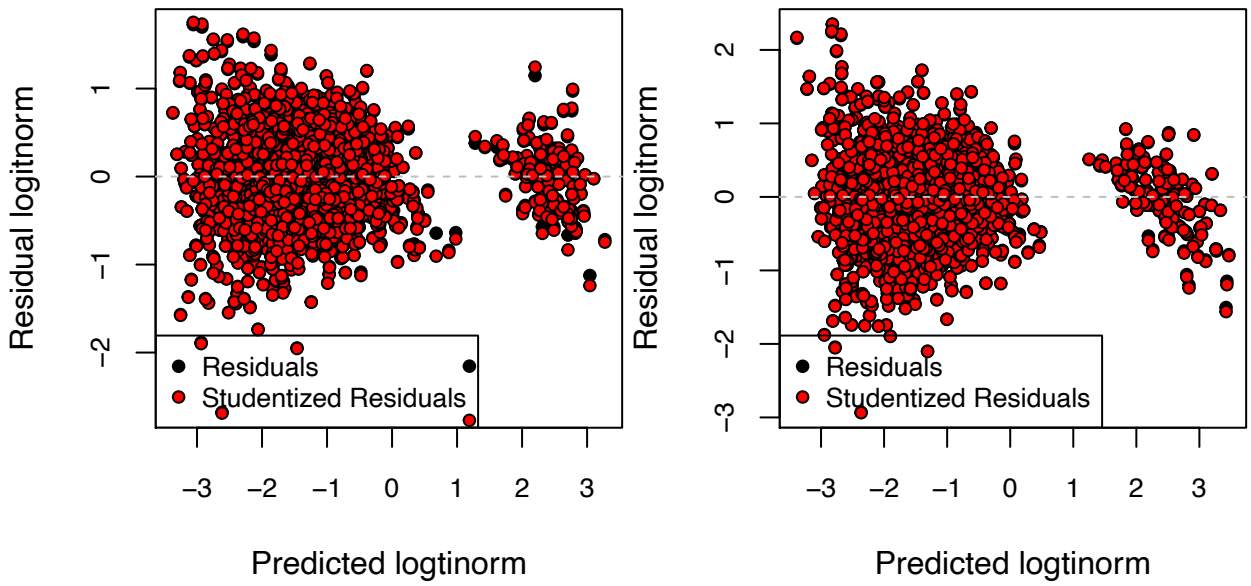

Figure 2: Residuals against Predicted Values for logitnorm

In Histogram and QQ-plot of studentized residuals with theoretical normal distribution Figure 3 and 4, the histograms show that both `Mstep` and `Mmanual` satisfy the assumption of normality, thus indicating both models might be correct. However, QQ-plots do not fully support this assumption, because even though most points lie on the 45 degree line, there are still some away from the line on both tails. `Mstep` has two points very far off the line with sample quantiles close to -6, while `Mmanual` only has one. These deviated points can be caused by outliers, which should be removed when designing the models. So with these points

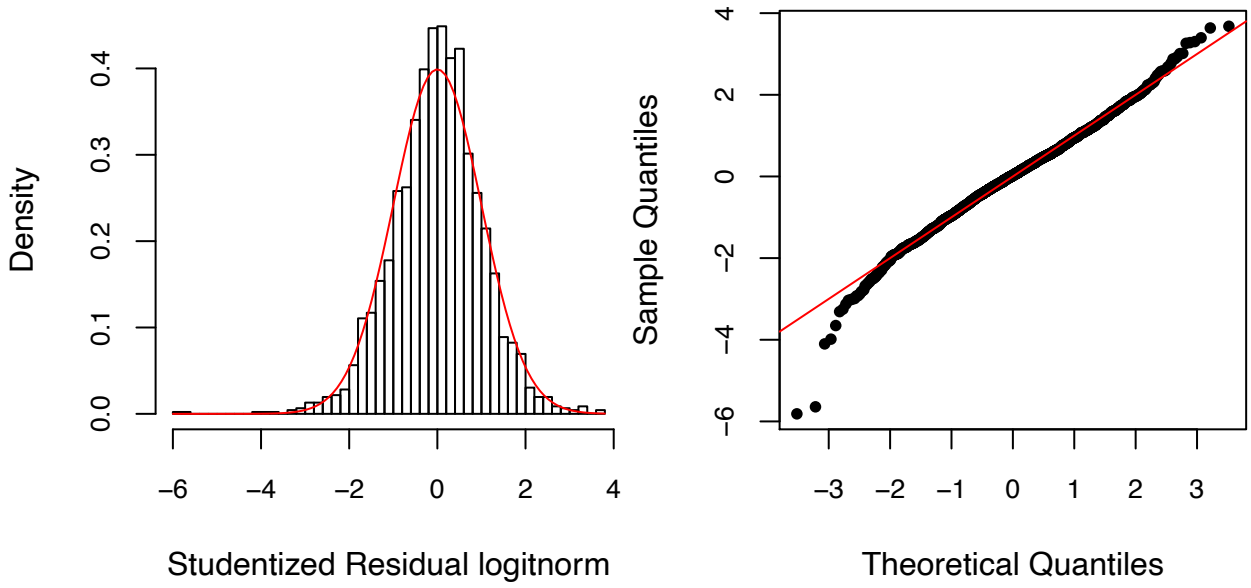

Figure 3: Histogram and QQ-plot1 of studentized residuals with theoretical normal distribution

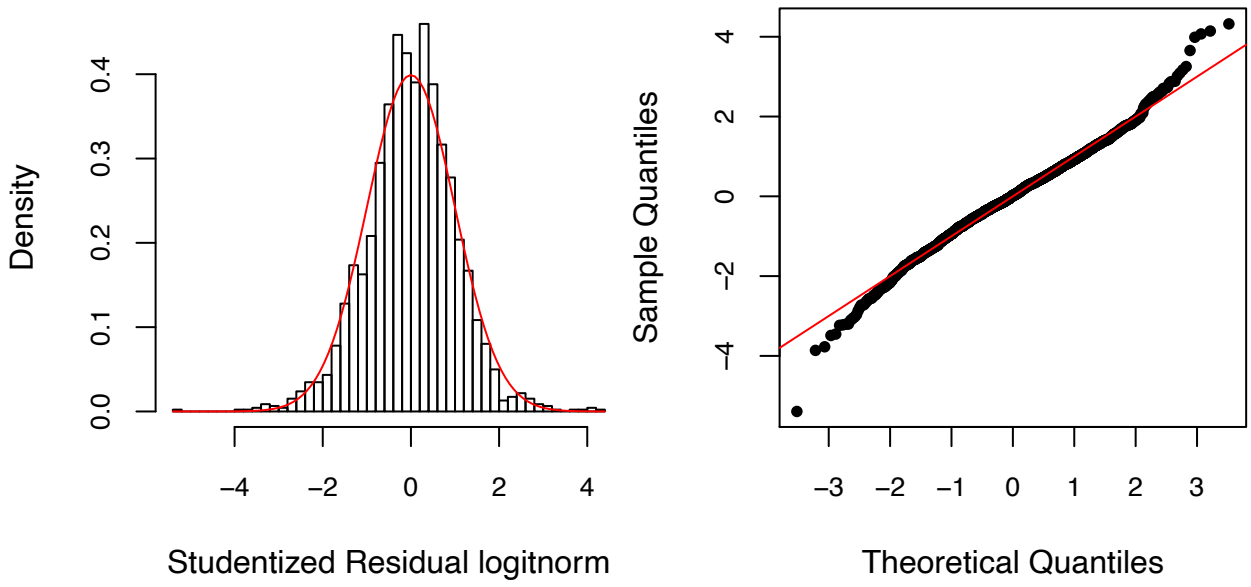

Figure 4: Histogram and QQ-plot2 of studentized residuals with theoretical normal distribution

scratched off, it is clear that `Mstep` has points closer to the line compared to `Mmanual`, meaning `Mstep` indeed satisfies the assumption of normality better.

4.2 Leverage and Influence Measures

From Cook's distance against leverage Figure 7, one of the observations has more than 3 times the Cook's distance of the others (in red) and several observations have 1 to 2 times the average leverage (in blue). For stepwise selection, both figures show that all observations are closer aligned along the 0 line, while for manaul selection model, they are more scatterd, especially in Cook's distance against leverage figure. Maunal selection model appears to have higher influence observations on the other hand. Based on the figures, stepwise selection is better than manual selection model because it has smaller Cook's distance and would be less influenced by outliers.

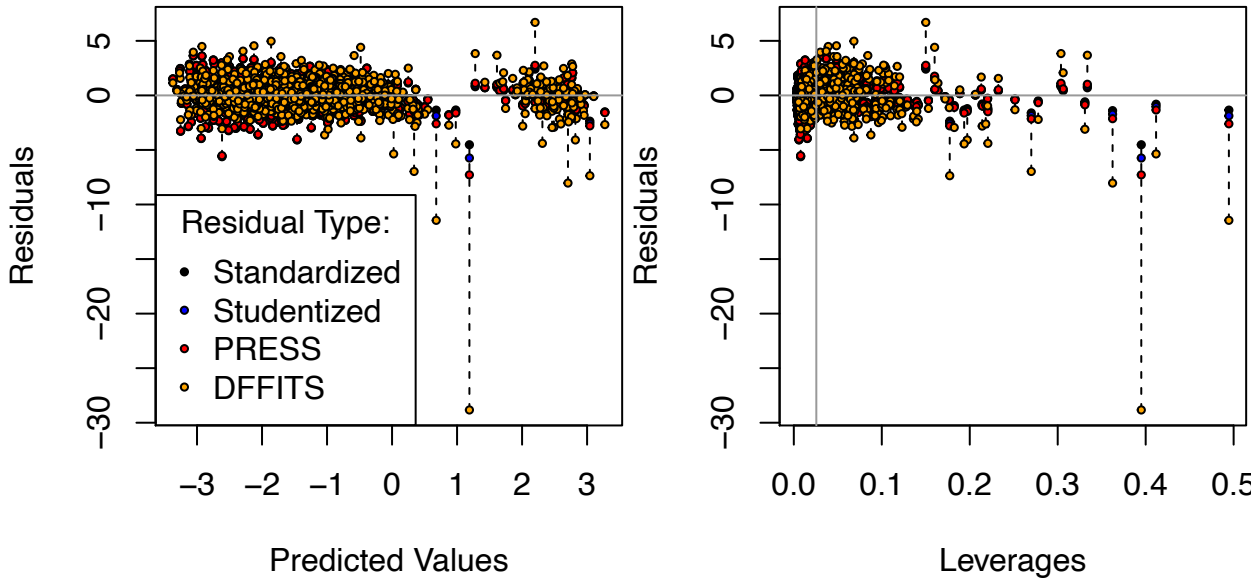

Figure 5: Residuals against fitted values for stepwise selection

5. Model Selection

5.1 Predictive Cross-Validation Analysis

Predictive power can be assesed by a cross-validation analysis. From Figure 8, the boxplot for root mean prediction error (rPMSE), it shows that median of stepwise selection is slightly less than the median of manual model. Model with smaller value of median in rPMSE is a better model. Average Abs. CV Residual figure shows similar results, therefore, `Mstep` performs better than `Mmanual`.

5.2 Explanatory power analysis

According to the previous sections, for stepwise selection model, $R^2 = 0.8447$, $R^2_{adj} = 0.8407$, while for manual model, $R^2 = 0.794$, $R^2_{adj} = 0.792$. Both R^2 and R^2_{adj} of `Mstep` are larger than that of `Mmanual`. Since R^2_{adj} measures the proportion of the total variatrion in response variable that is explaned by the

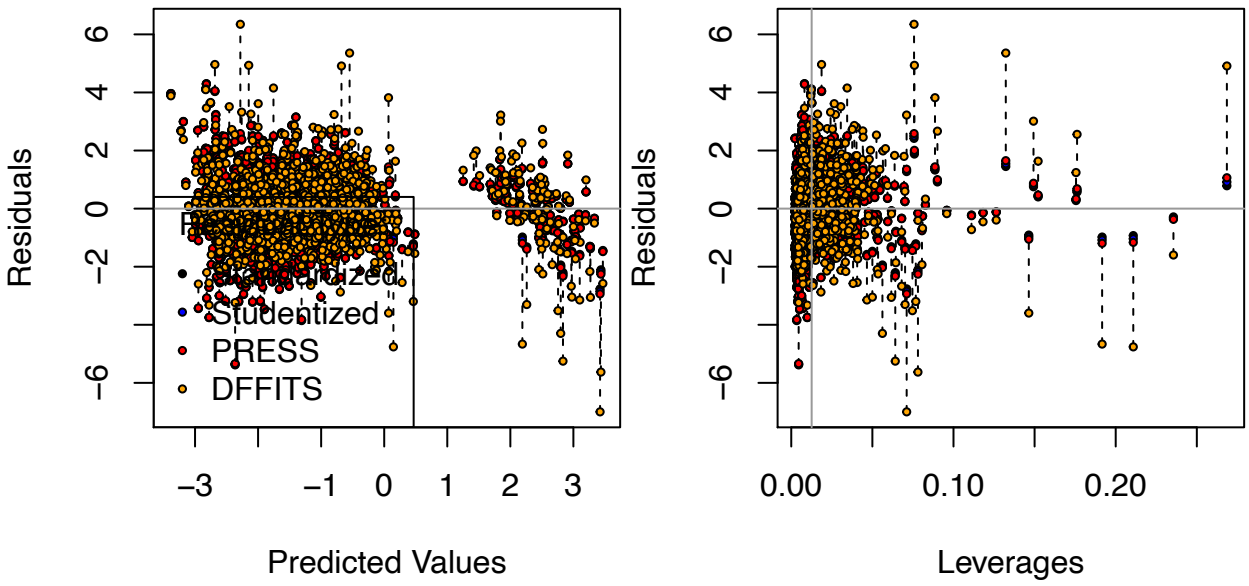

Figure 6: Residuals against fitted values for manual model

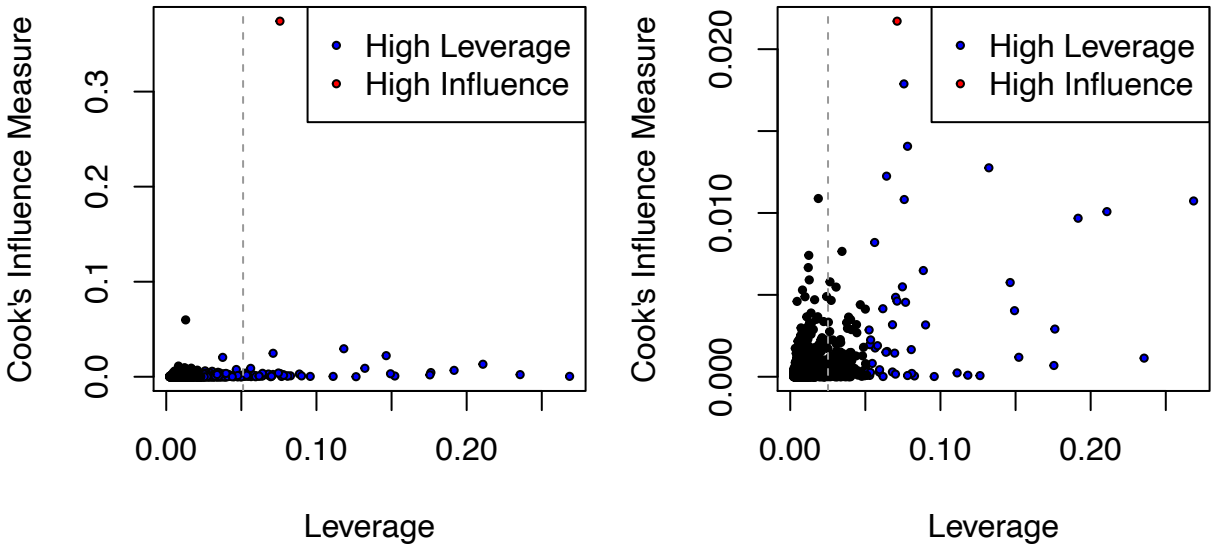

Figure 7: Cook's distance against leverage

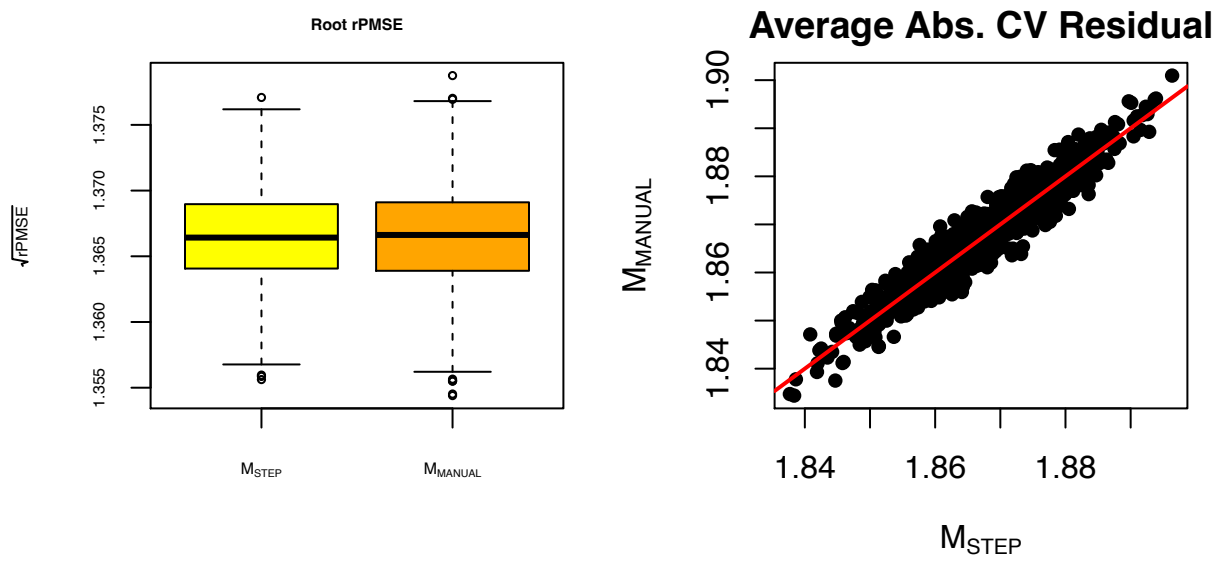

Figure 8: Cross-Validation model comparison

explanatry variable, the model with larger R^2 and R^2_{adj} has better explanatory power. Hence, `Mstep` should have better explanatory power.

5.3 Final Model

Based on the figures and analysis above, stepwise selection model `Mstep` is in general a better model compared to `Mmanual`. Its parameter estimates, standard errors and p-values are listed on next page.

6. Discussion

Based on our analysis, ther are several important factors associated to high CHD risk: `sexMale`, `diabeteYes`, `cursmokeYes`,`prevmiYes`, `prevhypYes`, `prevstrkYes` and `age`. There are also several important factors associated to low CHD risk: `diabp`, `hdlc`, `diabetesYes:prevmiYes`, `prevmiYes:prevhypYes`, `diabp:prevhypYes`, `age:prevmiYes`, `bmi:prevmiYes`, `prevmiYes:prevstrkYes`, `bmi:bpmedsYes` and `cursmokeYes:bpmedsYes`.

From Figure3 and Figure7, it is observable that there are still outliers exist in the model. In Figure3, two observations fell under the bottom left of the QQ-plot, and these two observations are outlying observations and thus might be appropriate to remove. In Figure7, there is one observation which has very high influence and its Cook's Influence Measure is more than 3 times of others. This observation is also an outlying observation which might be appropriate to remove as well.

In order to change individuals' behaviour to obtain lower risk of CHD, it is important that individuals control their diet to lower the amount of sugar they consume, and furtherly prevent diabetes. Individuals might also need to exercise regularly to keep their BMI in good condition. This greatly helps to lower the probability of having a stroke. Last but not least, routined body exams are crutial in lowering the risk of CHD, because many other diseases have a high possibility of triggering CHD.

Table 1: Summary of Parameter Estimates, Std Errors, and P-Value of Mstep

	Estimate	Std. Error	P-value		Estimate	Std. Error	P-value
(Intercept)	-7.7159973	1.0502002	0.0000000	bmi:ldlc	0.0002043	0.0000591	0.0005547
sexMale	0.8495120	0.1951078	0.0000140	prevmiYes:prevhypYes	-0.3685784	0.1336110	0.0058521
totchol	-0.0017943	0.0022109	0.4171143	sysbp:heartrte	-0.0001140	0.0000380	0.0027645
age	0.0759942	0.0125465	0.0000000	sexMale:glucose	-0.0020598	0.0007073	0.0036234
sysbp	0.0035876	0.0050279	0.4755918	age:cigpday	-0.0002890	0.0001294	0.0256089
diabp	-0.0166479	0.0101176	0.1000176	prevmiYes:hdlc	0.0111742	0.0038192	0.0034701
cursmokeYes	0.1788552	0.2133895	0.4020282	sysbp:hdlc	-0.0001075	0.0000344	0.0018156
cigpday	0.0402971	0.0091029	0.0000100	age:ldlc	0.0000603	0.0000288	0.0364602
bmi	-0.0216540	0.0110878	0.0509482	sexMale:sysbp	-0.0036803	0.0012456	0.0031637
diabetesYes	1.1549479	0.2852880	0.0000533	prevmiYes:ldlc	-0.0028631	0.0009315	0.0021387
bpmedsYes	0.7494415	0.3267862	0.0219187	age:heartrte	-0.0002506	0.0001011	0.0132792
heartrte	0.0494807	0.0078309	0.0000000	sysbp:bpmedsYes	-0.0057522	0.0017772	0.0012271
glucose	-0.0022810	0.0028608	0.4253406	sysbp:cursmokeYes	-0.0015526	0.0010986	0.1577297
prevmiYes	6.2665793	0.6165701	0.0000000	age:glucose	0.0000807	0.0000443	0.0685234
prevstrkYes	0.1829019	0.0805467	0.0232559	diabp:prevhypYes	-0.0103044	0.0037522	0.0060768
prevhypYes	3.6486727	0.4043154	0.0000000	age:prevmiYes	-0.0116205	0.0065464	0.0760193
hdlc	-0.0275879	0.0058037	0.0000021	bmi:prevmiYes	-0.0211912	0.0124283	0.0883185
ldlc	-0.0005422	0.0029155	0.8524786	diabetesYes:hdlc	0.0057545	0.0024473	0.0187903
sysbp:prevmiYes	-0.0086764	0.0026147	0.0009200	cigpday:hdlc	-0.0003008	0.0000957	0.0016993
age:diabp	-0.0004557	0.0001295	0.0004440	cursmokeYes:hdlc	0.0053438	0.0023704	0.0242682
totchol:prevhypYes	-0.0061712	0.0012267	0.0000005	age:prevhypYes	-0.0068353	0.0033864	0.0436631
totchol:hdlc	0.0003009	0.0000205	0.0000000	cursmokeYes:ldlc	-0.0007778	0.0004990	0.1191750
hdlc:ldlc	-0.0002441	0.0000188	0.0000000	sexMale:totchol	0.0010652	0.0005156	0.0389336
diabetesYes:prevmiYes	-0.6662294	0.1349383	0.0000009	prevmiYes:prevstrkYes	-0.3239239	0.1985547	0.1029442
prevhypYes:ldlc	0.0030046	0.0011826	0.0111323	cigpday:glucose	-0.0000667	0.0000379	0.0786037
sysbp:diabetesYes	-0.0067797	0.0016547	0.0000433	diabp:bpmedsYes	0.0071241	0.0033199	0.0319879
totchol:heartrte	-0.0000672	0.0000191	0.0004286	totchol:ldlc	0.0000043	0.0000025	0.0876705
sysbp:prevhypYes	-0.0089529	0.0020562	0.0000140	bmi:bpmedsYes	-0.0133370	0.0071772	0.0632654
sysbp:diabp	0.0003543	0.0000527	0.0000000	sexMale:prevhypYes	0.0849749	0.0563287	0.1315535

Appendix: R Code

```
library(knitr)
library(kableExtra)

fhs <- read.csv("fhs.csv") # Import data from local drive

## -----------------------------
## 2. Descriptive Statistics
## -----------------------------

# Summary Statistics
summary(fhs)

# Create pair plots (for continuous variables only)
pairs(~chdrisk+totchol+age+sysbp+diabp+cigpday+bmi+heartrte+
        glucose+hdlc+ldlc,data=fhs)

# Calculate VIF for explanatory variables: We will calculate VIFs using R^2
#'@param VarName Name of covariate for which to calculate R^2
#'@return The VIF for that covariate
VIF.R2 <- function(VarName){
  regform <- formula(paste0(VarName, " ~ totchol+age+sysbp+diabp+cigpday+bmi+
                                        heartrte+glucose+hdlc+ldlc"))
  FitModel <- lm(regform, data = fhs)
  R2 <- summary(FitModel)$r.square # Extract R2
  1/(1-R2) #Calculate VIF for the covariate
}
# Apply funtion to each of the covariates
# Design matrix excluding intercept
ModelMatrix <- model.matrix(lm(chdrisk ~
               totchol+age+sysbp+diabp+cigpday+bmi+heartrte+glucose+hdlc+ldlc-1,data = fhs))
VIF <- sapply(colnames(ModelMatrix), VIF.R2)

# Method 2: We will caculate VIFs using correlation matrix
#VIF2 <- diag(solve(cor(ModelMatrix)))
#VIF2

## -----------------------------
## 3. Candidate Models
## -----------------------------

# Construct response variable
chdrisk = fhs$chdrisk
logit_chdrisk = log(chdrisk) - log(1-chdrisk)

# Construct Maximal Model, Mininal Model, and Starting Model
Mmax <- lm(logit_chdrisk ~ (.-chdrisk)^2, data = fhs) # All main effects and interactions
M0 <- lm(logit_chdrisk ~ 1, data = fhs) # Mininal Model contains only intercept
Mstart <- lm(logit_chdrisk ~ .-chdrisk, data = fhs) # Starting Model for Stepwise selection, containing

# Now we want to remove all NA's and construct a simplified maximal model, Mmax1
beta.max <- coef(Mmax)
```

```r
length(beta.max) # Number of coefficients
names(beta.max)[is.na(beta.max)] # Coefficients that couldn't be estimated
table(fhs[c("cursmoke","cigpday")]) # No interaction between these
table(fhs[c("bpmeds","prevhyp")]) # No interaction between these

# Remove the interactions betwwen cursmoke & cigpday, bpmeds & prevhyp
Mmax1 <- lm(logit_chdrisk ~ (.-chdrisk)^2 - cursmoke:cigpday - bpmeds:prevhyp, data = fhs)
anyNA(coef(Mmax1)) # Result is FALSE, now Mmax1 doesn't contain any NA's

## Automated Model Selection
##============================

# Forward Selection
if(!params$load_calcs) {
  Mfwd <- step(object = M0, # Starting at Mininal model
               scope = list(lower = M0, upper = Mmax1), # Minimal and Maximal model
               direction = 'forward',
               trace = FALSE)
   saveRDS(list(Mfwd=Mfwd), file = "3_1_Mfwd.rds")
} else {
  # just load the calculations
  tmp <- readRDS("3_1_Mfwd.rds")
  Mfwd <- tmp$Mfwd
  rm(tmp) # optionally remove tmp from workspace
}
system.time(Mfwd)
summary(Mfwd)

# Backward Selection
if(!params$load_calcs) {
 Mback <- step(object = Mmax1, # Starting at Maximal model
               scope = list(lower = M0, upper = Mmax1), # Minimal and Maximal model
               direction = 'backward',
               trace = FALSE)
   saveRDS(list(Mback=Mback), file = "3_2_Mback.rds")
} else {
  # just load the calculations
  tmp <- readRDS("3_2_Mback.rds")
  Mback <- tmp$Mback
  rm(tmp) # optionally remove tmp from workspace
}
system.time(Mback)
summary(Mback)

# Stepwise Selection
if(!params$load_calcs) {
 Mstep <- step(object = Mstart, # Starting at Starting model
               scope = list(lower = M0, upper = Mmax1), # Minimal and Maximal model
               direction = 'both',
               trace = FALSE)
   saveRDS(list(Mstep=Mstep), file = "3_3_Mstep.rds")
} else {
  # just load the calculations
```

```
  tmp <- readRDS("3_3_Mstep.rds")
  Mstep <- tmp$Mstep
  rm(tmp) # optionally remove tmp from workspace
}
system.time(Mstep)
summary(Mstep)

## Manual Model Selection
##============================

# from pair plot: totchol & ldlc has very strong linear relationship
# vif>10: totchol & ldlc
Mred <- lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + hdlc +
    sex:sysbp + sex:glucose + sex:prevstrk + sex:prevhyp +
    age:diabp + age:cursmoke + age:heartrte + age:prevmi + age:prevhyp +
    age:hdlc + sysbp:diabp + sysbp:cigpday + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:cigpday + diabp:bpmeds + diabp:glucose +
    diabp:prevhyp + diabp:hdlc + cursmoke:bmi + cursmoke:hdlc +
    cigpday:bmi + cigpday:glucose + cigpday:hdlc +
    bmi:bpmeds + bmi:prevmi + diabetes:prevmi + diabetes:hdlc +
    bpmeds:glucose + bpmeds:prevstrk + heartrte:glucose + prevmi:prevhyp +
    prevmi:hdlc, data = fhs)
# remove interactions which are two c.v.:
Mred1 <- lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + hdlc +
    sex:sysbp + sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevmi + age:prevhyp +
    age:hdlc + sysbp:diabp + sysbp:cigpday + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:cigpday + diabp:bpmeds + diabp:glucose +
    diabp:prevhyp + diabp:hdlc + cursmoke:bmi + cursmoke:hdlc +
    cigpday:bmi + cigpday:glucose + cigpday:hdlc +
    bmi:bpmeds + bmi:prevmi + diabetes:hdlc +
    bpmeds:glucose + heartrte:glucose +
    prevmi:hdlc, data = fhs)
summary(Mred1)
# remove large pvalue b/t two r.v
Mred2 <- lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + hdlc +
    sex:sysbp + sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevmi + age:prevhyp +
    age:hdlc + sysbp:diabp + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:bpmeds + diabp:glucose +
    diabp:prevhyp + diabp:hdlc + cursmoke:bmi + cursmoke:hdlc +
    cigpday:bmi + cigpday:glucose + cigpday:hdlc +
    bmi:bpmeds + bmi:prevmi + diabetes:hdlc +
    bpmeds:glucose + heartrte:glucose +
```

```
    prevmi:hdlc, data = fhs)
summary(Mred2)
anova(Mred2,Mred1)
# remove large pvalue b/t r.v & c.v
Mred3 <- lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + hdlc +
    sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevhyp +
    age:hdlc + sysbp:diabp + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:glucose +
    diabp:hdlc +
    cigpday:bmi + cigpday:glucose + cigpday:hdlc +
    bmi:prevmi + heartrte:glucose, data = fhs)
summary(Mred3)
anova(Mred3,Mred2)
# remove large pvalue in interaction
Mred4 <- lm(formula = logit_chdrisk ~ sex + age + sysbp + diabp +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + hdlc +
    sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevhyp +
    sysbp:diabp + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:glucose +
    diabp:hdlc + cigpday:hdlc +
    bmi:prevmi, data = fhs)
summary(Mred4)
anova(Mred4,Mred3)
# remove large pvalue in cv and rv
Mred5 <- lm(formula = logit_chdrisk ~ sex + age +
    cursmoke + cigpday + bmi + diabetes + bpmeds + heartrte +
    glucose + prevmi + prevstrk + prevhyp + sex:glucose +
    age:diabp + age:cursmoke + age:heartrte + age:prevhyp +
    sysbp:diabp + sysbp:diabetes +
    sysbp:bpmeds + sysbp:heartrte + sysbp:prevmi + sysbp:prevhyp +
    diabp:cursmoke + diabp:glucose + diabp:hdlc + cigpday:hdlc +
    bmi:prevmi, data = fhs)
summary(Mred5)
anova(Mred5,Mred4)
Mmanual <- Mred5

## -----------------------------
## 4. Model Diagnostics
## -----------------------------

## Residual Plots
##============================

# Usual residuals of two candidate models
Res_Auto <- residuals(Mstep)
Res_Man <- residuals(Mred5)
```

```r
# Design matrix for two candidate models
X_Auto <- model.matrix(Mstep)
X_Man <- model.matrix(Mred5)
# Hat matrix
H_Auto <- X_Auto %*% solve(crossprod(X_Auto),t(X_Auto))
H_Man <- X_Man %*% solve(crossprod(X_Man),t(X_Man))
h_Auto <- diag(H_Auto)
h_Man <- diag(H_Man)
# Studentized residuals on data scale for two models
Res.stu_Auto <- resid(Mstep)/sqrt(1-h_Auto)
Res.stu_Man <- resid(Mred5)/sqrt(1-h_Man)
# Plot Residuals against Predicted Values plots
cex <- .8 # controls the size of the points and labels
# For auto
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
plot(predict(Mstep), Res_Auto, pch = 21, bg = "black", cex = cex, cex.axis = cex,
     xlab = "Predicted logtinorm", ylab = "Residual logitnorm")
points(predict(Mstep), Res.stu_Auto, pch = 21, bg = "red", cex = cex)
abline(h = 0, lty = 2, col = "grey") # add horizontal line at 0
legend(x = "bottomleft", c("Residuals", "Studentized Residuals"),
       pch = 21, pt.bg = c("black", "red"), pt.cex = cex, cex = cex)
# For Manual
plot(predict(Mred5), Res_Man, pch = 21, bg = "black", cex = cex, cex.axis = cex,
     xlab = "Predicted logtinorm", ylab = "Residual logitnorm")
points(predict(Mred5), Res.stu_Man, pch = 21, bg = "red", cex = cex)
abline(h = 0, lty = 2, col = "grey") # add horizontal line at 0
legend(x = "bottomleft", c("Residuals", "Studentized Residuals"),
       pch = 21, pt.bg = c("black", "red"), pt.cex = cex, cex = cex)
# Plot Studentized Residuals
Sigma.hat_Auto <- sigma(Mstep)
Sigma.hat_Man <- sigma(Mred5)
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
# Auto model
# histogram
hist(Res.stu_Auto/Sigma.hat_Auto, breaks = 50, freq = FALSE, cex.axis = cex,
     xlab = "Studentized Residual logitnorm", main = "")
curve(dnorm(x), col = "red", add = TRUE) # theoretical normal curve
# qq-plot
qqnorm(Res.stu_Auto/Sigma.hat_Auto, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line
# Manual model
# histogram
hist(Res.stu_Man/Sigma.hat_Man, breaks = 50, freq = FALSE, cex.axis = cex,
     xlab = "Studentized Residual logitnorm", main = "")
curve(dnorm(x), col = "red", add = TRUE) # theoretical normal curve
# qq-plot
qqnorm(Res.stu_Man/Sigma.hat_Man, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line

## Leverage and Influential
##============================

# for Mstep
```

```r
y.hat <- predict(Mstep) # predicted values
sigma.hat <- sigma(Mstep)
res <- resid(Mstep) # original residuals
stan.res <- res/sigma.hat # standardized residuals
# compute leverages
X <- model.matrix(Mstep)
H <- X %*% solve(crossprod(X), t(X)) # HAT matrix
head(diag(H))
h <- hatvalues(Mstep) # the R way
range(h - diag(H))
# studentized residuals (response scale)
stud.res <- stan.res/sqrt(1-h)
# PRESS residuals: response minus
press <- res/(1-h)
# DFFITS residuals
dfts <- dffits(Mstep) # the R way
# standardize each of these such that they are identical at the average leverage value
p <- length(coef(Mstep))
n <- nobs(Mstep)
hbar <- p/n # average leverage
stud.res <- stud.res*sqrt(1-hbar) # at h = hbar, stud.res = stan.res
press <- press*(1-hbar)/sigma.hat # at h = hbar, press = stan.res
dfts <- dfts*(1-hbar)/sqrt(hbar) # at hbar dfts = stan.res
# plot all residuals
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
cex <- .5
# against predicted values
plot(y.hat, rep(0, length(y.hat)), type = "n", # empty plot to get the axis range
     ylim = range(stan.res, stud.res, press, dfts),
     xlab = "Predicted Values", ylab = "Residuals")
# dotted line connecting each observations residuals for better visibility
segments(x0 = y.hat,
         y0 = pmin(stan.res, stud.res, press, dfts),
         y1 = pmax(stan.res, stud.res, press, dfts),
         lty = 2)
points(y.hat, stan.res, pch = 21, bg = "black", cex = cex)
points(y.hat, stud.res, pch = 21, bg = "blue", cex = cex)
points(y.hat, press, pch = 21, bg = "red", cex = cex)
points(y.hat, dfts, pch = 21, bg = "orange", cex = cex)
abline(h = 0, col = "grey60")
legend("bottomleft",
       legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.cex = cex,
       pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:")
# against leverages
plot(h, rep(0, length(y.hat)), type = "n",
     ylim = range(stan.res, stud.res, press, dfts),
     xlab = "Leverages", ylab = "Residuals")
segments(x0 = h,
         y0 = pmin(stan.res, stud.res, press, dfts),
         y1 = pmax(stan.res, stud.res, press, dfts),
         lty = 2)
```

```r
points(h, stan.res, pch = 21, bg = "black", cex = cex)
points(h, stud.res, pch = 21, bg = "blue", cex = cex)
points(h, press, pch = 21, bg = "red", cex = cex)
points(h, dfts, pch = 21, bg = "orange", cex = cex)
abline(v = hbar, col = "grey60", lty = 1)
abline(h = 0, col = "grey60")
#for Mmanual
Mmanual <- Mred5
y.hat <- predict(Mmanual) # predicted values
sigma.hat <- sigma(Mmanual)
res <- resid(Mmanual) # original residuals
stan.res <- res/sigma.hat # standardized residuals
# compute leverages
X <- model.matrix(Mmanual)
H <- X %*% solve(crossprod(X), t(X)) # HAT matrix
head(diag(H))
h <- hatvalues(Mmanual) # the R way
range(h - diag(H))
# studentized residuals (response scale)
stud.res <- stan.res/sqrt(1-h)
# PRESS residuals: response minus
press <- res/(1-h)
# DFFITS residuals
dfts <- dffits(Mmanual) # the R way
# standardize each of these such that they are identical at the average leverage value
p <- length(coef(Mmanual))
n <- nobs(Mmanual)
hbar <- p/n # average leverage
stud.res <- stud.res*sqrt(1-hbar) # at h = hbar, stud.res = stan.res
press <- press*(1-hbar)/sigma.hat # at h = hbar, press = stan.res
dfts <- dfts*(1-hbar)/sqrt(hbar) # at hbar dfts = stan.res
# plot all residuals
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
cex <- .5
# against predicted values
plot(y.hat, rep(0, length(y.hat)), type = "n", # empty plot to get the axis range
     ylim = range(stan.res, stud.res, press, dfts),
     xlab = "Predicted Values", ylab = "Residuals")
# dotted line connecting each observations residuals for better visibility
segments(x0 = y.hat,
         y0 = pmin(stan.res, stud.res, press, dfts),
         y1 = pmax(stan.res, stud.res, press, dfts),
         lty = 2)
points(y.hat, stan.res, pch = 21, bg = "black", cex = cex)
points(y.hat, stud.res, pch = 21, bg = "blue", cex = cex)
points(y.hat, press, pch = 21, bg = "red", cex = cex)
points(y.hat, dfts, pch = 21, bg = "orange", cex = cex)
abline(h = 0, col = "grey60")
legend("bottomleft",
       legend = c("Standardized", "Studentized", "PRESS", "DFFITS"),
       pch = 21, pt.cex = cex,
       pt.bg = c("black", "blue", "red", "orange"),
       title = "Residual Type:")
```

```r
# against leverages
plot(h, rep(0, length(y.hat)), type = "n",
     ylim = range(stan.res, stud.res, press, dfts),
     xlab = "Leverages", ylab = "Residuals")
segments(x0 = h,
         y0 = pmin(stan.res, stud.res, press, dfts),
         y1 = pmax(stan.res, stud.res, press, dfts),
         lty = 2)
points(h, stan.res, pch = 21, bg = "black", cex = cex)
points(h, stud.res, pch = 21, bg = "blue", cex = cex)
points(h, press, pch = 21, bg = "red", cex = cex)
points(h, dfts, pch = 21, bg = "orange", cex = cex)
abline(v = hbar, col = "grey60", lty = 1)
abline(h = 0, col = "grey60")

# cook's distance vs. leverage

#for Mstep
cex <- .5
D <- cooks.distance(Mstep)
# flag some of the points
infl.ind <- rep(FALSE, n) # top influence point
infl.ind[which.max(D)] <- TRUE
lev.ind <- h > 2*hbar # leverage more than 2x the average
clrs <- rep("black", len = n)
clrs[lev.ind] <- "blue"
clrs[infl.ind] <- "red"
par(mfrow = c(1,2), mar = c(4,4,1,1))
plot(h, D, pch = 21, bg = clrs, cex = cex,
     xlab = "Leverage", ylab = "Cook's Influence Measure")
p <- length(coef(Mstep))
n <- nrow(fhs)
hbar <- p/n # average leverage
abline(v = 2*hbar, col = "grey60", lty = 2) # 2x average leverage
legend("topright", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), pt.cex = cex)
#for Mmanual
cex <- .5
D <- cooks.distance(Mmanual)
# flag some of the points
infl.ind <- rep(FALSE, n) # top influence point
infl.ind[which.max(D)] <- TRUE
lev.ind <- h > 2*hbar # leverage more than 2x the average
clrs <- rep("black", len = n)
clrs[lev.ind] <- "blue"
clrs[infl.ind] <- "red"
plot(h, D, pch = 21, bg = clrs, cex = cex,
     xlab = "Leverage", ylab = "Cook's Influence Measure")
p <- length(coef(Mmanual))
n <- nrow(fhs)
hbar <- p/n # average leverage
abline(v = 2*hbar, col = "grey60", lty = 2) # 2x average leverage
legend("topright", legend = c("High Leverage", "High Influence"), pch = 21,
```

```
       pt.bg = c("blue", "red"), pt.cex = cex)

## ----------------------------
## 5. Model Selection
## ----------------------------

## Cross Validation
##==========================

require(statmod)
M1 <- Mstep
M2 <- Mmanual
Mnames <- expression(M[STEP], M[MANUAL])
# number of cross-validation replications
nreps <- 1e3
ntot <- nrow(fhs) # total number of observations
ntrain <- 400 # for fitting MLE's
ntest <- ntot-ntrain # for out-of-sample prediction

# Calculate mean of the Logit-Normal Distribution
#'@param mu mean of data set: train
#'@param sigma standard deviation of data set: train
#'@return Value fo E[Gamma] for Gamma~N(mu,signma^2)
logitnorm_mean <- function(mu, sigma){

  v <- 1/(1+exp(-mu))
  a1 <- 1/(I(sigma^2)*(1-v))
  a2 <- 1/(v*I(sigma^2))

  gqp <- gauss.quad.prob(n = 400, dist = "beta", alpha = a1, beta = a2) # ntest=400
  x <- gqp$nodes #row vector with n elements
  w <- gqp$weights
  logit.x <- log(x)-log(1-x) # row vector of logit x for all x

  gx <- dnorm(logit.x, mean = mu, sd = sigma, log = TRUE)-log(1-x)-
    dbeta(x, shape1 = a1, shape2 = a2, log = TRUE)
  Ey <- w*exp(gx)
  sum(Ey)
}

# storage space
mspe1 <- rep(NA, nreps) # mspe for M1
mspe2 <- rep(NA, nreps) # mspe for M2
rPMSE1 <- rep(NA, nreps) # rPMSE for M1
rPMSE2 <- rep(NA, nreps) # rPMSE for M1

if(!params$load_calcs) {
    for(ii in 1:nreps) {
    if(ii%%100 == 0) message("ii = ", ii)
    train.ind <- sample(ntot, ntrain) # training observations
    # long-form cross-validation
    # using R functions
    M1.cv <- update(M1, subset = train.ind)
```

```
    M2.cv <- update(M2, subset = train.ind)
    # cross-validation residuals

    mu1 <- predict(M1.cv, newdata = fhs[-train.ind,]) # prediction with training data
    mu2 <- predict(M2.cv, newdata = fhs[-train.ind,])

    # out-of-sample log-likelihoods
    M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
    M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)

    M1.res <- logit_chdrisk[-train.ind] - # test observations
      sapply(1:400,
             function(ii) logitnorm_mean(mu1[ii], M1.sigma)) # prediction with training data

    M2.res <- logit_chdrisk[-train.ind] - # test observations
      sapply(1:400,
             function(ii) logitnorm_mean(mu2[ii], M2.sigma)) # prediction with training data

    #rPMSE for each model
    rPMSE1[ii] <- sqrt(mean(M1.res^2))
    rPMSE2[ii] <- sqrt(mean(M2.res^2))
  }
   saveRDS(list(rPMSE1=rPMSE1, rPMSE2=rPMSE2), file = "5_1_Logitnorm_Mean.rds")
} else {
  # just load the calculations
  tmp <- readRDS("5_1_Logitnorm_Mean.rds")
  rPMSE1 <- tmp$rPMSE1
  rPMSE2 <- tmp$rPMSE2
  rm(tmp) # optionally remove tmp from workspace
}
#rPMSE1
#rPMSE2

# boxplot
par(mfrow = c(1,2), mar = c(4, 4, 2, 1.5))
cex <- 0.5
boxplot(x = list(sqrt(rPMSE1), sqrt(rPMSE2)), names = Mnames,
        main = "Root rPMSE",
        ylab = expression(sqrt(rPMSE)),
        ## ylab = expression(SSE[CV]),
        col = c("yellow", "orange"),
        cex = cex, cex.lab = cex, cex.axis = cex, cex.main = cex)

# compare predictions by training set
plot(rPMSE1, rPMSE2, pch = 16,
     xlab = Mnames[1], ylab = Mnames[2],
     main = "Average Abs. CV Residual")
abline(a = 0, b = 1, col= "red", lwd = 2)

# Summary table for Mstep
ctable <- summary(Mstep)[["coefficients"]][,c("Estimate", "Std. Error", "Pr(>|t|)")]
colnames(ctable)[3] <- "P-value"
ctable_left <- ctable[1:29, ]
```

```r
ctable_right <- ctable[30:58,]
kable(list(ctable_left, ctable_right), caption = "Summary of Parameter Estimates, Std Errors, and P-Valu
      booktabs = T) %>%
  kable_styling(latex_option = "striped", font_size = 6, position = "center")
```

References

- Martin, L.(2020). STAT 331: Module II – Multiple Linear Regression: Theory
- Martin, L.(2020). STAT 331: Module III – Multiple Linear Regression: Case Studies
- Martin, L.(2020). STAT 331: Module IV – Model Diagnostics
- Martin, L.(2020). STAT 331: Module V – Model Selection
- Martin, L.(2020). stat331-model_selection.R
- Martin, L.(2020). stat331-interactions.R
- Martin, L.(2020). stat331-influence.R