

# STAT 331 Final Project

Krishna Prem Pasumathy & Islam Amin

April 15, 2020

## 1 Summary

## 2 Descriptive Statistics

First, take a look at summary statistics of the Framingham Heart Study dataset.

Table 1: Summary Statistics

chdrisk	sex	totchol	age	sysbp	diabp	cursmoke	cigpday	bmi
Min. :0.0050	Female:1305	Min. :112.0	Min. :44.00	Min. : 86.0	Min. : 30.00	No :1504	Min. : 0.00	Min. :14.43
1st Qu.:0.1320	Male :1001	1st Qu.:207.0	1st Qu.:53.00	1st Qu.:122.5	1st Qu.: 73.00	Yes: 802	1st Qu.: 0.00	1st Qu.:23.22
Median :0.2240		Median :235.5	Median :60.00	Median :136.0	Median : 80.00		Median : 0.00	Median :25.40
Mean :0.2655		Mean :237.8	Mean :60.23	Mean :139.2	Mean : 81.07		Mean : 6.84	Mean :25.78
3rd Qu.:0.3448		3rd Qu.:265.0	3rd Qu.:67.00	3rd Qu.:153.0	3rd Qu.: 88.00		3rd Qu.:10.00	3rd Qu.:27.91
Max. :0.9770		Max. :625.0	Max. :81.00	Max. :246.0	Max. :130.00		Max. :80.00	Max. :46.52

diabetes	bpmeds	hearttrte	glucose	prevmi	prevstrk	prevhyp	hdlc	ldlc
No :2142	No :1973	Min. : 44.00	Min. : 46.00	No :2189	No :2260	No : 957	Min. : 10.00	Min. : 20.0
Yes: 164	Yes: 333	1st Qu.: 70.00	1st Qu.: 75.00	Yes: 117	Yes: 46	Yes:1349	1st Qu.: 38.00	1st Qu.:152.0
		Median : 76.00	Median : 83.00				Median : 47.00	Median :180.0
		Mean : 77.61	Mean : 89.07				Mean : 48.89	Mean :183.1
		3rd Qu.: 85.00	3rd Qu.: 95.00				3rd Qu.: 57.00	3rd Qu.:210.0
		Max. :150.00	Max. :478.00				Max. :189.00	Max. :565.0

First observation we make from the summary is that the median and average ages are around 60, which means the survey seems to have been done on a relatively old group of people. We also have a significantly higher number of females in the study, almost 30% more than the number of males. This might affect the nature of the data to be skewed towards behaviours and physical attributes associated with females.

A further inspection of the expected coronary heart disease (CHD) risk against certain categorical variates, gives more insights.

For instance, if we take a look at expected CHD risk against whether or not an individual has hypertension, we get the following result:

```
## fhsd$prevhyp: No
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.005  0.077   0.140   0.176  0.216   0.944
## -----
## fhsd$prevhyp: Yes
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.0320  0.1980  0.2890  0.3291  0.4010  0.9770
```

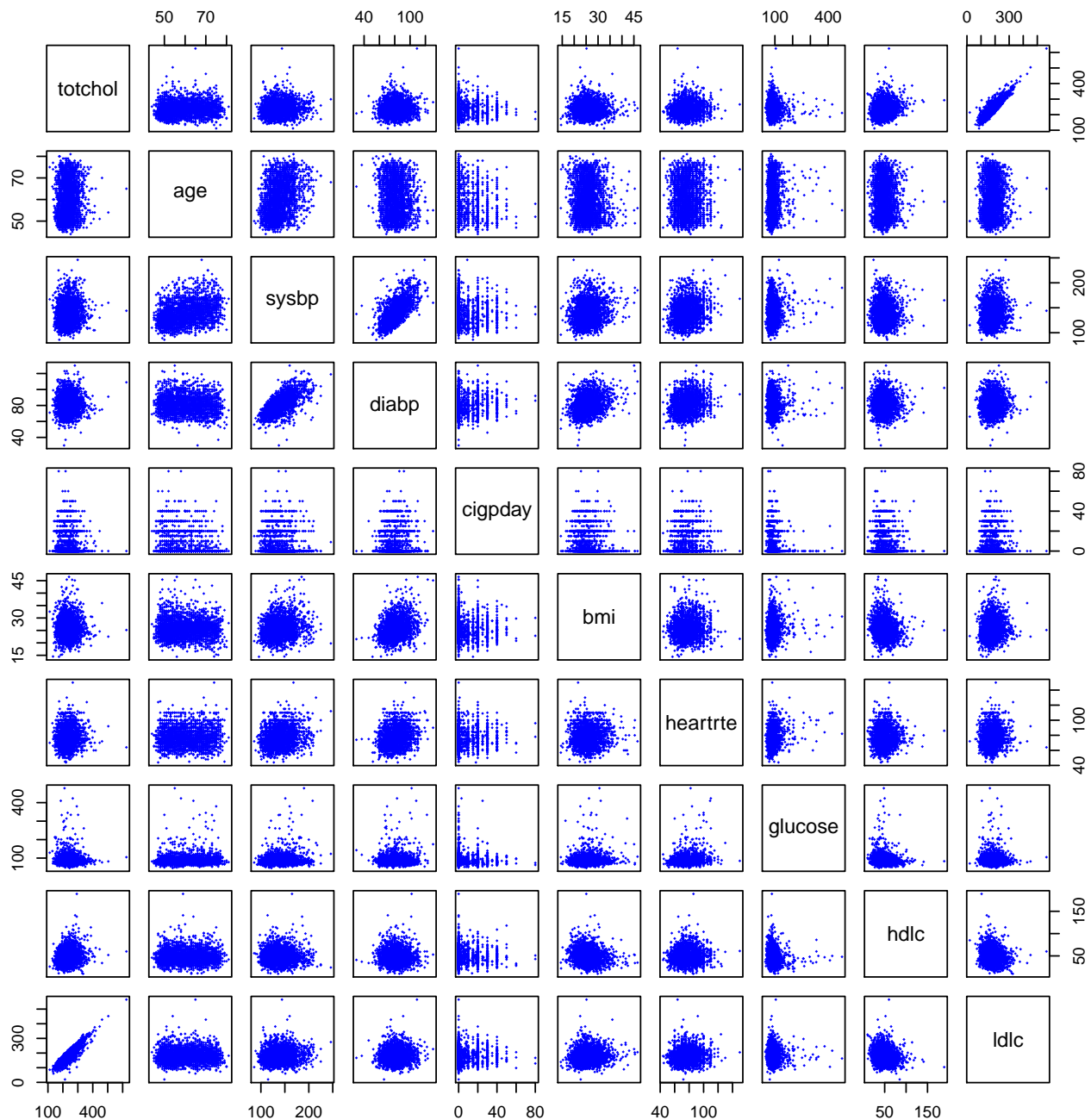
Indeed, we have that mean CHD risk given that a person has hypertension is significantly higher than the mean for people who did not have hypertension.

```
## fhds$prevstrk: No
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##  0.0050  0.1300  0.2200  0.2611  0.3392  0.9770
## -----
## fhds$prevstrk: Yes
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##  0.2020  0.3412  0.4410  0.4820  0.5060  0.9660
```

Again, we see the same results with people who had a stroke before the study, with even a higher difference between the two groups.

Now take a look at pair plots of all numeric explanatory variates i.e. variates excluding response variate `chdrisk` and logical variates such as `cursmoke`.

## Pair Plots of Continuous Variates



From the pair plots, we can observe a strong correlation between low density lipoprotein cholesterol and serum total cholesterol. This correlation could be explained by the fact that there could be a relationship between the amount [TO BE CONTINUED]

Now take a look at the VIFs of these variates.

##	sexMale	totchol	age	sysbp	diabp	cursmokeYes
##	1.225191	10.634882	1.489926	2.918660	2.406411	2.978609
##	cigpday	bmi	diabetesYes	bpmedsYes	hearttrte	glucose
##	2.973594	1.181865	1.286401	1.214744	1.105902	1.308923
##	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc	
##	1.067134	1.045746	1.823014	2.287571	10.367649	

[ADD COMMENTS]

## 3 Candidate Models

### 3.1 Automated Model Selection

```
suppressWarnings(library(gtools))
load_calcs = TRUE
# model with only intercept
M0 <- lm(I(logit(chdrisk)) ~ 1, data = fhds)
Mmax <- lm(I(logit(chdrisk)) ~ (. )^2, data = fhds)
# starting model for stepwise selection
Mstart <- lm(I(logit(chdrisk)) ~ ., data = fhds)
# find model coefficients which are NA
beta.max <- coef(Mmax)
names(beta.max)[is.na(beta.max)]
```

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

```
# find the problem with the NA coeffs
kable(table(fhds[c("cursmoke", "cigpday")]), "latex")
```

	0	1	2	3	4	5	6	7	8	9	10	12	14	15	16	17	18	19	20	23	25	26	27
No	1504	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

```
kable(table(fhds[c("bpmeds", "prevhyp")]), "latex")
```

	No	Yes
No	957	1016
Yes	0	333

```
# remove the coeffs with the problem and add quadratic terms for the continuous variables
Mmax <- lm(I(logit(chdrisk)) ~ (. )^2 - cursmoke:cigpday - bpmeds:prevhyp +
  I(totchol ^ 2) + I(sysbp ^ 2) + I(diabp ^ 2)
  + I(bmi ^ 2) + I(glucose ^ 2)
  + I(hdlc ^ 2) + I(ldlc ^ 2), data = fhds)
anyNA(coef(Mmax)) # check if there are any remaining NAs
```

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

```
if(!load_calcs){
  #forward model selection
  system.time({
    Mfwd <- step(object = M0,
      scope = list(lower = M0, upper = Mmax),
      direction = "forward", trace = FALSE)
  })

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

  #stepwise model selection
  system.time({
```

```

Mstep <- step(object = Mstart,
              scope = list(lower = M0, upper = Mmax),
              direction = "both", trace = FALSE)
})
}
# the caching/loading block
if(!load_calcs) {
  saveRDS(list(Mfwd = Mfwd, Mback = Mback, Mstep = Mstep), file = "models_automated.rds")
} else {
  # just load the calculations
  tmp <- readRDS("models_automated.rds")
  Mfwd <- tmp$Mfwd
  Mback <- tmp$Mback
  Mstep <- tmp$Mstep
  rm(tmp) # optionally remove tmp from workspace
}
# Stepwise model selection
Mstep$call

```

```

## lm(formula = I(logit(chdrisk)) ~ sex + totchol + age + sysbp +
##     diabp + cursmoke + cigpday + bmi + diabetes + bpmeds + heart rte +
##     glucose + prevmi + prevstrk + prevhyp + hdlc + ldlc + I(hdlc^2) +
##     I(bmi^2) + I(diabp^2) + I(sysbp^2) + sysbp:prevmi + totchol:prevhyp +
##     diabetes:prevmi + prevhyp:ldlc + sysbp:prevhyp + totchol:heart rte +
##     sysbp:diabetes + diabp:bmi + diabp:hdlc + prevmi:hdlc + prevmi:prevhyp +
##     sex:glucose + age:ldlc + age:heart rte + cigpday:hdlc + bmi:ldlc +
##     totchol:hdlc + totchol:prevmi + sysbp:heart rte + sysbp:bpmeds +
##     cursmoke:hdlc + prevmi:prevstrk + diabetes:hdlc + sex:sysbp +
##     cigpday:glucose + heart rte:glucose + diabp:glucose + cursmoke:ldlc +
##     age:cigpday + age:hdlc + hdlc:ldlc + age:prevhyp + diabp:prevhyp +
##     diabp:cursmoke + diabp:cigpday + bmi:bpmeds + bpmeds:glucose +
##     age:prevmi + sex:ldlc + cigpday:heart rte + cigpday:prevmi +
##     glucose:prevmi + heart rte:prevmi + bpmeds:prevstrk, data = fhds)

```

```

# Forward model selection
Mfwd$call

```

```

## lm(formula = I(logit(chdrisk)) ~ prevmi + sysbp + sex + age +
##     ldlc + prevhyp + diabetes + hdlc + I(hdlc^2) + cigpday +
##     I(bmi^2) + bmi + totchol + I(glucose^2) + I(sysbp^2) + bpmeds +
##     heart rte + cursmoke + prevstrk + prevmi:sysbp + sysbp:age +
##     prevhyp:hdlc + prevmi:diabetes + sysbp:prevhyp + prevhyp:totchol +
##     sysbp:diabetes + prevmi:hdlc + prevmi:prevhyp + age:ldlc +
##     age:cigpday + hdlc:cigpday + prevhyp:bmi + ldlc:bmi + prevmi:totchol +
##     ldlc:prevhyp + sysbp:bpmeds + sysbp:hdlc + hdlc:totchol +
##     totchol:heart rte + age:heart rte + diabetes:hdlc + sysbp:heart rte +
##     bmi:bpmeds + sysbp:sex + ldlc:hdlc + prevmi:bmi + age:bmi +
##     prevmi:age + sysbp:cursmoke + hdlc:cursmoke + ldlc:cursmoke +
##     prevmi:cigpday + sex:diabetes + prevmi:prevstrk, data = fhds)

```

```

# Backward model selection
Mback$call

```

```

## lm(formula = I(logit(chdrisk)) ~ sex + totchol + age + sysbp +
##     diabp + cursmoke + cigpday + bmi + diabetes + bpmeds + heart rte +

```

```
## glucose + prevmi + prevstrk + prevhyp + hdlc + ldlc + I(totchol^2) +
## I(sysbp^2) + I(diabp^2) + I(bmi^2) + I(hdlc^2) + I(ldlc^2) +
## sex:totchol + sex:sysbp + sex:glucose + sex:prevstrk + sex:prevhyp +
## totchol:age + totchol:bpmeds + totchol:hearttrte + totchol:prevmi +
## totchol:prevstrk + totchol:prevhyp + totchol:hdlc + totchol:ldlc +
## age:cursmoke + age:bmi + age:hearttrte + age:prevmi + age:prevhyp +
## age:hdlc + sysbp:diabetes + sysbp:bpmeds + sysbp:hearttrte +
## sysbp:prevmi + sysbp:prevhyp + diabp:cursmoke + diabp:cigpday +
## diabp:bmi + diabp:glucose + diabp:prevhyp + diabp:hdlc +
## cursmoke:bmi + cursmoke:hdlc + cursmoke:ldlc + cigpday:bmi +
## cigpday:hearttrte + cigpday:glucose + cigpday:prevmi + cigpday:hdlc +
## bmi:prevmi + bmi:prevhyp + bmi:ldlc + diabetes:prevmi + diabetes:hdlc +
## bpmeds:glucose + bpmeds:prevstrk + bpmeds:ldlc + hearttrte:glucose +
## hearttrte:prevmi + glucose:prevmi + prevmi:prevhyp + prevmi:hdlc +
## prevhyp:ldlc, data = fhsd)

beta.fwd = coef(Mfwd)
beta.back = coef(Mback)
beta.step = coef(Mstep)
identical(names(beta.fwd)[names(beta.fwd) %in% names(beta.back)], names(beta.fwd))

## [1] FALSE

identical(names(beta.fwd)[names(beta.fwd) %in% names(beta.step)], names(beta.fwd))

## [1] FALSE

identical(names(beta.back)[names(beta.back) %in% names(beta.step)], names(beta.back))

## [1] FALSE
```

## 3.2 Manual Model Selection

The following table lists terms in the stepwise model that result in insignificance when F-test is performed by removing them solely from the stepwise model along with corresponding p-values in a sorted order.

Table 2: Variates/Interactions with significant p-values from F-test

cigpday:hearttrte	bpmeds:prevstrk	bpmeds:glucose	diabp:cigpday	cigpday	sex:ldlc	age:prevmi	cigpday:prevmi	hdlc:ldlc
0.1506282	0.1492283	0.1189197	0.1155989	0.1151079	0.1141483	0.1097987	0.1051865	0.0923568

bmi:bpmeds	prevmi:prevstrk	hearttrte:prevmi	glucose:prevmi	I(sysbp^2)	cursmoke:hdlc	age:hearttrte	age:hdlc
0.0855445	0.0699776	0.0645195	0.0588312	0.0585469	0.0566094	0.0556206	0.0510796

Looking at the above table, removing highly insignificant continuous variate interactions `cigpday:hearttrte` and `diabp:cigpday`, we have the following p-value from F-test.

```
# Remove as many insignificant continuous variate interactions as possible
anova(Mstep, update(Mstep, . ~ . - cigpday:hearttrte - diabp:cigpday))$`Pr(>F)`[2]

## [1] 0.0729871
```

Assuming the insignificance threshold of 0.05, removing categorical/continuous variate interaction `bpmeds:prevstrk` results in the following p-value.

```
# Now remove insignificant interactions from categorical variates
anova(Mstep, update(Mstep, . ~ . - cigday:hearttrte - diabp:cigday - bpmeds:prevstrk))$`Pr(>F)`[2]

## [1] 0.05655719
```

Since above p-value is just slightly greater than 0.05, removing the above interactions from stepwise model is insignificant. Therefore a reduced model can be obtained from stepwise in the following way.

```
# Thus we have the following manually constructed model
Mmanual <- update(Mstep, . ~ . - cigday:hearttrte - diabp:cigday - bpmeds:prevstrk)
```

## 4 Model Diagnostics

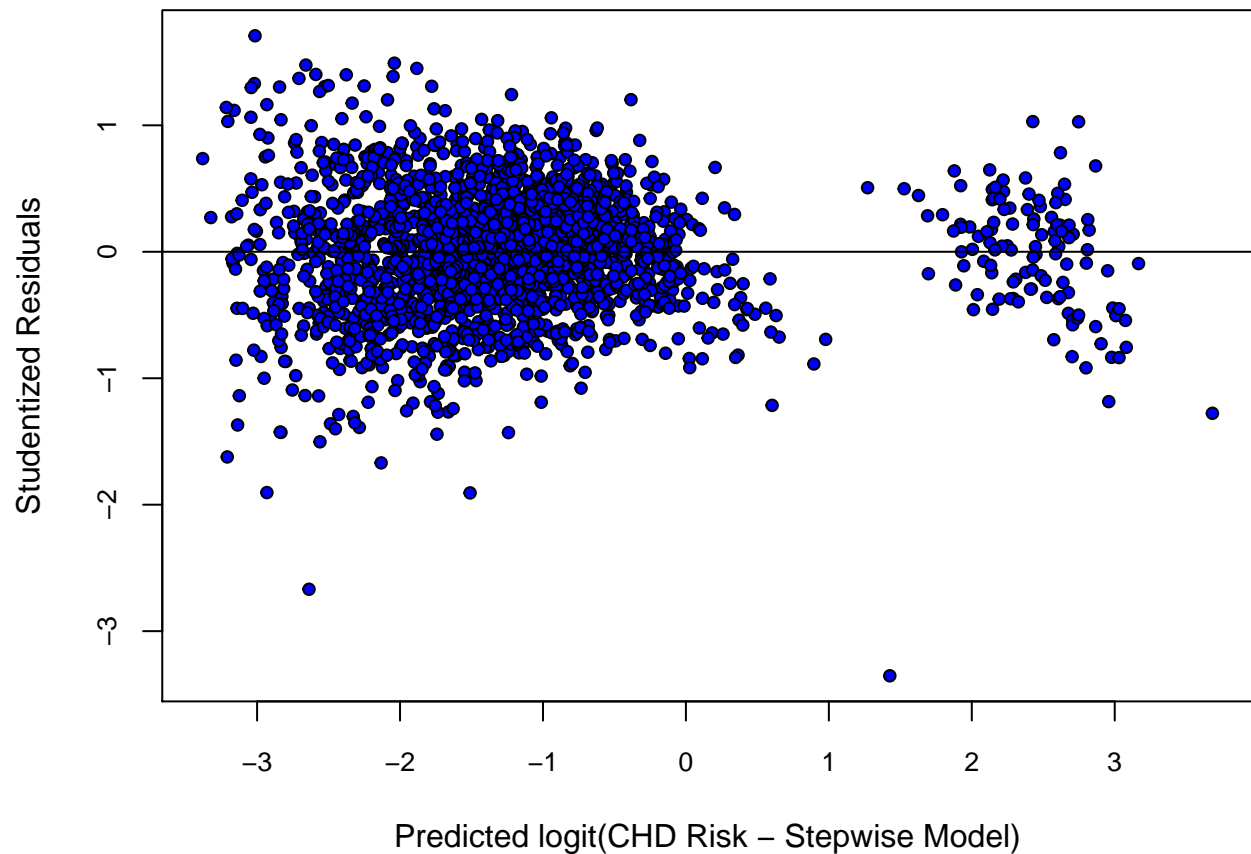
### 4.1 Residual Plots

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

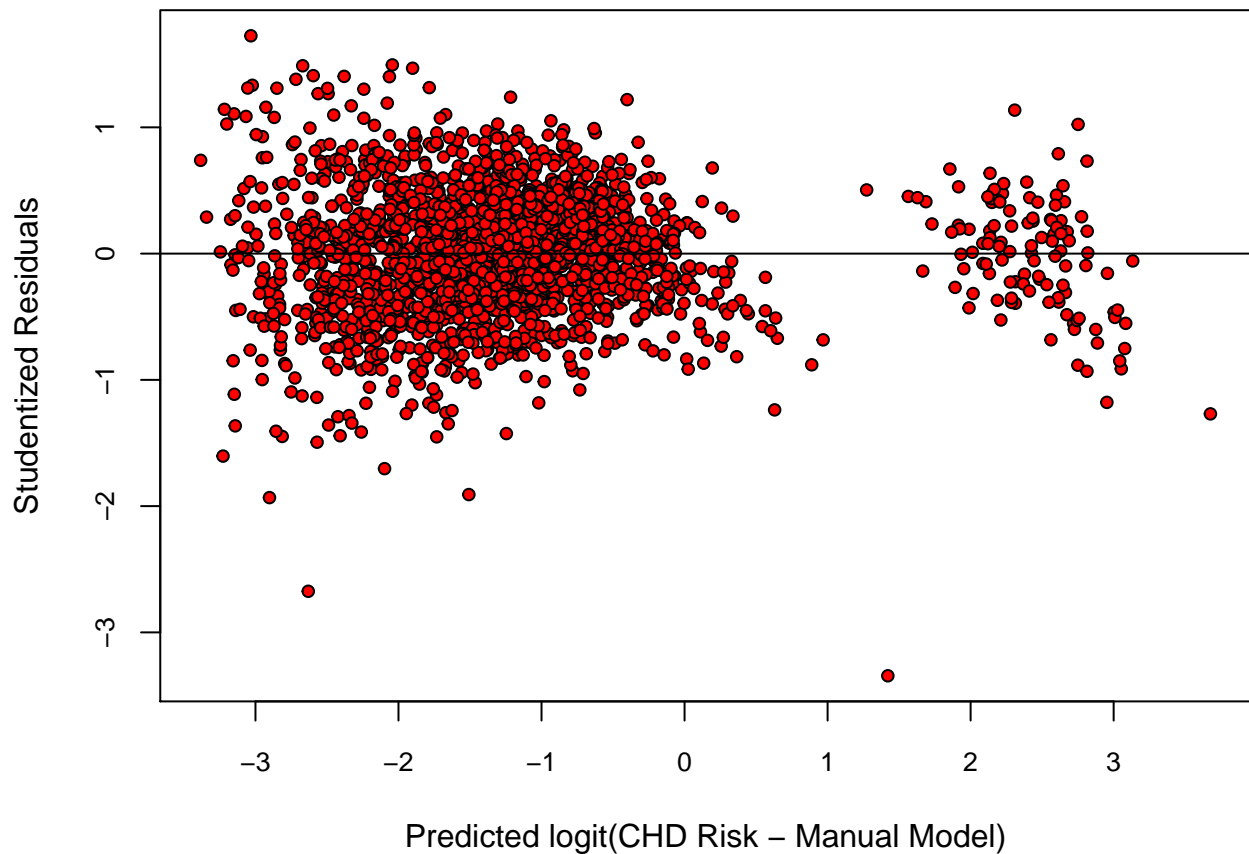
First, we have that the most normal looking residuals assuming that the model is true, would be the studentized residuals, so to check the homoscedasity we plot those values against the predicted values, as shown below:

```
# First we analyze Mstep
# get the hat values
h <- hatvalues(Mstep)
res.step <- resid(Mstep)/sqrt(1-h) # studentized residuals, but on the data scale
cex <- .8 # controls the size of the points and labels
par(mar = c(4,4,.5,.1))
plot(predict(Mstep), res.step, pch = 21, bg = "blue", cex = cex, cex.axis = cex, xlab = "Predicted logit
abline(h = 0, lty = 1, col = "black") # add horizontal line at 0
```



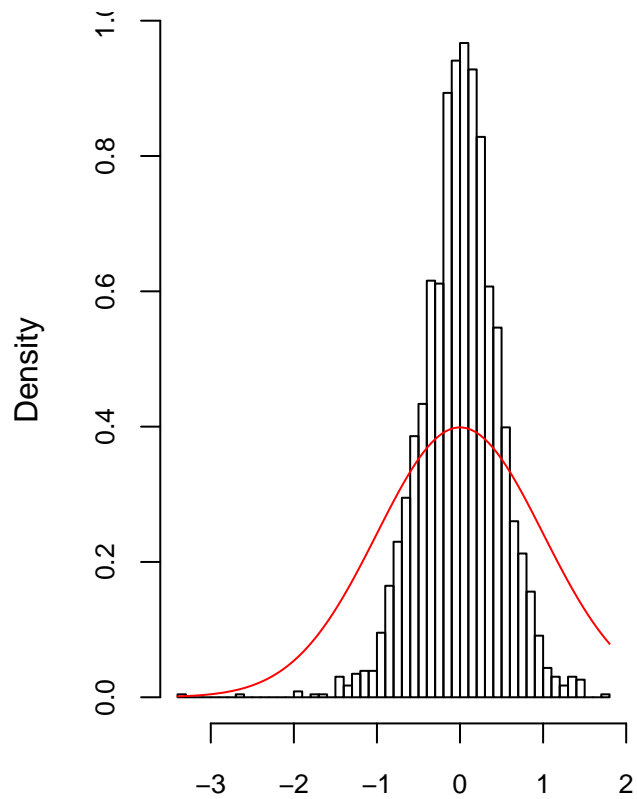


```
# Then we analyze Mdl_manual
# get the hat values
h <- hatvalues(Mmanual)
res.manual <- resid(Mmanual)/sqrt(1-h) # studentized residuals, but on the data scale
cex <- .8 # controls the size of the points and labels
par(mar = c(4,4,.5,.1))
plot(predict(Mmanual), res.manual, pch = 21, bg = "red", cex = cex, cex.axis = cex, xlab = "Predicted logit(CHD Risk – Stepwise Model)", ylab = "Studentized Residuals")
abline(h = 0, lty = 1, col = "black") # add horizontal line at 0
```

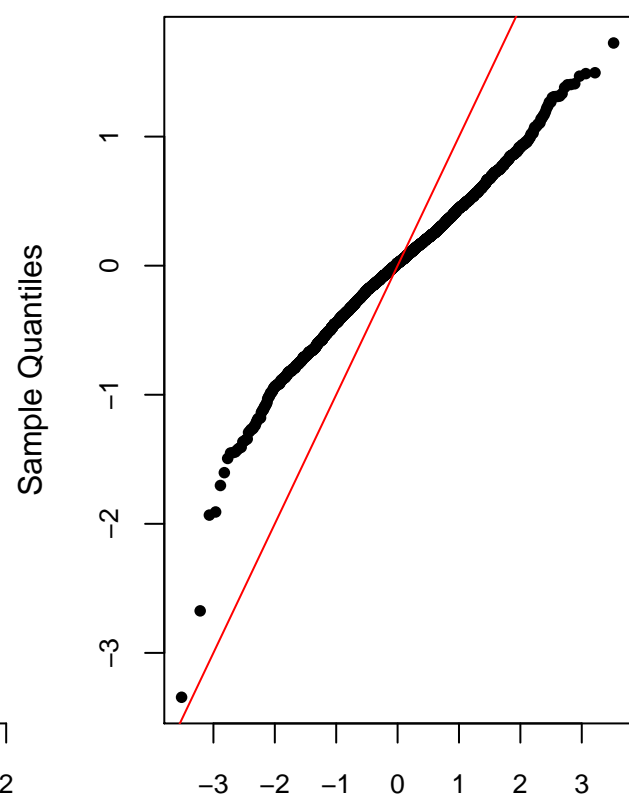


Then to check our assumption of normality of residuals we plot the residuals on a QQPlot and a histogram:

```
# plot standardized residuals
sigma.hat <- sigma(Mmanual)
cex <- .8
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
# histogram
hist(res.manual, breaks = 50, freq = FALSE, cex.axis = cex, xlab = "Studentized Residual CHD Risk (Manual Model)")
# theoretical normal curve
curve(dnorm(x), col = "red", add = TRUE)
# qq-plot
qqnorm(res.manual, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line
```

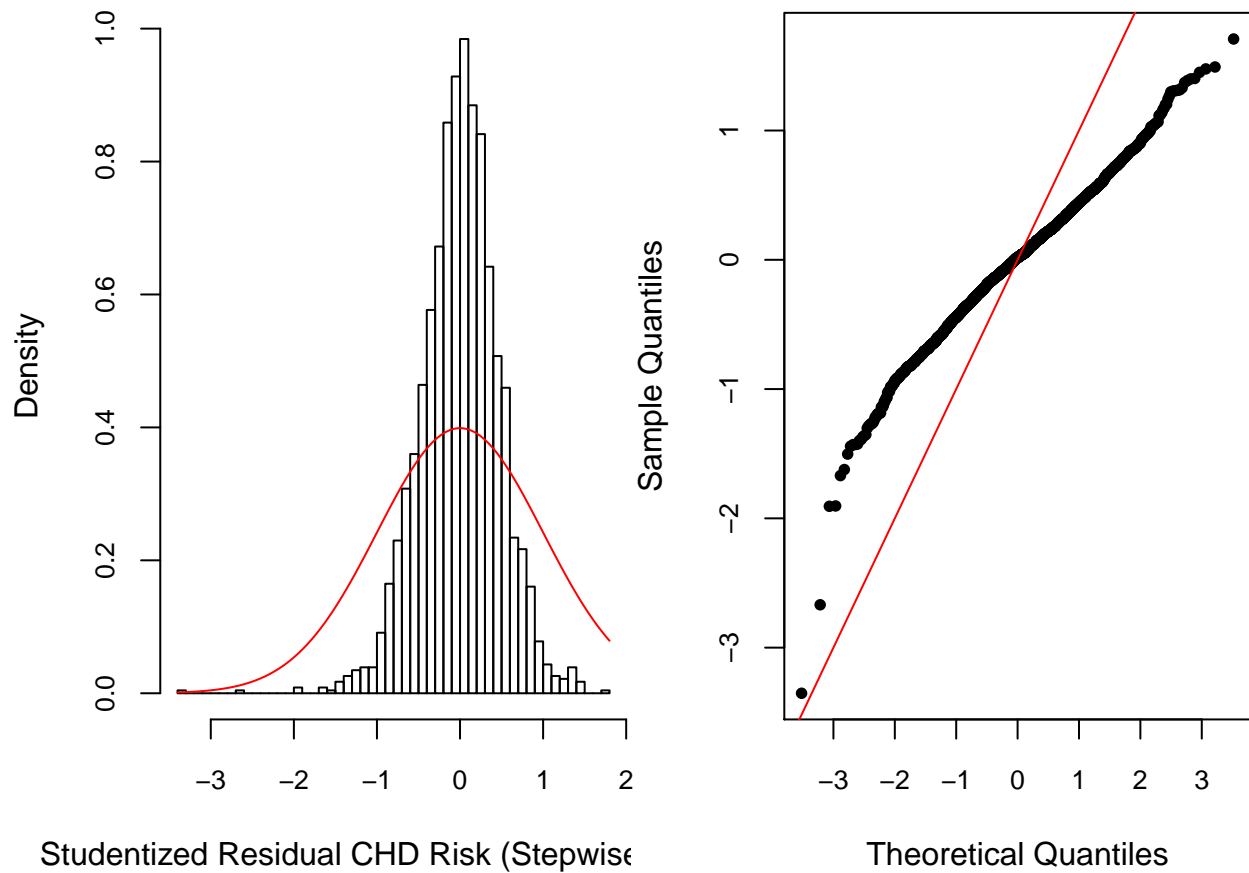


Studentized Residual CHD Risk (Manual



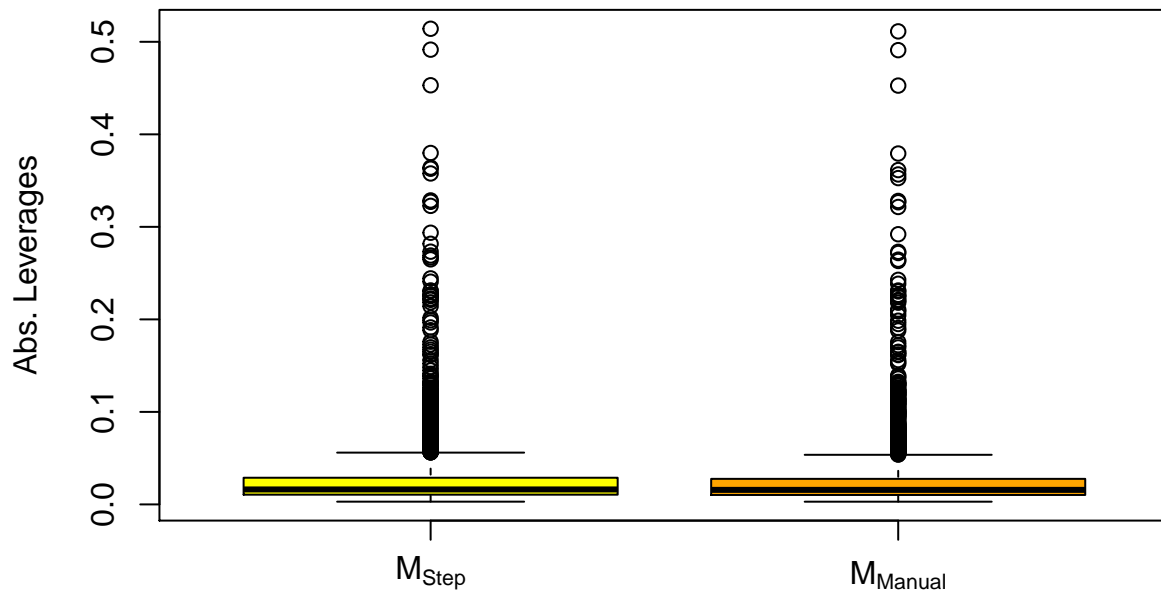
Theoretical Quantiles

```
# plot standardized residuals
sigma.hat <- sigma(Mstep)
cex <- .8
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
# histogram
hist(res.step, breaks = 50, freq = FALSE, cex.axis = cex,xlab = "Studentized Residual CHD Risk (Stepwise)
# theoretical normal curve
# qq-plot
qqnorm(res.step, main = "", pch = 16, cex = cex, cex.axis = cex)
abline(a = 0, b = 1, col = "red") # add 45 degree line
```

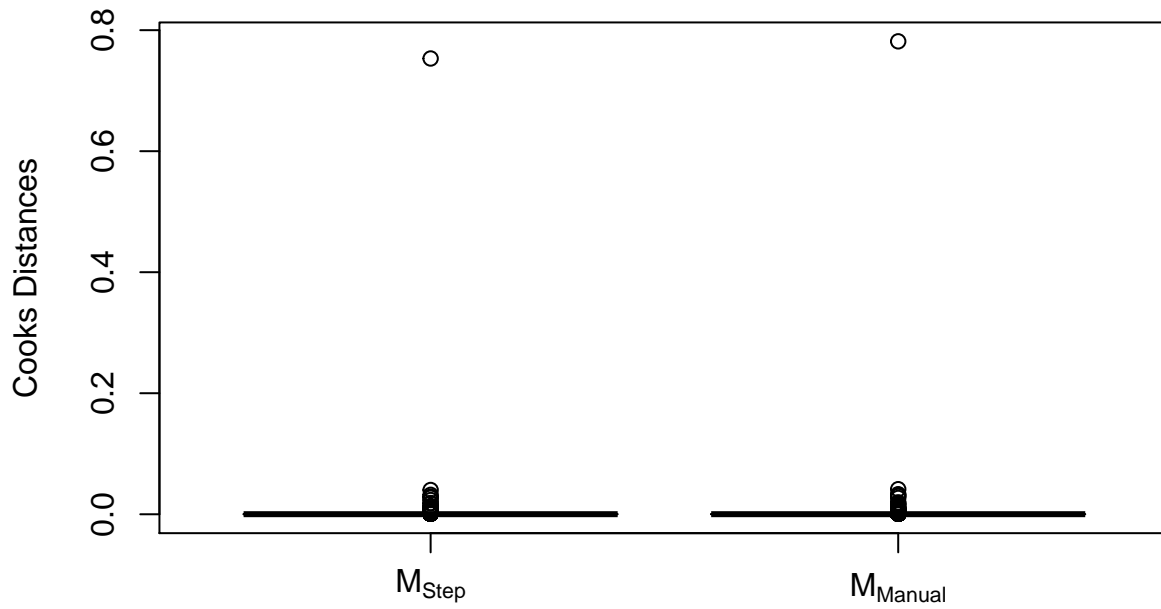


## 4.2 Leverage and Influence Measures

We have the following boxplot of absolute values of leverages of both the step-wise and manual models.



Similarly, we have the following boxplot of cook's distances of both the models.



## 5 Model Selection

### 5.1 Cross Validation

Before performing cross-validation analysis, function `logitnorm_mean` is created to approximate the conditional mean  $E[\text{chdrisk}|\mathbf{x}]$  based on the regression model  $\text{logit}(\text{chdrisk})|\mathbf{x} \sim N(\mathbf{x}'\beta, \sigma^2)$  (look into the Appendix for code). The following output is produced when tested.

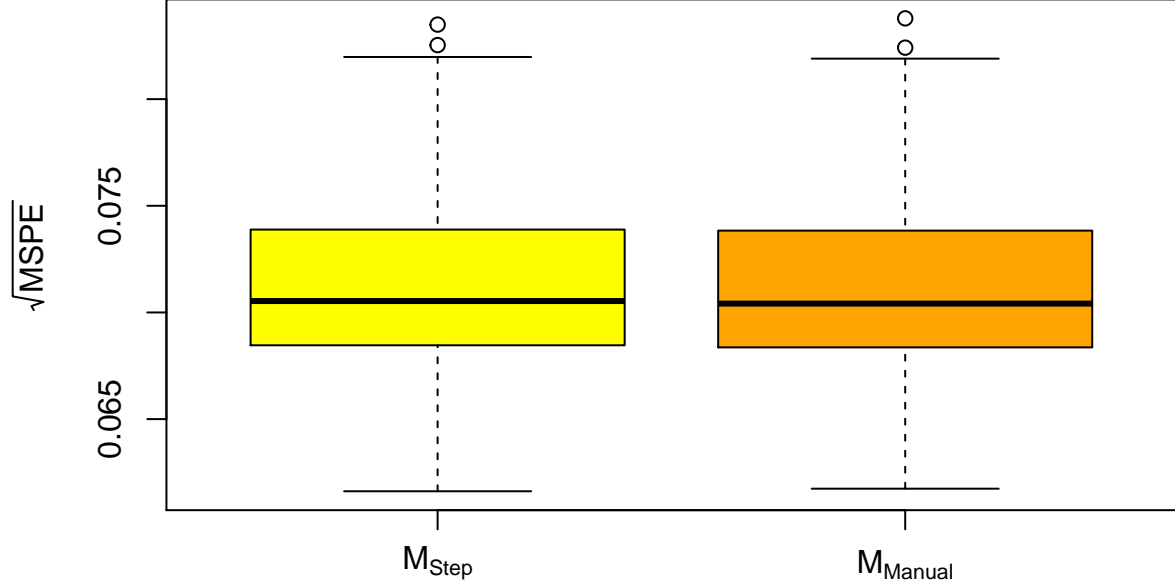
```
# Test the function
mu <- c(0.7,3.2,-1.1)
sigma <- c(0.8,0.1,2.3)
# Returns results expected in the project description
sapply(1:3, function(i) logitnorm_mean(mu[i],sigma[i]))

## [1] 0.6491002 0.9606606 0.3530580
```

The above function is then used to perform cross-validation analysis and the following boxplot that shows MSPE of the both the models is produced.

```
## user system elapsed
## 84.832 1.616 94.841
```

## Root MSPE



Model chosen will be the manual one [ADD MORE] These are the parameter estimates, std.errors and p-values of the manually constructed model.

Table 3: Summary of chosen manually constructed model

	(Intercept)	sexMale	totchol	age	sysbp	diabp	cursmokeYes	cigpday	bmi
Estimate	-6.841121	0.7509308	0.0086801	0.0562327	0.0125123	-0.0426462	0.4903809	0.0463907	-0.0843445
Std. Error	1.017652	0.1622031	0.0029290	0.0101841	0.0061648	0.0106013	0.2180640	0.0087763	0.0244266
Pr(> t )	0.000000	0.0000039	0.0030739	0.0000000	0.0425109	0.0000594	0.0246223	0.0000001	0.0005647

	diabetesYes	bpmedsYes	hearttrte	glucose	prevmiYes	prevstrkYes	prevhypYes	hdlc	ldlc
Estimate	1.0311734	0.8595578	0.0440175	0.0023589	5.6983941	0.1404944	3.9567705	-0.0201085	-0.0063626
Std. Error	0.2918380	0.3038036	0.0078437	0.0026598	0.5685983	0.0792694	0.3583848	0.0084351	0.0035374
Pr(> t )	0.0004186	0.0047062	0.0000000	0.3752314	0.0000000	0.0764694	0.0000000	0.0172127	0.0722106

	I(hdlc^2)	I(bmi^2)	I(diabp^2)	I(sysbp^2)	sysbp:prevmiYes	totchol:prevhypYes	diabetesYes:prevmiYes	prevhypYes:ldlc	sysbp:prevhypYes
Estimate	0.0001820	0.0028956	5.97e-04	0.0000426	-0.0109765	-0.0070567	-0.6245291	0.0038926	-0.0087679
Std. Error	0.0000508	0.0004308	6.57e-05	0.0000219	0.0025360	0.0011582	0.1511829	0.0011277	0.0023122
Pr(> t )	0.0003497	0.0000000	0.00e+00	0.0520383	0.0000157	0.0000000	0.0000375	0.0005671	0.0001534

	totchol:hearttrte	sysbp:diabetesYes	diabp:bmi	diabp:hdlc	prevmiYes:hdlc	prevmiYes:prevhypYes	sexMale:glucose	age:ldlc	age:hearttrte
Estimate	-7.64e-05	-0.0061063	-0.0010390	-0.0002321	0.0149580	-0.2985780	-0.0018741	0.0000606	-0.0002377
Std. Error	1.86e-05	0.0017038	0.0002507	0.0000625	0.0038971	0.1305121	0.0007338	0.0000285	0.0000989
Pr(> t )	4.09e-05	0.0003457	0.0000353	0.0002120	0.0001274	0.0222452	0.0107144	0.0337545	0.0163262

	cigpday:hdlc	bmi:ldlc	totchol:ldlc	totchol:prevmiYes	sysbp:hearttrte	sysbp:bpmedsYes	cursmokeYes:hdlc	prevmiYes:prevstrkYes	diabetesYes:hdlc
Estimate	-0.0003505	0.0001499	0.0001294	-0.0026466	-0.0001125	-0.0035185	0.0043414	-0.3881888	0.0058461
Std. Error	0.0000943	0.0000589	0.0000504	0.0009274	0.0000375	0.0015038	0.0023248	0.1951360	0.0024167
Pr(> t )	0.0002059	0.0110647	0.0102686	0.0043574	0.0027333	0.0193871	0.0619759	0.0467867	0.0156403

## 6 Discussion

	sexMale:sysbp	cigpday:glucose	heart rte:glucose	diabp:glucose	cur smokeYes:ldlc	age:cigpday	age:hdlc	hdlc:ldlc	age:prevhypYes
Estimate	-0.0018228	-0.0000755	0.0000695	-0.0000682	-0.0010001	-0.0003358	-0.0001725	-0.0000803	-0.0103638
Std. Error	0.0009765	0.0000385	0.0000260	0.0000281	0.0004913	0.0001248	0.0000888	0.0000478	0.0029473
Pr(> t )	0.0620856	0.0500605	0.0075675	0.0154340	0.0418854	0.0071905	0.0522339	0.0927933	0.0004462

	diabp:prevhypYes	diabp:cur smokeYes	bmi:bpmedsYes	bpmedsYes:glucose	age:prevmiYes	sexMale:ldlc	cigpday:prevmiYes	glucose:prevmiYes	heart rte:prevmiYes
Estimate	-0.0110683	-0.0057752	-0.0128143	0.0014572	-0.0105250	0.0007191	-0.0087788	-0.0027211	0.0058756
Std. Error	0.0034311	0.0019148	0.0072263	0.0009960	0.0063892	0.0004619	0.0049761	0.0015063	0.0033641
Pr(> t )	0.0012740	0.0025893	0.0763179	0.1436117	0.0996373	0.1196056	0.0778372	0.0709783	0.0808493