

# STAT 331 Final Project

Krishna Prem Pasumarthu & Islam Amin

April 11, 2020

## 1 Summary

## 2 Descriptive Statistics

First, take a look at summary statistics of the `fhsc` dataset.

Table 1: Summary Statistics

| chdrisk        | sex         | totchol       | age           | sysbp         | diabp         | cursmoke | cigday        | bmi           | diabetes | hpmeds   | hearttrt       | glucose        | prevmi   | prevstrk | prevhbp  | hdlc           | ldlc          |
|----------------|-------------|---------------|---------------|---------------|---------------|----------|---------------|---------------|----------|----------|----------------|----------------|----------|----------|----------|----------------|---------------|
| 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   | No :2142 | No :1973 | Min. : 44.00   | Min. : 46.00   | No :2189 | No :2260 | No : 957 | Min. : 10.00   | Min. : 20.0   |
| 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 | Yes :164 | Yes :353 | 1st Qu.: 70.00 | 1st Qu.: 75.00 | Yes :117 | Yes :46  | Yes:1349 | 1st Qu.: 38.00 | 1st Qu.:152.0 |
| Median :0.2240 | NA          | Median :235.5 | Median :60.00 | Median :136.0 | Median :80.00 | NA       | Median : 0.00 | Median :25.40 | NA       | NA       | Median : 76.00 | Median : 83.00 | NA       | NA       | NA       | Median : 47.00 | Median :180.0 |
| Mean :0.2655   | NA          | Mean :237.8   | Mean :60.23   | Mean :139.2   | Mean :81.07   | NA       | Mean : 6.84   | Mean :25.78   | NA       | NA       | Mean : 77.61   | Mean : 89.07   | NA       | NA       | NA       | Mean : 48.89   | Mean :183.1   |
| 3rd Qu.:0.3448 | NA          | 3rd Qu.:265.0 | 3rd Qu.:67.00 | 3rd Qu.:153.0 | 3rd Qu.:88.00 | NA       | 3rd Qu.:10.00 | 3rd Qu.:27.91 | NA       | NA       | 3rd Qu.: 85.00 | 3rd Qu.: 95.00 | NA       | NA       | NA       | 3rd Qu.: 57.00 | 3rd Qu.:210.0 |
| Max. :0.9770   | NA          | Max. :625.0   | Max. :81.00   | Max. :246.0   | Max. :130.00  | NA       | Max. :80.00   | Max. :46.52   | NA       | NA       | Max. :150.00   | Max. :478.00   | NA       | NA       | NA       | Max. :189.00   | Max. :565.0   |

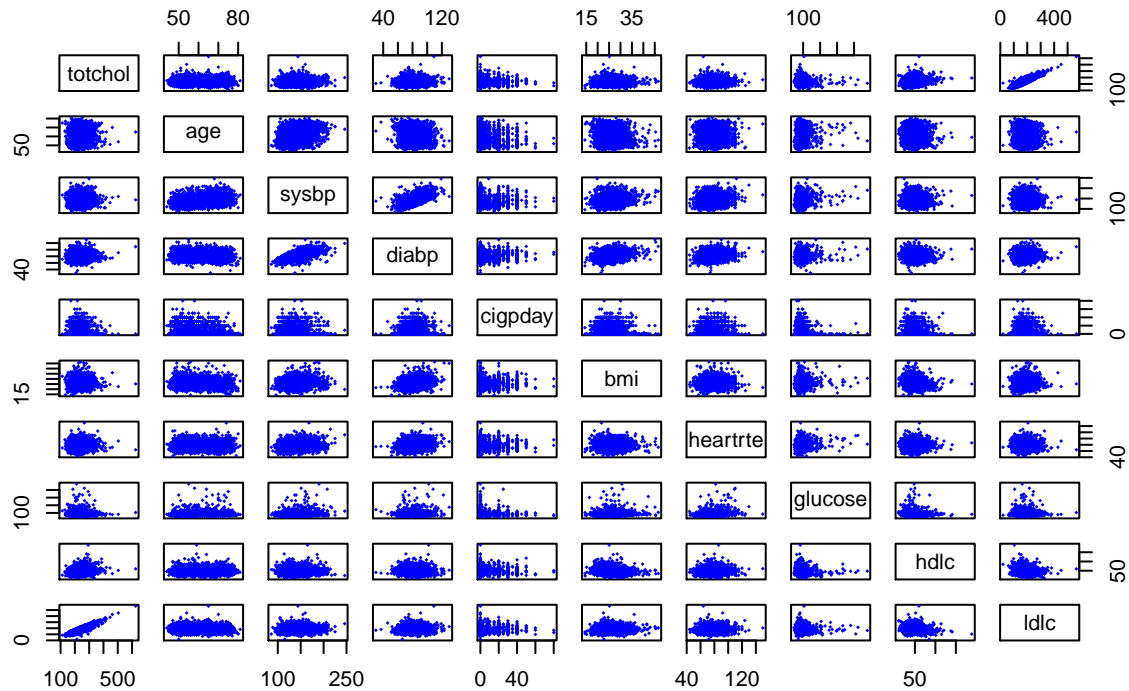
Then take a look at `chdrisk` grouped by `sex` as well as `chdrisk` grouped by `cursmoke`.

```
## fhsc$sex: Female
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.005   0.104   0.179   0.215   0.285   0.949
## -----
## fhsc$sex: Male
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.0210  0.1860  0.2860  0.3314  0.4060  0.9770
## -----
## fhsc$cursmoke: No
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.0050  0.1390  0.2350  0.2754  0.3580  0.9770
## -----
## fhsc$cursmoke: Yes
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.0080  0.1220  0.1995  0.2471  0.3140  0.9710
```

[ADD SOME COMMENTS HERE REGARDING SUMMARY]

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

```
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 = M0,
      scope = list(lower = M0, upper = Mmax),
      direction = "forward", 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 + hearttrte +
##     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:hearttrte +
##     sysbp:diabetes + diabp:bmi + diabp:hdlc + prevmi:hdlc + prevmi:prevhyp +
##     sex:glucose + age:ldlc + age:hearttrte + cigpday:hdlc + bmi:ldlc +
##     totchol:hdlc + totchol:prevmi + sysbp:hearttrte + sysbp:bpmeds +
##     cursmoke:hdlc + prevmi:prevstrk + diabetes:hdlc + sex:sysbp +
##     cigpday:glucose + hearttrte: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:hearttrte + cigpday:prevmi +
##     glucose:prevmi + hearttrte: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 +
##     hearttrte + 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:hearttrte + age:hearttrte + diabetes:hdlc + sysbp:hearttrte +
##     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)) ~ prevmi + sysbp + sex + age +  
##     ldlc + prevhyp + diabetes + hdlc + I(hdlc^2) + cigpday +  
##     I(bmi^2) + bmi + totchol + I(glucose^2) + I(sysbp^2) + bpmeds +  
##     heartрте + cursmoke + prevstrk + prevmi:sysbp + sysbp:age +  
##     prevhyp:ldlc + 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:heartрте + age:heartрте + diabetes:hdlc + sysbp:heartрте +  
##     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)
```

### 3.2 Manual Model Selection

## 4 Model Diagnostics

## 5 Model Selection

## 6 Discussion