FHS 4240 Logistic Regression

Uohna Thiessen

January 18, 2017

Step 1: Load data and run numerical and graphical summaries

```
framingham_edx <- read.csv("framingham_edx.csv")
fhs = framingham_edx
str(fhs)
```

```
## 'data.frame':    4240 obs. of  16 variables:
##  $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
##  $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
##  $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
##  $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

```
summary(fhs)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54    Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.07    1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.40    Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.80    Mean   : 75.88   Mean   : 81.96   Mean   :0.1519
##  3rd Qu.:28.04    3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80    Max.   :143.00   Max.   :394.00   Max.   :1.0000
##  NA's   :19       NA's   :1        NA's   :388
```

```
library(caTools)
library(GGally)
library(ggplot2)
library(corrplot)
```

Step 2: Convert categorical variables to factors

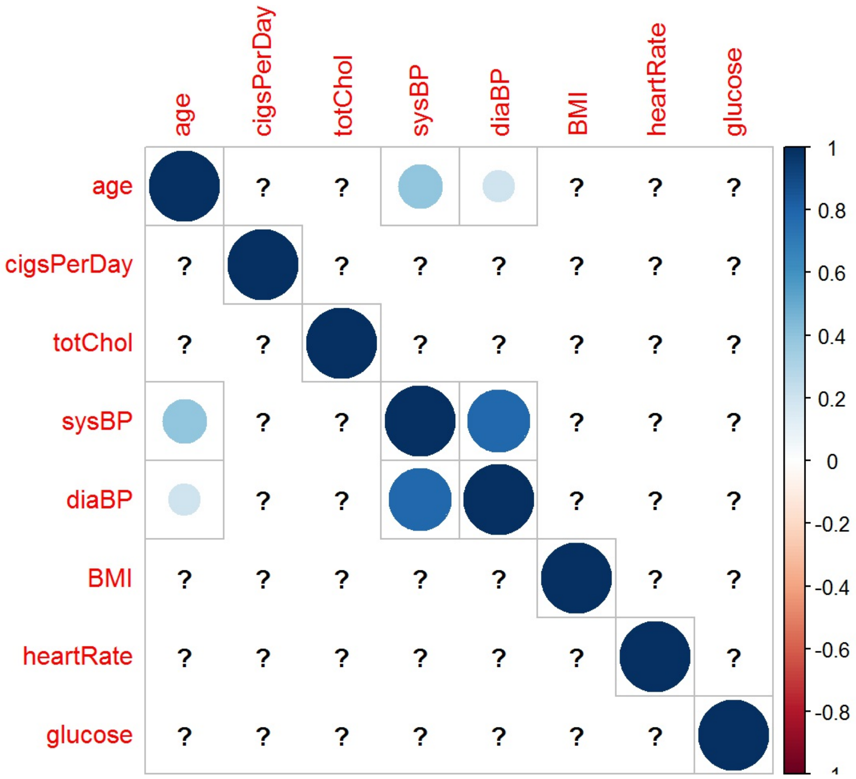
```
fhs$BPMeds = as.factor(fhs$BPMeds)
fhs$prevalentStroke = as.factor(fhs$prevalentStroke)
fhs$prevalentHyp = as.factor(fhs$prevalentHyp)
fhs$diabetes = as.factor(fhs$diabetes)
```

Step 2: Scaling/Centering quantitative variables

```
#fhs$age = scale(fhs$age)
#fhs$cigsPerDay = scale(fhs$cigsPerDay)
#fhs$totChol = scale(fhs$cigsPerDay)
#fhs$sysBP = scale(fhs$sysBP)
#fhs$diaBP = scale(fhs$diaBP)
#fhs$BMI = scale(fhs$BMI)
#fhs$glucose = scale(fhs$glucose)
```

Step 3: Generate scatterplots

```
fhs_num = subset(fhs[c(2,5,10:15)])
corrplot(cor(fhs_num), method = "circle")
```



```
ggpairs(fhs_num)
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 29 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 19 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 388 rows containing missing values
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```

```
## Warning: Removed 29 rows containing non-finite values (stat_density).
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 79 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 29 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 29 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 48 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 30 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 413 rows containing missing values
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

```
## Warning: Removed 79 rows containing missing values (geom_point).
```

```
## Warning: Removed 50 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 50 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 68 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 51 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 398 rows containing missing values
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 19 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 388 rows containing missing values
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 19 rows containing missing values
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 388 rows containing missing values
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```

```
## Warning: Removed 48 rows containing missing values (geom_point).
```

```
## Warning: Removed 68 rows containing missing values (geom_point).
```

```
## Warning: Removed 19 rows containing missing values (geom_point).

## Warning: Removed 19 rows containing missing values (geom_point).
```

```
## Warning: Removed 19 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 20 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 402 rows containing missing values
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 30 rows containing missing values (geom_point).
```

```
## Warning: Removed 51 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 389 rows containing missing values
```

```
## Warning: Removed 388 rows containing missing values (geom_point).
```

```
## Warning: Removed 413 rows containing missing values (geom_point).
```

```
## Warning: Removed 398 rows containing missing values (geom_point).
```

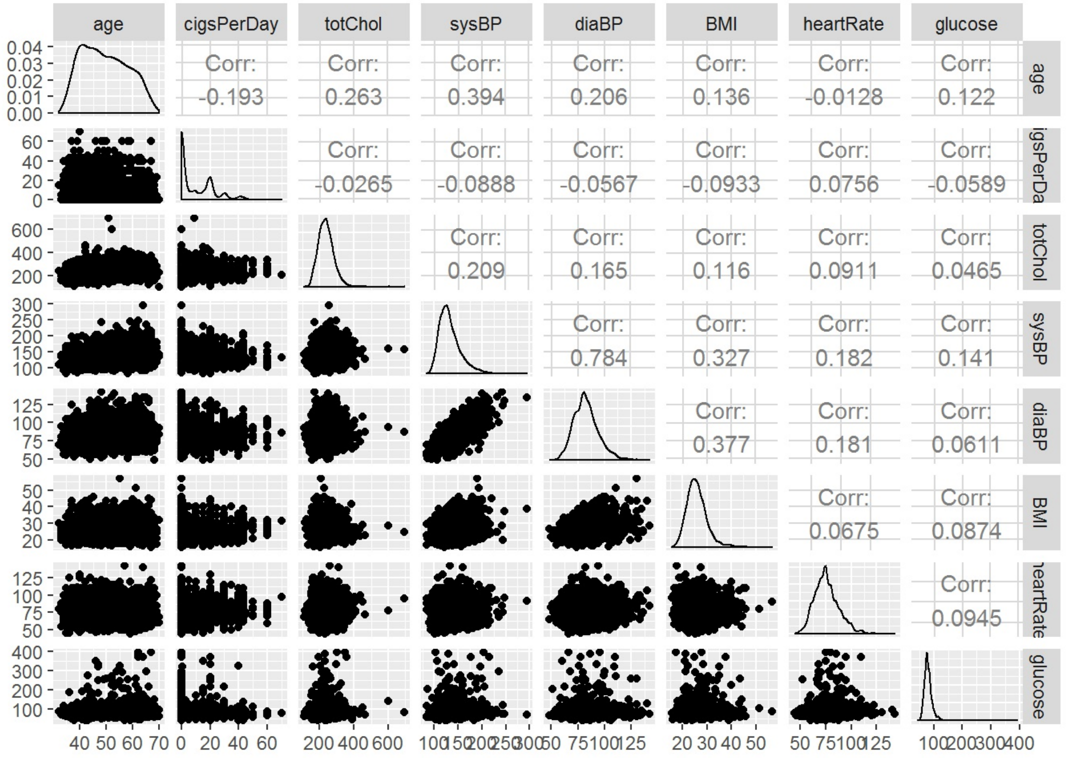
```
## Warning: Removed 388 rows containing missing values (geom_point).

## Warning: Removed 388 rows containing missing values (geom_point).
```

```
## Warning: Removed 402 rows containing missing values (geom_point).
```

```
## Warning: Removed 389 rows containing missing values (geom_point).
```

```
## Warning: Removed 388 rows containing non-finite values (stat_density).
```



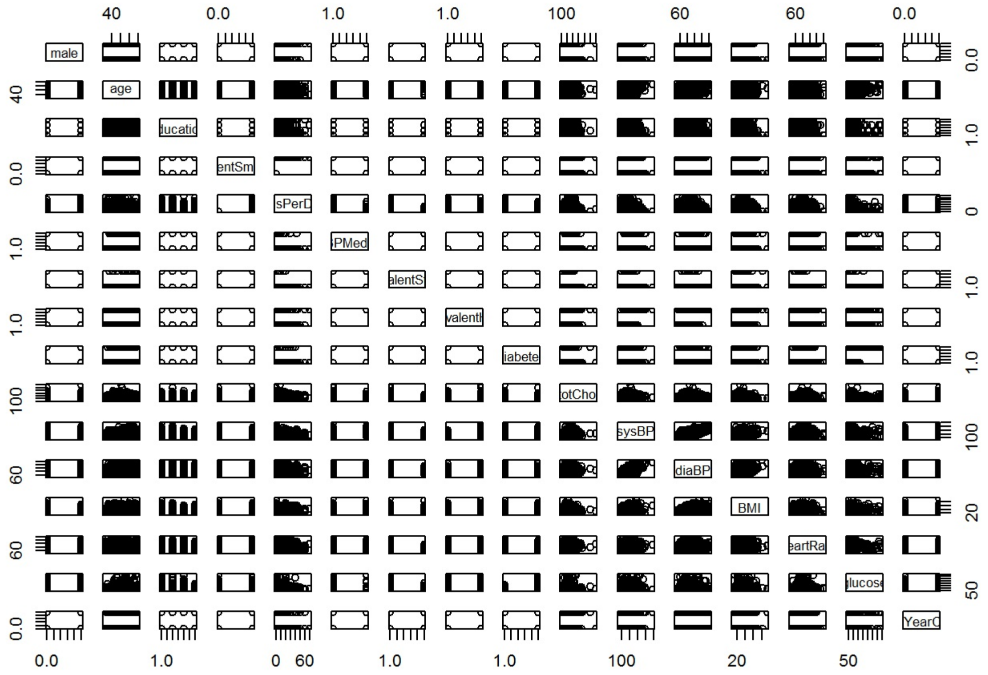
Step 5: Spliting the data for building and testing the model

```
set.seed(1000)
split = sample.split(fhs, SplitRatio = 0.75)
fhs_train = subset(fhs, split == TRUE)
fhs_test = subset(fhs, split == FALSE)
fhs_LR = glm(TenYearCHD ~., data = fhs_train, family = binomial)
summary(fhs_LR)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = fhs_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6392  -0.5908  -0.4153  -0.2797   2.8573
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.914454   0.837084  -9.455  < 2e-16 ***
## male               0.531158   0.125849   4.221 2.44e-05 ***
## age                0.067861   0.007831   8.665  < 2e-16 ***
## education         -0.042357   0.056380  -0.751   0.4525
## currentSmoker      0.030377   0.181626   0.167   0.8672
## cigsPerDay         0.017325   0.007269   2.383   0.0172 *
## BPMeds1            0.334846   0.278809   1.201   0.2298
## prevalentStroke1  0.550715   0.604485   0.911   0.3623
## prevalentHyp1      0.238585   0.161295   1.479   0.1391
## diabetes1          0.018776   0.353446   0.053   0.9576
## totChol            0.001925   0.001319   1.459   0.1445
## sysBP              0.017334   0.004418   3.924 8.72e-05 ***
## diaBP             -0.004571   0.007479  -0.611   0.5411
## BMI               -0.001039   0.014983  -0.069   0.9447
## heartRate         -0.007550   0.004990  -1.513   0.1303
## glucose            0.004391   0.002637   1.665   0.0959 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2328.6  on 2755  degrees of freedom
## Residual deviance: 2049.8  on 2740  degrees of freedom
##   (424 observations deleted due to missingness)
## AIC: 2081.8
##
## Number of Fisher Scoring iterations: 5
```

Step 6: Testing and improving the accuracy of the model

```
pairs(fhs)
```



```
table(fhs$TenYearCHD)
```

```
##
##    0    1
## 3596  644
```

```
fhs_LR_2 = glm(TenYearCHD ~ +male + age + cigsPerDay + sysBP + glucose, data = fhs_train, family = binomial)
summary(fhs_LR_2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ +male + age + cigsPerDay + sysBP +
##     glucose, family = binomial, data = fhs_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7777  -0.5858  -0.4197  -0.2870   2.8688
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.890138   0.482578 -18.422  < 2e-16 ***
## male         0.489661   0.118515   4.132 3.60e-05 ***
## age          0.072611   0.007301   9.945  < 2e-16 ***
## cigsPerDay   0.017904   0.004753   3.767 0.000165 ***
## sysBP        0.019116   0.002459   7.775 7.57e-15 ***
## glucose      0.004698   0.001963   2.394 0.016667 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2443.0  on 2883  degrees of freedom
## Residual deviance: 2159.3  on 2878  degrees of freedom
##   (296 observations deleted due to missingness)
## AIC: 2171.3
##
## Number of Fisher Scoring iterations: 5
```