Prevalence of Heart Disease in Framington, MA

Dustin Leatherman

2/15/2020

Introduction

Data Analysis

```
heart.clean <-
  heart %>%
  # there are several discrete features with low cardinality so
  # treating them as factors
  mutate(
    isMale = factor(male),
    education.factor = factor(education),
    currentSmoker.factor = factor(currentSmoker),
    prevalentStroke.factor = factor(prevalentStroke),
    prevalentHyp.factor = factor(prevalentHyp),
    diabetes.factor = factor(diabetes),
    TenYearCHD.factor = factor(TenYearCHD),
    BPMeds.factor = factor(BPMeds)
  ) %>%
  select(-c(prevalentStroke, male, education, prevalentHyp, diabetes, currentSmoker, TenYearCHD, BPMeds
heart.clean %>% ggpairs(
  aes(color = TenYearCHD.factor, alpha = 0.3),
    # correlation text is off so this makes it readable
   upper = list(continuous = wrap("cor", size = 3, hjust=0, alignPercent=1)),
   title = c("Scatterplot by Ten Year CHD"),
  ) + labs(caption = "Figure 1. Relationships between features are explored in regards to whether or no
```

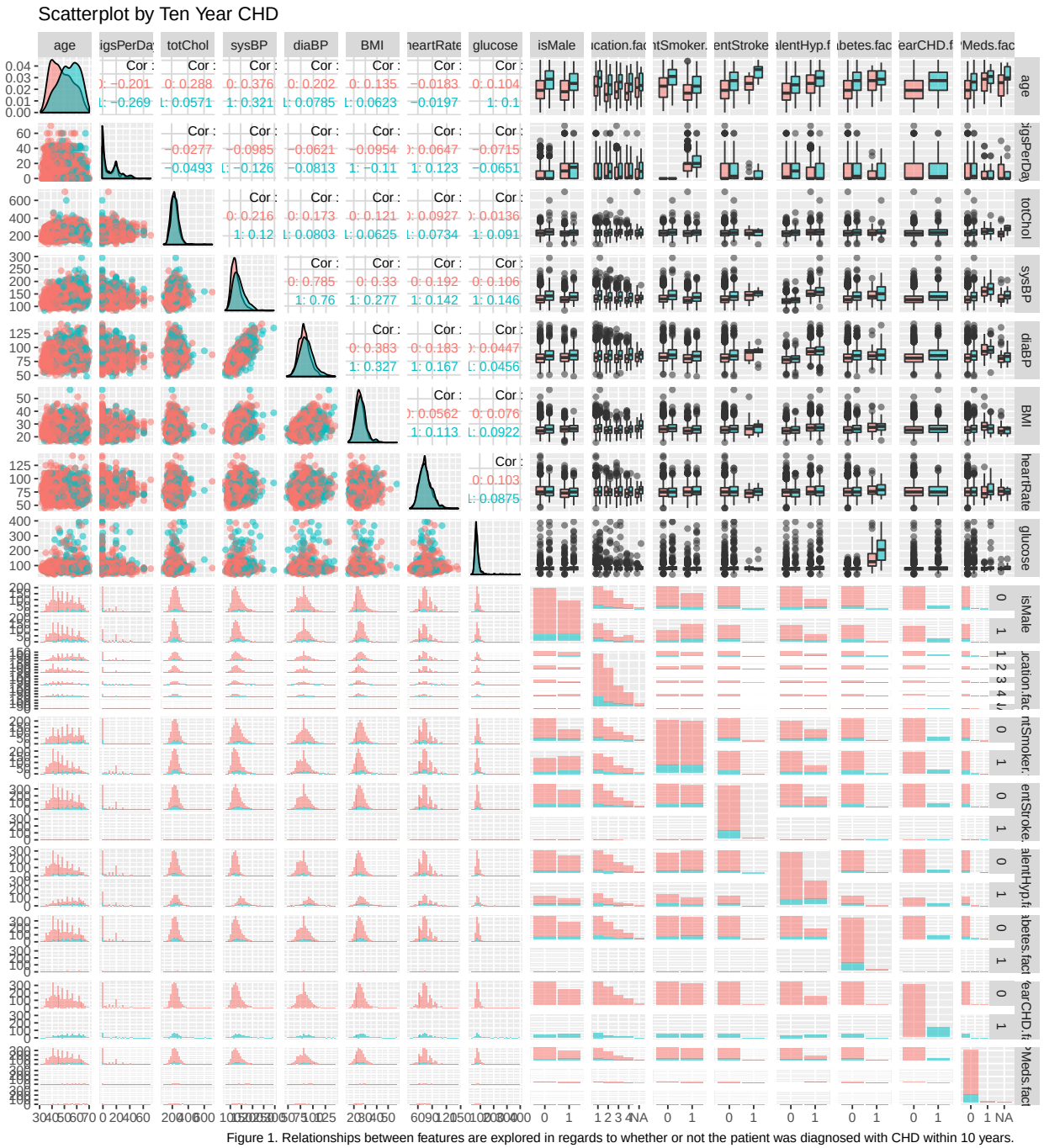


Figure 1. Relationships between features are explored in regards to whether or not the patient was diagnosed with CHD within 10 years.

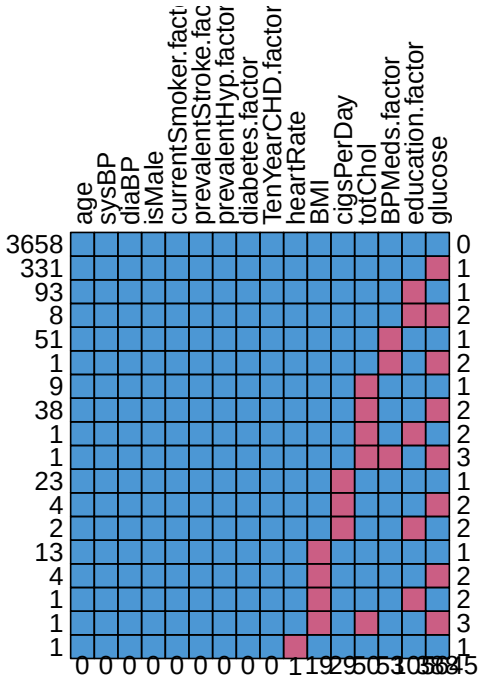
Observations

- The Age density chart indicates that a larger number of older patients were diagnosed than younger patients. The age in which both meet is around 50.
- Cigs Per Day, total Cholesterol, Systolic BP, Diastolic BP, BMI, Heart Rate, and Glucose levels *all* appear to be right-skewed according to their Density Charts. Typically a log transformation is applied in such cases, but may not be the best choice since we will be applying Logistic Regression later on.
- The only strong correlation present is between Diastolic BP and Systolic BP. These are similar for patients with and without CDH. Likely one of these variables can be discarded.
- The scatterplots comparing Glucose vs Systolic BP, Diastolic BP, and BMI show that there appear to be higher glucose levels for those who have been diagnosed with CHD.

- A similar pattern is seen in the Diabetes vs Glucose box plots. There is a wider range of Glucose levels between the 25th and 75th quantiles and a significantly higher median for those diagnosed with CDH. This indicates glucose levels may be a relevant predictor.
- The prevalent Hypertension vs Age Boxplot indicates that the 25th, 50yh, and 75th quantile values are larger for older patients with prevalent Hypertension. The values are even larger for those who have been diagnosed with CHD indicating Prevalent Hypertension may be associated with CHD.
- The Prevalent Stroke vs Age Boxplot shows similar characteristics to the Prevalent Hypertension vs Age Boxplot indicating that there may be a relationship with CHD.

Missing Data

```
# show missingness Graph
md.pattern(heart.clean, rotate.names = TRUE)
```



There are 645 rows which contain missing data. The indicator graph shows that missing data typically falls into a select few fields. This indicates that the data is not Missing Completely at Random (MCAR).

```
# Run Little's Test to determine if the data is Missing Completely at Random (MCAR)
LittleMCAR(heart.clean)$p.value
```

```
## this could take a while
```

```
## [1] 4.955219e-08
```

There is convincing evidence that the missing data is not completely random (Little's Test). Thus, it is inappropriate to drop the data as it would be dropping meaningful patterns from the analysis. Therefore, the missing values will be imputed using Multiple Imputation with Markov Chain Monte Carlo simulations.

Analysis

```
# create a 30% sample for training data. The 30% is arbitrary
heart.samples <-
  stratified(heart.clean, c("TenYearCHD.factor"), .3, bothSets = TRUE)

heart.testing <- heart.samples$SAMP1
heart.training <- heart.samples$SAMP2
```

Dropped Data

```
heart.clean.w.TenYearCHD <-
  heart.training %>%
    mutate(TenYearCHD = as.numeric(as.character(TenYearCHD.factor))) %>%
    select(-education.factor)
# as.numeric(as.character(heart.clean$TenYearCHD.factor)): converts factor back to 0 and 1. Otherwise,
# 1 and 2.
fit.reg <- glm(
    TenYearCHD ~ age + sysBP + BMI + glucose +  prevalentHyp.factor + prevalentStroke.factor + BPMeds.fa
    family = "binomial",
    data = heart.clean.w.TenYearCHD %>% drop_na()
  )
```

```
tidy(fit.reg) %>%
  kable(
    digits = 4
  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-8.8428	0.7387	-11.9704	0.0000
age	0.0692	0.0077	8.9824	0.0000
sysBP	0.0105	0.0034	3.0987	0.0019
BMI	0.0295	0.0143	2.0689	0.0386
glucose	0.0084	0.0019	4.3728	0.0000
prevalentHyp.factor1	0.2642	0.1600	1.6509	0.0988
prevalentStroke.factor1	1.2112	0.5974	2.0273	0.0426
BPMeds.factor1	0.1930	0.2716	0.7108	0.4772
totChol	0.0017	0.0013	1.3540	0.1757
cigsPerDay	0.0294	0.0046	6.3495	0.0000
heartRate	-0.0022	0.0048	-0.4641	0.6426

There are a handful of variables that are considered not significant in predicting risk for CHD.

```
car::vif(fit.reg) %>% kable(
  caption = "Variance Inflation Factors for CHD Predictors"
)
```

The low Variance Inflation Factors indicate that multicollinearity between predictors is not significantly present. This indicates that the high p-values are likely related to being statistically insignificant opposed to its information already being included in the model via other predictors.

Table 1: Variance Inflation Factors for CHD Predictors

	x
age	1.215330
sysBP	2.126120
BMI	1.117642
glucose	1.023262
prevalentHyp.factor	1.920238
prevalentStroke.factor	1.013310
BPMeds.factor	1.108083
totChol	1.049704
cigsPerDay	1.112941
heartRate	1.076032

Reduced Model

```
# calculate drop in deviance based on a full and reduced glm model
dind <- function (glm.full, glm.reduced) {
  lrt <- glm.reduced$deviance - glm.full$deviance
  lrt.df <- glm.reduced$df.residual - glm.full$df.residual
  1 - pchisq(lrt, lrt.df)
}

fit.reg.reduced <- glm(
    TenYearCHD ~ age + sysBP + glucose +  prevalentHyp.factor + cigsPerDay,
    family = "binomial",
    data = heart.clean.w.TenYearCHD %>% drop_na()
  )

dind(fit.reg, fit.reg.reduced)
```

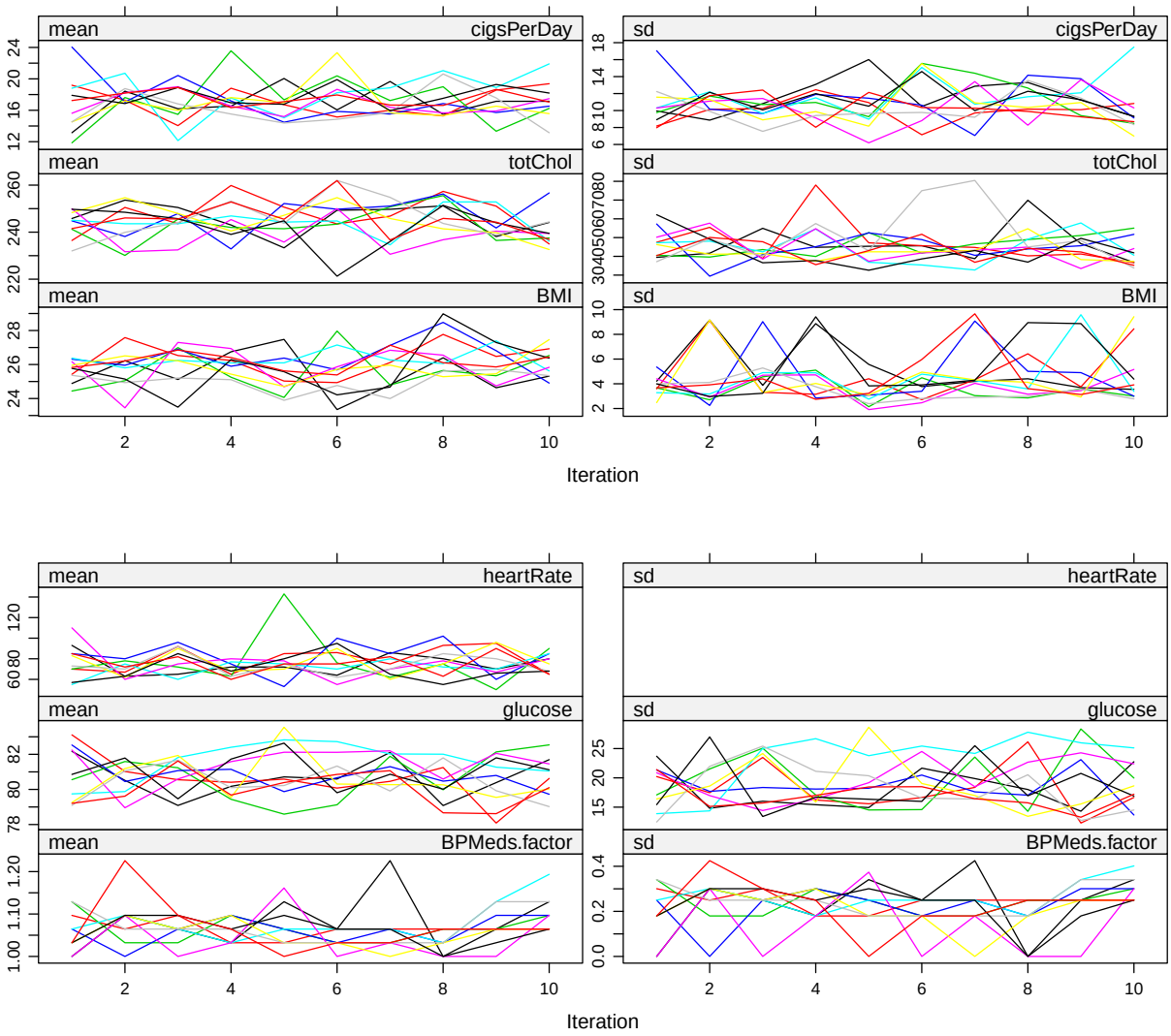
```
## [1] 0.06055354
```

Removing the insignificant parameters from the full model and comparing with the significant values in a reduced model shows that there is no evidence that the full model explains more deviance than the reduced model (Drop-in-Deviance Test. p-value = 0.2241). Going forward, this reduced model is what will be used.

Imputed Data

```
imp <- mice(heart.clean.w.TenYearCHD, m = 10, maxit = 10, seed = 123)

plot(imp)
```



Over 10 iterations of imputed values, it is ideal to see that the lines in both the mean and standard deviation intermingle and be free of any trends as the number of iterations increase. A seed is used in order to provide reproducability in the generation of values. The lines appear to intermingle and no significant trends are visible.

```
# fit our logistic regression model on the imputed values
fit.imp <- with(data = imp, glm(TenYearCHD ~ age + sysBP + glucose +  prevalentHyp.factor + cigsPerDay,

# pool the imputations together
fit.pool <- pool(fit.imp)

fit.pool$pooled %>% as_tibble(rownames = "term") %>%
  kable(
    digits = 4
  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

term	estimate	ubar	b	t	dfcom	df	riv	lambda	fmi
(Intercept)	-7.9979	0.2612	0.001	0.2623	2962	2929.953	0.0043	0.0043	0.0049
age	0.0710	0.0001	0.000	0.0001	2962	2958.061	0.0006	0.0006	0.0012
sysBP	0.0106	0.0000	0.000	0.0000	2962	2956.130	0.0010	0.0010	0.0017
glucose	0.0090	0.0000	0.000	0.0000	2962	1927.910	0.0404	0.0388	0.0398
prevalentHyp.factor1	0.3619	0.0223	0.000	0.0223	2962	2958.397	0.0005	0.0005	0.0011
cigsPerDay	0.0276	0.0000	0.000	0.0000	2962	2879.436	0.0079	0.0078	0.0085

The fractional information missing due to nonresponse (fmi) and the relative in crease in variance due to nonresponse are low which indicates the imputed data doesn't have a significant effect on the shape of the data itself.

```
summary(fit.pool) %>% as_tibble(rownames = "term") %>%
  kable (
    digits = 4
  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	-7.9979	0.5121	-15.6169	2929.953	0.0000
age	0.0710	0.0072	9.8924	2958.061	0.0000
sysBP	0.0106	0.0031	3.4145	2956.130	0.0006
glucose	0.0090	0.0019	4.7657	1927.910	0.0000
prevalentHyp.factor1	0.3619	0.1492	2.4251	2958.397	0.0154
cigsPerDay	0.0276	0.0044	6.3163	2879.436	0.0000

Results

```
pred <- predict(fit.reg.reduced, newdata = heart.testing %>% mutate(TenYearCHD = as.numeric(as.characte
confusionMatrix(table(as.numeric(pred > 0.5), heart.testing %>% mutate(TenYearCHD = as.numeric(as.chara
```

```
## Confusion Matrix and Statistics
##
##
##       0   1
##   0 936 152
##   1   3  11
##
##                Accuracy : 0.8593
##                  95% CI : (0.8374, 0.8793)
##     No Information Rate : 0.8521
##     P-Value [Acc > NIR] : 0.2642
##
##                   Kappa : 0.1033
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99681
##             Specificity : 0.06748
##          Pos Pred Value : 0.86029
##          Neg Pred Value : 0.78571
##              Prevalence : 0.85209
##          Detection Rate : 0.84936
```

```
##    Detection Prevalence : 0.98730
##       Balanced Accuracy : 0.53214
##
##        'Positive' Class : 0
##
```