HW6

Ben Tankus

2/12/2021

1. (Adapted from Exercise #3 in Chapter 13 of Linear Models in R by J. Faraway.) The pima dataset contains information on 768 adult female Pima Indians living near Phoenix.

(a) (2 points) It has been suggested that the zeros in diastolic, glucose, triceps, insulin and bmi are actually missing values. Replace these zeros with NAs and describe (quantitatively, visually, and in words) the distribution of missing values in the data.

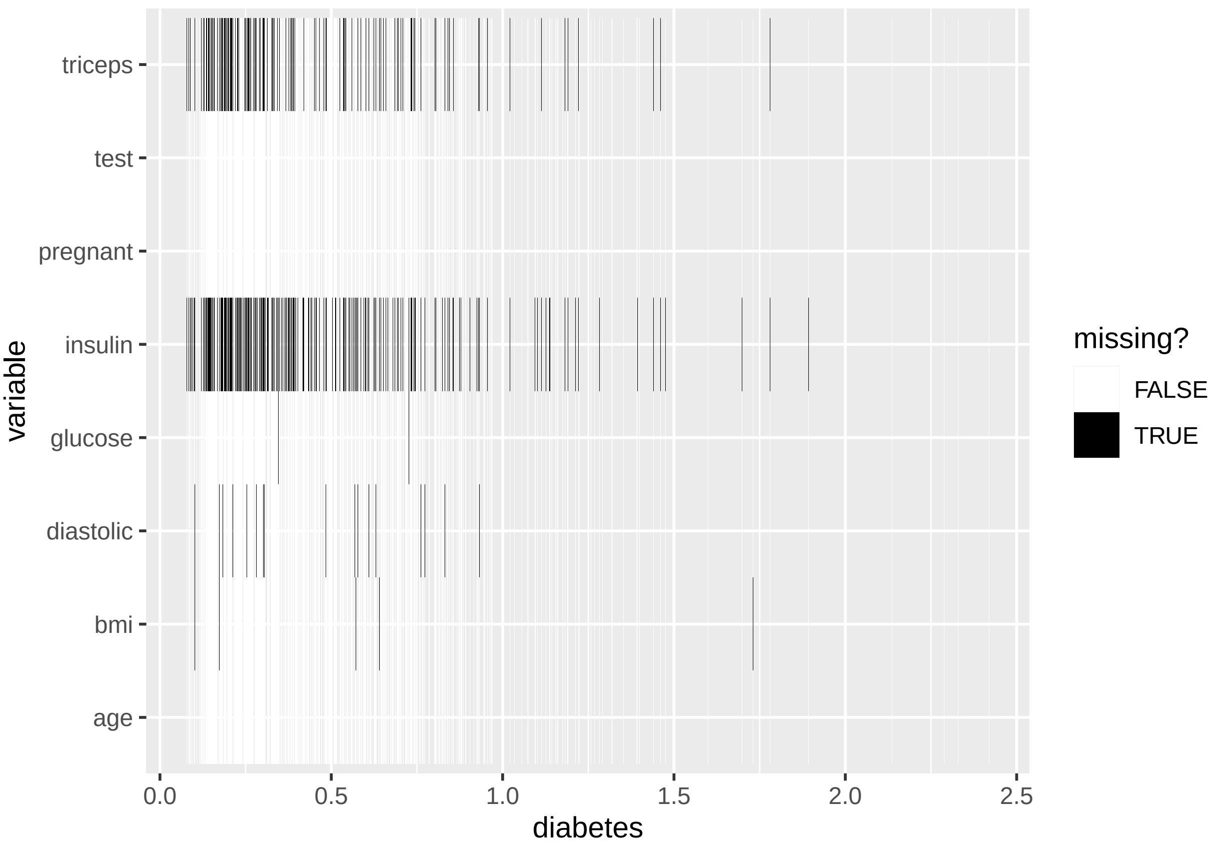
```
library(faraway)
?pima
```

```
## starting httpd help server ... done
```

```
df <- pima
#df[ == 0] <- NA
dfResponse <- df[, c(’pregnant’, ’diabetes’, ’test’, ’age’)]
dfReplaceNulls <-  df[, c(’diastolic’, ’glucose’, ’triceps’, ’insulin’,  ’bmi’)]

dfReplaceNulls[dfReplaceNulls == 0] <- NA

dfNulls <- cbind(dfReplaceNulls, dfResponse)

head(dfNulls)
```

```
##   diastolic glucose triceps insulin  bmi pregnant diabetes test age
## 1        72     148      35      NA 33.6        6    0.627    1  50
## 2        66      85      29      NA 26.6        1    0.351    0  31
## 3        64     183      NA      NA 23.3        8    0.672    1  32
## 4        66      89      23      94 28.1        1    0.167    0  21
## 5        40     137      35     168 43.1        0    2.288    1  33
## 6        74     116      NA      NA 25.6        5    0.201    0  30
```

```
# By Diabetes
dfNulls_long <- gather(dfNulls, variable, value, -diabetes)

qplot(diabetes, variable, data = dfNulls_long, geom = "tile",
  fill = is.na(value)) +
  scale_fill_manual("missing?",
  values = c(‘TRUE‘ = "black", ‘FALSE‘ = "white")) +
  theme(axis.text.x = element_text(angle = 0))
```



```r
sapply(dfNulls, function(x) sum(is.na(x)))
```

```
## diastolic   glucose   triceps   insulin       bmi  pregnant  diabetes      test 
##        35         5       227       374        11         0         0         0 
##       age 
##         0
```

It looks as diabetes pedigree increases it is less likely to have missing values in insulin concentration, tricep fold thickness, and diastolic pressure.

There are 652 values missing, 35, 5, 227, 374, and 11 from diastolic, glucose, triceps, insulin, and BMI respectively.

(b) (1 point) Suggest, in the context of the study, a mechanism such that the missing values in diastolic might be considered missing completely at random.

One mechanisim for missingness completely at random would be the diastolic measurement tool losing the ability to record at random times.

(c) (1 point) Suggest, in the context of the study, another mechanism such that the missing values in diastolic might be considered missing not at random.

Mising diastolic not at random would occur when the measurement device has a recording error at diastolic values over a certain threshold value.

(d) (2 points) Fit a linear model with diastolic as the response and the other variables as predictors. Summarize the fit.

```
fit <- lm(diastolic ~ . , data = dfNulls)

summary(fit)
```

```
##
## Call:
## lm(formula = diastolic ~ ., data = dfNulls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.420  -6.956  -0.604   7.432  29.268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.004185   4.043536  10.141  < 2e-16 ***
## glucose      0.047134   0.025848   1.824 0.069003 .
## triceps     -0.005719   0.074506  -0.077 0.938851
## insulin     -0.008268   0.006027  -1.372 0.170913
## bmi          0.532806   0.112798   4.724 3.26e-06 ***
## pregnant     0.183487   0.247575   0.741 0.459064
## diabetes    -3.213760   1.722406  -1.866 0.062826 .
## test         0.047652   1.508849   0.032 0.974822
## age          0.284048   0.081494   3.485 0.000548 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.38 on 383 degrees of freedom
##   (376 observations deleted due to missingness)
## Multiple R-squared:  0.1882, Adjusted R-squared:  0.1712
## F-statistic:  11.1 on 8 and 383 DF,  p-value: 3.94e-14
```

(e) (2 points) Use mean value imputation for the missing values, and refit the model. Compare the resulting estimates to the estimates from the previous fit: are the coefficient estimates similar, or do they differ substantially?

```
#copy df
dfMeanImputation <- dfNulls

#get col means / names
dfMeans <- colMeans(dfMeanImputation, na.rm = T)
names <- colnames(dfMeanImputation)

# replace nulls with column means
i <- 1
for (e in names){

  dfMeanImputation[e][is.na(dfMeanImputation[e])] <- mean(dfMeans[i], na.rm = TRUE)
  i <- i + 1

}

# REFIT MODEL

fitMeanImputation <- lm(diastolic ~ . , data = dfMeanImputation)

summary(fitMeanImputation)
```

```
##
## Call:
## lm(formula = diastolic ~ ., data = dfMeanImputation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.879  -6.599  -0.694   6.369  56.998
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.205265   2.620456  16.488  < 2e-16 ***
## glucose      0.048453   0.016310   2.971  0.00306 **
## triceps      0.006457   0.054022   0.120  0.90489
## insulin     -0.007388   0.005139  -1.438  0.15095
## bmi          0.476441   0.071163   6.695 4.19e-11 ***
## pregnant     0.157970   0.141327   1.118  0.26402
## diabetes    -2.127135   1.221251  -1.742  0.08195 .
## test        -0.868070   1.002583  -0.866  0.38686
## age          0.285792   0.041421   6.900 1.10e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.92 on 759 degrees of freedom
## Multiple R-squared:  0.1938, Adjusted R-squared:  0.1853
## F-statistic:  22.8 on 8 and 759 DF,  p-value: < 2.2e-16
```

The estimates compared to the null-excluded model are mostly higher, but very similar. The biggest difference is the standard error is lower because there are 760 degrees of freedom compared to 384. This comes from the ~650 null values replaced by mean values.

(f) (2 points) Use multiple imputation to handle missing values and fit the same model again. Compare the resulting estimates to the estimates from the previous two models (with no imputation, and with mean imputation): are the coefficient estimates similar, or do they differ substantially?

```
set.seed(123)

n.imp <- 25 # Set the number of imputed datasets
chimp <- amelia(dfNulls, m=n.imp, p2s=0)
```

```r
betas <- matrix(0, nrow=n.imp, ncol=9)
ses <- matrix(0, nrow=n.imp, ncol=9)


for(i in 1:n.imp){
  newMod <- lm(diastolic ~ . , data=chimp$imputations[[i]])
  betas[i,] <- coef(newMod)
  ses[i,] <- coef(summary(newMod))[,2]
}


mi.meld(q=betas, se=ses)
```

```
## $q.mi
##          [,1]       [,2]        [,3]         [,4]      [,5]      [,6]      [,7]
## [1,] 39.98622 0.06154447 -0.03572059 -0.008560325 0.5517072 0.1445007 -2.087615
##            [,8]      [,9]
## [1,] -0.9722926 0.3028191
##
## $se.mi
##          [,1]       [,2]       [,3]        [,4]       [,5]      [,6]     [,7]
## [1,] 2.987596 0.02097666 0.06973579 0.006529058 0.09218245 0.1543267 1.275688
##          [,8]      [,9]
## [1,] 1.055991 0.0453438
```

The coefficients between the mean-replacement model and the linear regression-replacement model are similar, but both the coefficients and standard error are lower in the linear regression model. This suggests that the data has outliers in the higher values, and drags up the mean value for each column.