ST517-HW6

Nora Quick

1. The **pima** dataset contains information on 768 adult female Pima Indians living near Phoenix.

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

pregnant <- pima$pregnant
glucose <- pima$glucose
diastolic <- pima$diastolic
triceps <- pima$triceps
insulin <- pima$insulin
bmi <- pima$bmi
diabetes <- pima$diabetes
age <- pima$age
test <- pima$test
```

   (a) 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.

```
diastolic[diastolic == 0] <- NA
glucose[glucose == 0] <- NA
triceps[triceps == 0] <- NA
insulin[insulin == 0] <- NA
bmi[bmi == 0] <- NA

summary(diastolic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   24.00   64.00   72.00   72.41   80.00  122.00      35
```

```
summary(glucose)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    44.0    99.0   117.0   121.7   141.0   199.0       5
```

```
summary(triceps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    7.00   22.00   29.00   29.15   36.00   99.00     227
```

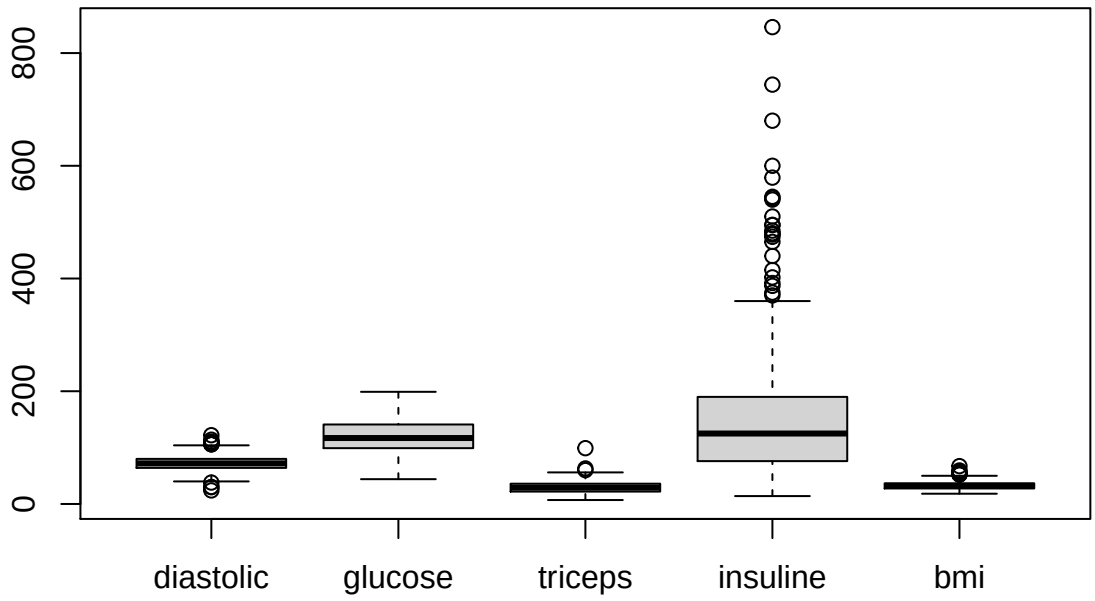
```r
summary(insulin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   14.00   76.25  125.00  155.55  190.00  846.00     374
```

```r
summary(bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.20   27.50   32.30   32.46   36.60   67.10      11
```

```r
boxplot(diastolic, glucose, triceps, insulin, bmi, names = c("diastolic", "glucose", "triceps", "insuli
```



As we can see from the summaries and boxplot most of the variables are normally distributed (diastolic, glucose, triceps, bmi). The only variable that is skewed is insuline which is right skewed.

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

While it is required for all medical visits to get blood pressure checked there could be missing values because a nurse or doctor forgot to check the woman's blood pressure. In other words, human error.

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

Age (or any of the other factors) could cause a woman to not want her blood pressure checked. For example, if the woman is much older and doesn't have the best health then they ask that blood pressure isn't checked at that time.

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

```
pima$test <- as.factor(ifelse(pima$test == 1, "positive", "negative"))
fit <- lm(diastolic ~ pregnant + glucose + insulin + bmi + diabetes + age + test, data = pima)

summary(fit)
```

```
##
## Call:
## lm(formula = diastolic ~ pregnant + glucose + insulin + bmi +
##     diabetes + age + test, data = pima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.239  -5.479   1.833   9.404  62.323
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.031343   3.938626   7.625 7.30e-14 ***
## pregnant      0.203131   0.232875   0.872  0.38333
## glucose       0.041505   0.024775   1.675  0.09429 .
## insulin       0.006233   0.006163   1.011  0.31217
## bmi           0.698737   0.087787   7.959 6.28e-15 ***
## diabetes     -0.071402   2.028444  -0.035  0.97193
## age           0.363745   0.068090   5.342 1.22e-07 ***
## testpositive -4.682499   1.622446  -2.886  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.99 on 760 degrees of freedom
## Multiple R-squared:  0.1443, Adjusted R-squared:  0.1365
## F-statistic: 18.32 on 7 and 760 DF,  p-value: < 2.2e-16
```

With a null hypothesis of everything having the same (no) effect on a womans blood pressure we can reject the mull hypothesis.

The overall p-value is quite small (2.2e-16) so we will conclude that the different effects on the body all effect blood pressure in different ways. Specifically, bmi, age, and testing positive for signs of diabetes can all effect the blood pressure of a woman.

(e) 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?

```
pima_Mean <- pima

pima_Mean$insulin[is.na(pima$insulin)] <- mean(pima$insulin, na.rm=TRUE)
pima_Mean$glucose[is.na(pima$glucose)] <- mean(pima$glucose, na.rm=TRUE)
```

```r
pima_Mean$triceps[is.na(pima$triceps)] <- mean(pima$triceps, na.rm=TRUE)
pima_Mean$bmi[is.na(pima$bmi)] <- mean(pima$bmi, na.rm=TRUE)

fit <- lm(diastolic ~ pregnant + glucose + insulin + bmi + diabetes + age + test, data = pima_Mean)

summary(fit)
```

```
##
## Call:
## lm(formula = diastolic ~ pregnant + glucose + insulin + bmi +
##     diabetes + age + test, data = pima_Mean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.239  -5.479   1.833   9.404  62.323
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.031343   3.938626   7.625 7.30e-14 ***
## pregnant      0.203131   0.232875   0.872  0.38333
## glucose       0.041505   0.024775   1.675  0.09429 .
## insulin       0.006233   0.006163   1.011  0.31217
## bmi           0.698737   0.087787   7.959 6.28e-15 ***
## diabetes     -0.071402   2.028444  -0.035  0.97193
## age           0.363745   0.068090   5.342 1.22e-07 ***
## testpositive -4.682499   1.622446  -2.886  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.99 on 760 degrees of freedom
## Multiple R-squared:  0.1443, Adjusted R-squared:  0.1365
## F-statistic: 18.32 on 7 and 760 DF,  p-value: < 2.2e-16
```

With a null hypothesis of everything having the same (no) effect on a womans blood pressure we can reject the mull hypothesis.

In part (d) we have three different factors that effect blood pressure (bmi, age, and testing positive for signs of diabetes). Here we can see that those three are still the significant. In fact, they are all the same outcome in both part (d) and (e).

(f) 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?

```r
set.seed(1234)
n.imp <- 50

pima$diastolic[pima$diastolic == 0] <- NA
pima$glucose[pima$glucose == 0] <- NA
pima$triceps[pima$triceps == 0] <- NA
pima$insulin[pima$insulin == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA
```

```r
new_pima <- amelia(pima, m = n.imp, p2s = 0, idvars = "test")

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

for(i in 1:n.imp){
   new_mod <- lm(diastolic ~ pregnant + glucose + triceps + insulin + bmi + diabetes + age + test, data
   betas[i,] <- coef(new_mod)
   ses[i,] <- coef(summary(new_mod))[,2]
}

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

```
## $q.mi
##          [,1]      [,2]       [,3]        [,4]         [,5]      [,6]      [,7]
## [1,] 40.10093 0.1407406 0.06284053 -0.02955985 -0.009331859 0.5400033 -1.995045
##           [,8]       [,9]
## [1,] 0.3024994 -0.8775261
##
## $se.mi
##          [,1]      [,2]       [,3]       [,4]        [,5]       [,6]     [,7]
## [1,] 2.943275 0.1524555 0.01946332 0.06180605 0.005371583 0.08667134 1.271466
##            [,8]     [,9]
## [1,] 0.04455391 1.056364
```

Comparing this model to the models above with this one I would conclude that the variables are similar. Going through each of the betas and standard errors the new model varies up and down from the old models a bit, however, they are not what I would classify as substantial.