Module 6 Lab Submission

Nora Quick

Again we consider the Brain size data in the data set `case0902` from the `Sleuth3` library. You can read more about this data set by viewing the help file:

```
help(case0902)
head(case0902)
```

```
##            Species  Brain    Body Gestation Litter
## 1         Aardvark    9.6    2.20        31    5.0
## 2         Acouchis    9.9    0.78        98    1.2
## 3 African elephant 4480.0 2800.00       655    1.0
## 4          Agoutis   20.3    2.80       104    1.3
## 5        Axis deer  219.0   89.00       218    1.0
## 6           Badger   53.0    6.00        60    2.2
```

In the previous Module, we considered a model using the log scale for all four variables, where the goal is to model log-brain-size as a function of log-gestation, log-littersize, and log-body-size. Here, we are going to arbitrarily set some of those predictor values to be missing:

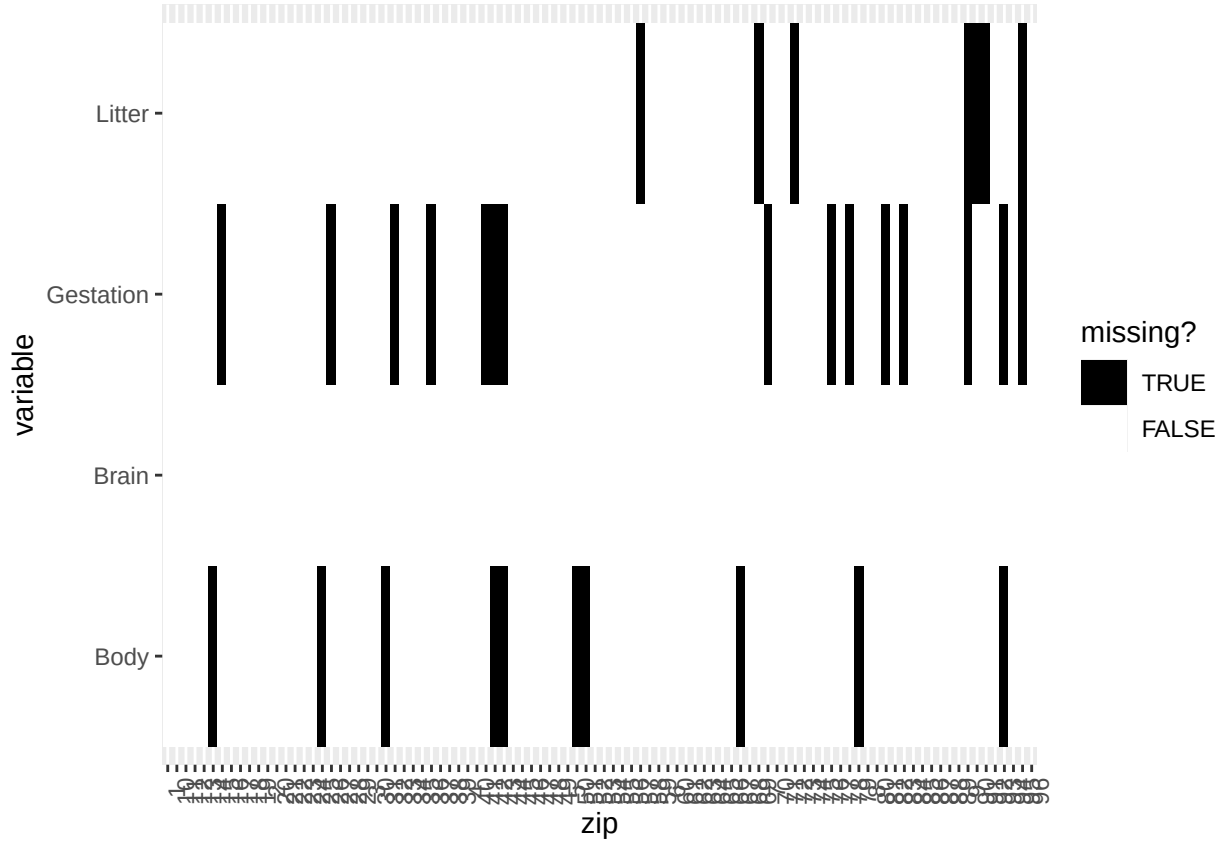
```
set.seed(123)
brainMiss <- case0902
n <- nrow(brainMiss)
brainMiss$Body[sample(1:n, 10, replace=FALSE)] <- NA
brainMiss$Litter[sample(1:n, 7, replace=FALSE)] <- NA
brainMiss$Gestation[sample(1:n, 15, replace=FALSE)] <- NA
```

1. **Create a visual representation of the missingness pattern in this data, using one of the methods demonstrated in the lab example.**

```
brainMiss$zip <- rownames(brainMiss)
#head(brainMiss)

brainmiss_long <- gather(brainMiss, variable, value, -zip, -Species)
# head(brainmiss_long)

qplot(zip, variable, data = brainmiss_long,
      geom = "tile",
      fill = is.na(value)) +
  scale_fill_manual("missing?",
                    values=c(`TRUE`="black", `FALSE`="white")) +
  theme(axis.text.x = element_text(angle=90))
```



2. **Use the `amelia` function to perform multiple imputation for these missing values, based on `m=50` imputed datasets.**

```
set.seed(123)
brainMiss <- case0902
n <- nrow(brainMiss)
brainMiss$Body[sample(1:n, 10, replace=FALSE)] <- NA
brainMiss$Litter[sample(1:n, 7, replace=FALSE)] <- NA
brainMiss$Gestation[sample(1:n, 15, replace=FALSE)] <- NA

n.imp <- 50
brainImp <- amelia(brainMiss, m=n.imp, p2s=0, idvars="Species")

names(brainImp)
```

```
##  [1] "imputations" "m"           "missMatrix"  "overvalues"  "theta"
##  [6] "mu"          "covMatrices" "code"        "message"     "iterHist"
## [11] "arguments"   "orig.vars"
```

3. **Fit models using each of the imputed datasets**

```
bounds_mat <- matrix(c(3,4,5,0.01,10,1,3000,700,10), ncol=3)
bounds_mat
```

```
##      [,1]  [,2] [,3]
## [1,]    3  0.01 3000
## [2,]    4 10.00  700
## [3,]    5  1.00   10
```

```
n.imp <- 50
brain_imput <- amelia(brainMiss, m = n.imp, p2s = 0, idvars = 'Species', bounds=bounds_mat)

betas <- matrix(0, nrow=n.imp, ncol=4)
ses <- matrix(0, nrow=n.imp, ncol=4)

for(i in 1:n.imp){
  newMod <- lm(log(Brain) ~ log(Gestation) + log(Litter) + log(Body), data=brain_imput$imputations[[i]])
  betas[i,] <- coef(newMod)
  ses[i,] <- coef(summary(newMod))[,2]
}
```

4. **Use the `mi.meld` function to find the multiple imputation estimates of the coefficients.**

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

```
## $q.mi
##           [,1]      [,2]       [,3]      [,4]
## [1,] -1.084791 0.8448951 -0.2050669 0.4576729
##
## $se.mi
##          [,1]      [,2]      [,3]      [,4]
## [1,] 1.160628 0.2412775 0.2135992 0.0540458
```

Recall the coefficients from the model constructed on the full data (with no missing values):

```
origMod <- lm(log(Brain) ~ log(Gestation) + log(Litter) + log(Body), data=case0902)
origMod
```

```
##
## Call:
## lm(formula = log(Brain) ~ log(Gestation) + log(Litter) + log(Body),
##     data = case0902)
##
## Coefficients:
##    (Intercept)  log(Gestation)     log(Litter)       log(Body)
##         0.8548          0.4179         -0.3101          0.5751
```

5. **How do the coefficient estimates from the multiple imputation compare to the coefficient estimates on the original full data?**

We can see that Brain and Gestation in the coefficient estimates from the multiple imputation are smaller than the original full data. However, we can see that Litter and Body are both very slightly larger. The missing data makes the overall value of Brain a negative.