Pairs Assignment 03/07

Kerem Tuncer and Nikhil Gopal

Question 17.11

For this question, we will be looking at our past analysis from 17.1 with the earnings dataset.

```
library(rstanarm)
```

```
## Loading required package: Rcpp

## This is rstanarm version 2.21.1

## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!

## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.

## - For execution on a local, multicore CPU with excess RAM we recommend calling

##   options(mc.cores = parallel::detectCores())
```

```
data <- read.csv("https://raw.githubusercontent.com/avehtari/ROS-Examples/master/Earnings/data/earnings
```

```
sum(is.na(data$height))
```

```
## [1] 0
```

```
sum(is.na(data$weight))
```

```
## [1] 27
```

```
sum(is.na(data$male))
```

```
## [1] 0
```

We have 27 missing observations in our dependent variable. Now, we will do multiple imputation using simple random imputation. Here is the code for the random imputation.

```
random_imp <- function(a) {
  missing <- is.na(a)
  n_missing <- sum(missing)
  a_obs <- a[!missing]
  imputed <- a
  imputed[missing] <- sample(a_obs, n_missing)
  imputed
}
```

Then, we will use a for-loop to run 10 iterations of our original regression. The loop will extract the coefficient estimates and their standard errors. In the end, we will have a dataframe with 10 rows and 6 columns, where each row gives us the three coefficient estimates (including the intercept) and their se.

```
vector <- matrix(ncol = 6, nrow = 10)

for (i in 1:10){
data$weight_imp <- random_imp(data$weight)
fit <- stan_glm(weight_imp ~ height+factor(male) , data = data, refresh=0)
vector[i,] <- c(fit$coefficients, fit$ses)
}

imputed_output <- as.data.frame(vector)
imputed_output
```

```
##           V1       V2       V3       V4        V5       V6
## 1  -103.62954 3.835549 12.00389 16.05735 0.2515649 1.953466
## 2   -97.51509 3.741032 12.45384 15.86451 0.2463870 1.955347
## 3   -98.91570 3.764363 12.31341 16.23764 0.2519942 2.037546
## 4  -102.32987 3.817016 12.01193 16.09392 0.2473544 1.897999
## 5  -102.70166 3.824736 11.87193 15.83297 0.2456835 2.013608
## 6  -105.63552 3.871786 11.50752 16.92912 0.2626480 2.039460
## 7  -100.61808 3.787120 12.37082 15.68654 0.2456800 2.053250
## 8  -103.35790 3.830281 12.25900 16.24704 0.2494941 1.942858
## 9  -103.41077 3.835749 11.90625 16.35749 0.2534965 1.938826
## 10 -102.77480 3.823530 11.95900 15.95662 0.2445939 1.990030
```

The V2 variable will be the height coefficient and the V3 variable will be the male coefficient. According to page 326 of the textbook, the average of these 10 iterations will give an overall point estimate.

```
average_beta_height <- mean(imputed_output$V2)
average_beta_male <- mean(imputed_output$V3)
```

Then, we will use the formula on page 326 to get the overall standard errors for height and male.

Let's first do it for height. The V5 variable is its standard error.

```
within_variance <- sum(imputed_output$V5^2)/10
between_variance <- sum((imputed_output$V2-average_beta_height)^2)/9

se_b_height <- sqrt(within_variance + (1+1/10)*between_variance)
```

Now, let's first do it for male. The V6 variable is its standard error.

```
within_variance <- sum(imputed_output$V6^2)/10
between_variance <- sum((imputed_output$V3-average_beta_male)^2)/9

se_b_male <- sqrt(within_variance + (1+1/10)*between_variance)
```

Finally, let's compare it to our original regression without imputation.

```
fit_1 <- stan_glm(weight ~ height+factor(male), data = data, refresh=0, )
fit_1$coefficients
```

```
##   (Intercept)        height factor(male)1
##   -106.957895      3.885779     11.883355
```

```
fit_1$ses
```

```
##   (Intercept)        height factor(male)1
##    16.4149138     0.2548495     2.0019723
```

```
average_beta_height
```

```
## [1] 3.813116
```

```
average_beta_male
```

```
## [1] 12.06576
```

```
se_b_height
```

```
## [1] 0.253164
```

```
se_b_male
```

```
## [1] 2.005385
```

There was a slight decrease in our estimate of the height variable. Meanwhile, there was a slight increase in our estimate of the male variable. Also, our new standard errors were higher than the original, which means that our original analysis without imputation had a higher certainty. Given that there were only 27 missing observations, it makes sense that the values did not change much.