# exercises_week11

Lindsey Greenhill

4/14/2021

## Question 25.2

### Part a

```r
# using built in iris data set. filtering to one species and selecting petal
# length and width. Petal length is x petal width is y.

iris_2 <- iris %>%
  filter(Species == "setosa") %>%
  select(Petal.Length, Petal.Width)


# MAR data. going to delete 25 out of 50 obvervations. Going to pick on the the
# higher petal widths. Creating rbinom vectors below with higher and lower
# probabilities of deletion

high_del <- rbinom(n = 25, size = 1, prob = 0.7)
low_del <- rbinom(n = 25, size = 1, prob = 0.3)

del_vec <- c(high_del, low_del)

# deleting based off of vectors above. Deleting based on petal.width (out
# outcome variable)

iris_available <- iris_2 %>%
  arrange(desc(Petal.Width)) %>%
  mutate(del_col = del_vec) %>%
  mutate(Petal.Length = ifelse(del_col == 1, NA, Petal.Length))
```

### Part b

In the code below, I perform a regression of petal length on width, or x on y for both the full data (mod_complete) frame and the available data frame (mod_available). The models are relatively consistent. The constant for the full model is 1.328 and the constant for the available model is 1.38. The coefficient for petal.width for the full model is .546 and the coefficient for petal width for the available data is .327. It makes sense that the two models are not drastically different because the missing data is random with respext to petal length.

```r
# creating regression with full data but with y as predictor on x

mod_complete <- lm(Petal.Length ~ Petal.Width, data = iris_2)

mod_available <- lm(Petal.Length ~ Petal.Width, data = iris_available)

# the models are relatively similar because we didn't have any missing x values.
# our deletion was random with regard to petal length.

stargazer(mod_complete, mod_available, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Apr 20, 2021 - 17:13:36

Table 1

| | *Dependent variable:* | |
|---|---|---|
| | Petal.Length | |
| | (1) | (2) |
| Petal.Width | $0.546^{**}$ | 0.327 |
| | (0.224) | (0.354) |
| Constant | $1.328^{***}$ | $1.380^{***}$ |
| | (0.060) | (0.086) |
| Observations | 50 | 28 |
| $R^2$ | 0.110 | 0.032 |
| Adjusted $R^2$ | 0.091 | −0.006 |
| Residual Std. Error | 0.166 (df = 48) | 0.178 (df = 26) |
| F Statistic | $5.931^{**}$ (df = 1; 48) | 0.852 (df = 1; 26) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

## Part c

In the code below, I perform a regression of petal width on petal length, or y on x, for both the full data frame (mod_complete_c) and the available data frame (mod_available_c). The models are not consistent with each other. The constant for the full model is -.048 and the constant for the available model is .084. The coefficient for petal length for the full model is .201 and the coefficient for petal length for the available model .097, less than half of the first model. It makes sense that these models are inconsistent because the missing data is not random with respect to petal width.

```r
mod_complete_c <- lm(Petal.Width ~ Petal.Length, data = iris_2)

mod_available_c <- lm(Petal.Width ~ Petal.Length, data = iris_available)

# the models are significantly different with regards to both the coefficients
# and the estimates.

stargazer(mod_complete_c, mod_available_c, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Apr 20, 2021 - 17:13:36

Table 2

| | *Dependent variable:* | |
|---|---|---|
| | Petal.Width | |
| | (1) | (2) |
| Petal.Length | 0.201$^{**}$ | 0.097 |
| | (0.083) | (0.105) |
| Constant | −0.048 | 0.084 |
| | (0.122) | (0.154) |
| Observations | 50 | 28 |
| R$^2$ | 0.110 | 0.032 |
| Adjusted R$^2$ | 0.091 | −0.006 |
| Residual Std. Error | 0.100 (df = 48) | 0.097 (df = 26) |
| F Statistic | 5.931$^{**}$ (df = 1; 48) | 0.852 (df = 1; 26) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

## Part d

The new model, shown in the third column of the stargazer table, has very different estimates from both the complete and partial models in part c. These differences show that how you treat missing data can have large effects for your model.

```r
# using random imputation function from slides

random.imp.vec <- function(V)  {
  gone <- is.na(V)
  there <- V[!gone]
  V[gone] <- sample(x=there,size=sum(gone),replace=TRUE)
  return(V)
}

# creating the new data fram

df_d <- random.imp.vec(iris_available)

# creating the model with the imputed data

mod_d <- lm(Petal.Width ~ Petal.Length, data = df_d)

# comparing the models

stargazer(mod_complete_c, mod_available_c, mod_d,
          type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu

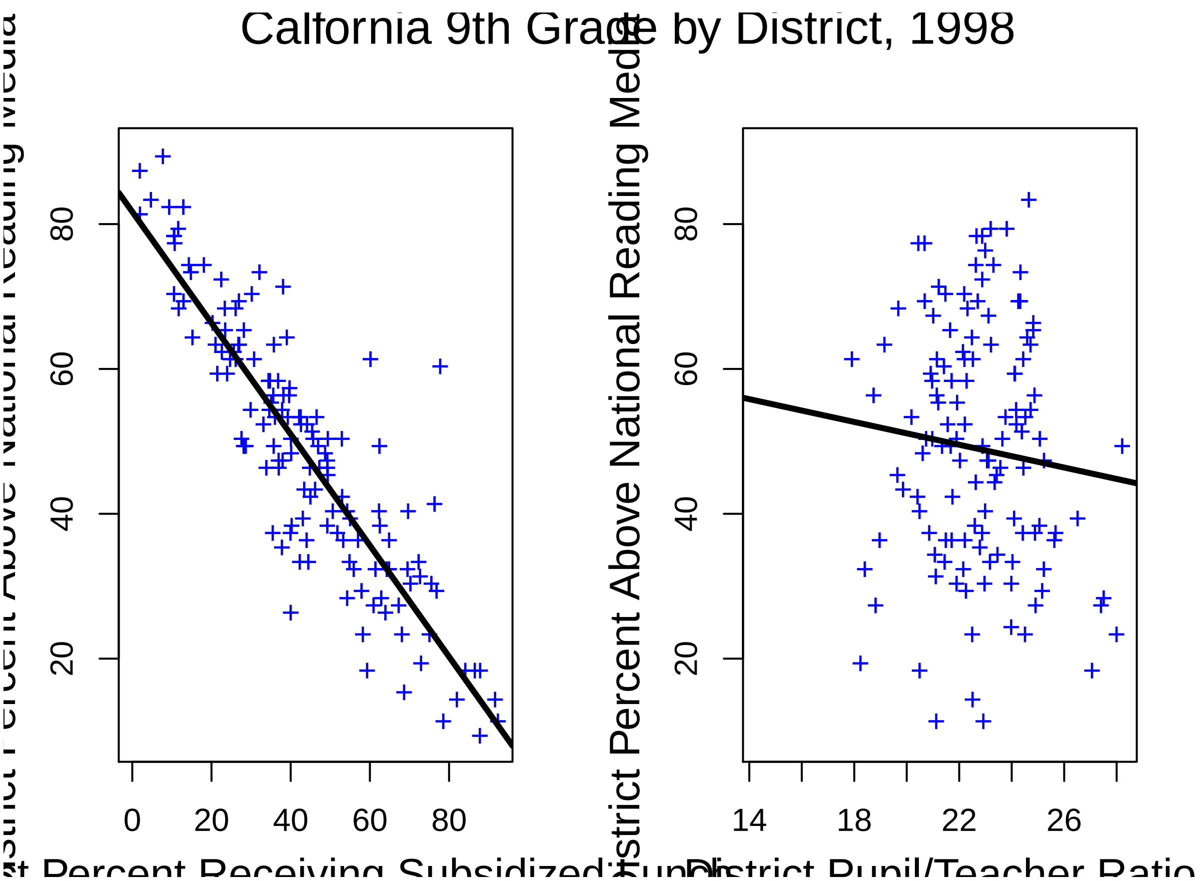
% Date and time: Tue, Apr 20, 2021 - 17:13:36

Table 3

| | *Dependent variable:* | | |
|---|---|---|---|
| | Petal.Width | | |
| | (1) | (2) | (3) |
| Petal.Length | 0.201$^{**}$ | 0.097 | −0.037 |
| | (0.083) | (0.105) | (0.026) |
| Constant | −0.048 | 0.084 | 0.286$^{***}$ |
| | (0.122) | (0.154) | (0.032) |
| Observations | 50 | 28 | 50 |
| R$^2$ | 0.110 | 0.032 | 0.041 |
| Adjusted R$^2$ | 0.091 | −0.006 | 0.021 |
| Residual Std. Error | 0.100 (df = 48) | 0.097 (df = 26) | 0.104 (df = 48) |
| F Statistic | 5.931$^{**}$ (df = 1; 48) | 0.852 (df = 1; 26) | 2.065 (df = 1; 48) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

# STAR data

```r
star98.missing <- read.table("star98.missing.dat.txt",header=TRUE)
par(mfrow=c(1,2),mar=c(3,3,3,3))
plot(star98.missing$SUBSIDIZED.LUNCH,star98.missing$READING.ABOVE.50,pch="+",col="blue")
abline(lm(star98.missing$READING.ABOVE.50~star98.missing$SUBSIDIZED.LUNCH),lwd=3)
mtext(side=1,cex=1.3,line=2.5,"District Percent Receiving Subsidized Lunch")
mtext(side=2,cex=1.3,line=2.5,"District Percent Above National Reading Median")
plot(star98.missing$PTRATIO,star98.missing$READING.ABOVE.50,pch="+",col="blue")
abline(lm(star98.missing$READING.ABOVE.50~star98.missing$PTRATIO),lwd=3)
mtext(side=1,cex=1.3,line=2.5,"District Pupil/Teacher Ratio")
mtext(side=2,cex=1.3,line=2.5,"District Percent Above National Reading Median")
mtext(side=3,cex=1.5,outer=TRUE,line=-1,"Calfornia 9th Grade by District, 1998")
```



```r
summary(star98.missing)
```

```
##  SUBSIDIZED.LUNCH     PTRATIO      READING.ABOVE.50
##  Min.   : 0.2653   Min.   :14.32   Min.   : 9.00   
##  1st Qu.:26.1143   1st Qu.:21.15   1st Qu.:36.00   
##  Median :40.0598   Median :22.59   Median :49.00   
##  Mean   :41.8263   Mean   :22.54   Mean   :49.07   
##  3rd Qu.:56.3312   3rd Qu.:24.15   3rd Qu.:63.00   
##  Max.   :92.3345   Max.   :28.21   Max.   :90.00   
##  NA's   :90        NA's   :104     NA's   :106
```

```r
# how to check if there is a pattern? for there is subsidized lunch
```

### Part a

determine how much missing data there is and if there is a discernable pattern

Looking at the summary above there appears to be some missing data.

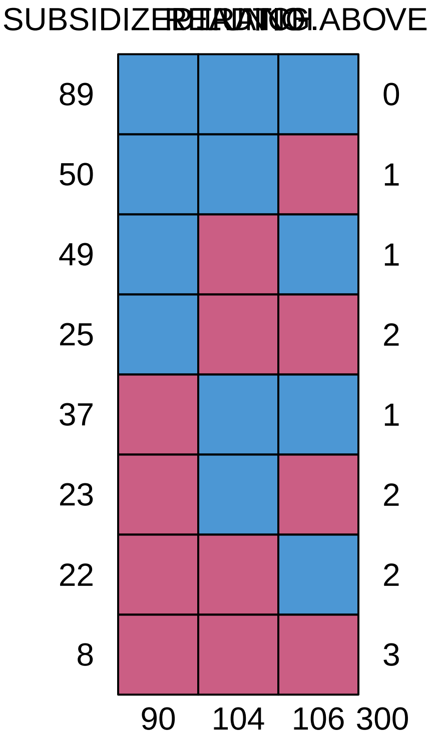
- SUBSIDIZED.LUNCH has 90 NA's
- PTRATION has 104 NA's
- READING.ABOVE.50 has 106 NA's

To see if there is any discernable pattern I used the md.pattern() function from the mice library (see output below).

From it we can tell that there are 89 complete samples. 50 samples that just miss the READING.ABOVE.50, 49 samples that just miss PTRATIO, 37 samples that just miss SUBSIDIZED.LUNCH, 25 samples that only have SUBSIZED LUNCH, 23 samples that only have PTRATIO, 22 samples that only have READING.ABOVE.50, and 8 samples that have no data.

```
# shows up the pattern of missing data. From it we can tell that there are 89
# complete samples. 50 samples that just miss the READING.ABOVE.50, 49 samples
# that just miss PTRATIO, 37 samples that just miss SUBSIDIZED.LUNCH, 25 samples
# that only have SUBSIZED LUNCH, 23 samples that only have PTRATIO, 22 samples
# that only have READING.ABOVE.50, and 8 samples that have no data.

md.pattern(star98.missing)
```



```
##    SUBSIDIZED.LUNCH PTRATIO READING.ABOVE.50
## 89                1       1                1   0
## 50                1       1                0   1
## 49                1       0                1   1
## 25                1       0                0   2
## 37                0       1                1   1
## 23                0       1                0   2
## 22                0       0                1   2
## 8                 0       0                0   3
##                  90     104              106 300
```

## Parts b and c

In the code below, I first create a case wise deletion model and then a mice model. The table below shows the results of the case wise deletion model on the left and the mice model on the right. The estimates for the intercept differ by about 5 points. The estimates for subsidized.lunch are almost equal, and the esttimates for ptratio differ by about .2. Overall, the models are relatively similar. The standard errors are a little bit smaller for the imputation model, so I am inclined to say it is the better model.

```
##
##  iter imp variable
##   1   1  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   1   2  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   1   3  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   1   4  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   1   5  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   2   1  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   2   2  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   2   3  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   2   4  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   2   5  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   3   1  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   3   2  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   3   3  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   3   4  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   3   5  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   4   1  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   4   2  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   4   3  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   4   4  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   4   5  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   5   1  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   5   2  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   5   3  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   5   4  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
##   5   5  SUBSIDIZED.LUNCH  PTRATIO  READING.ABOVE.50
```

```
## Warning: Use with(imp, lm(yourmodel).
```

| Estimate | Std. Error | term | estimate | std.error |
|---:|---:|:---:|---:|---:|
| 116.308829 | 8.62136932 | (Intercept) | 120.0884078 | 8.5109039 |
| -0.800961 | 0.03732617 | SUBSIDIZED.LUNCH | -0.7863431 | 0.0297244 |
| -1.510049 | 0.36842114 | PTRATIO | -1.6573803 | 0.3430673 |