HW9

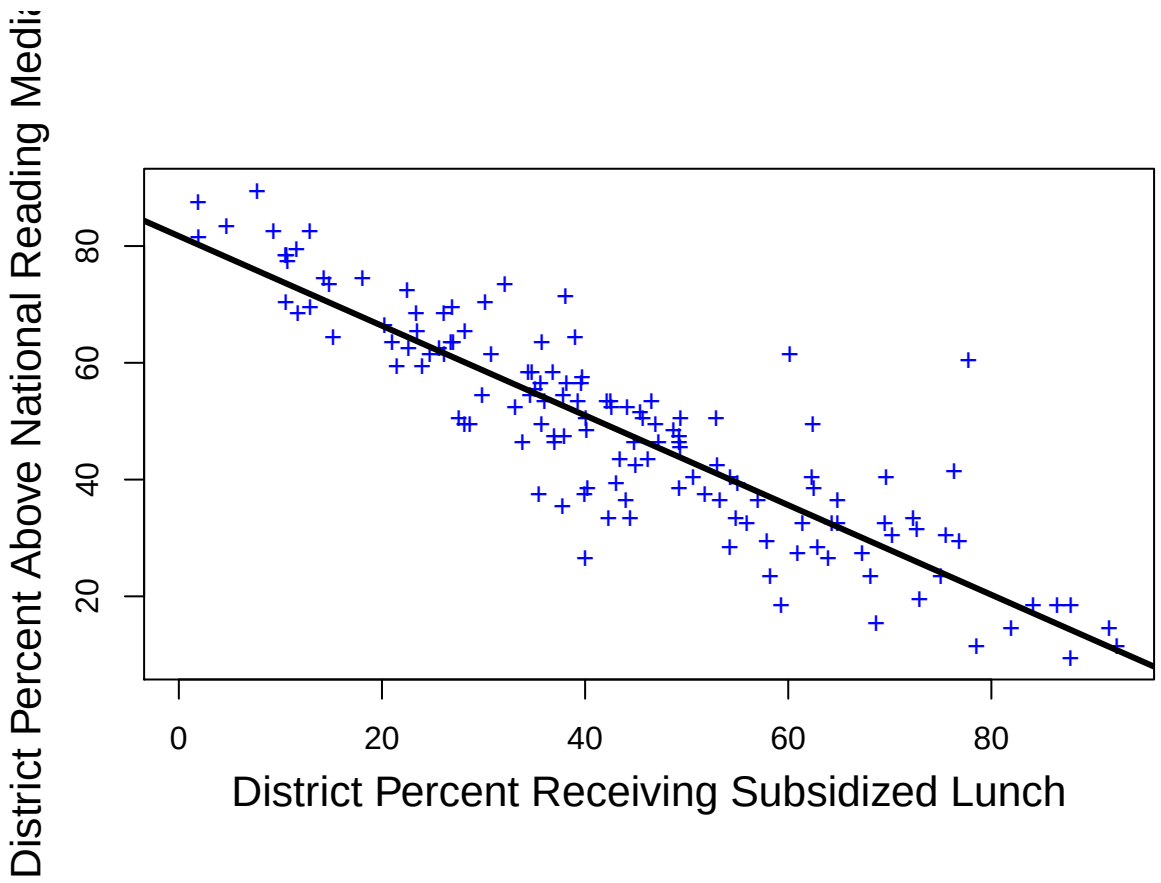
Avinash Ramu

November 13, 2016

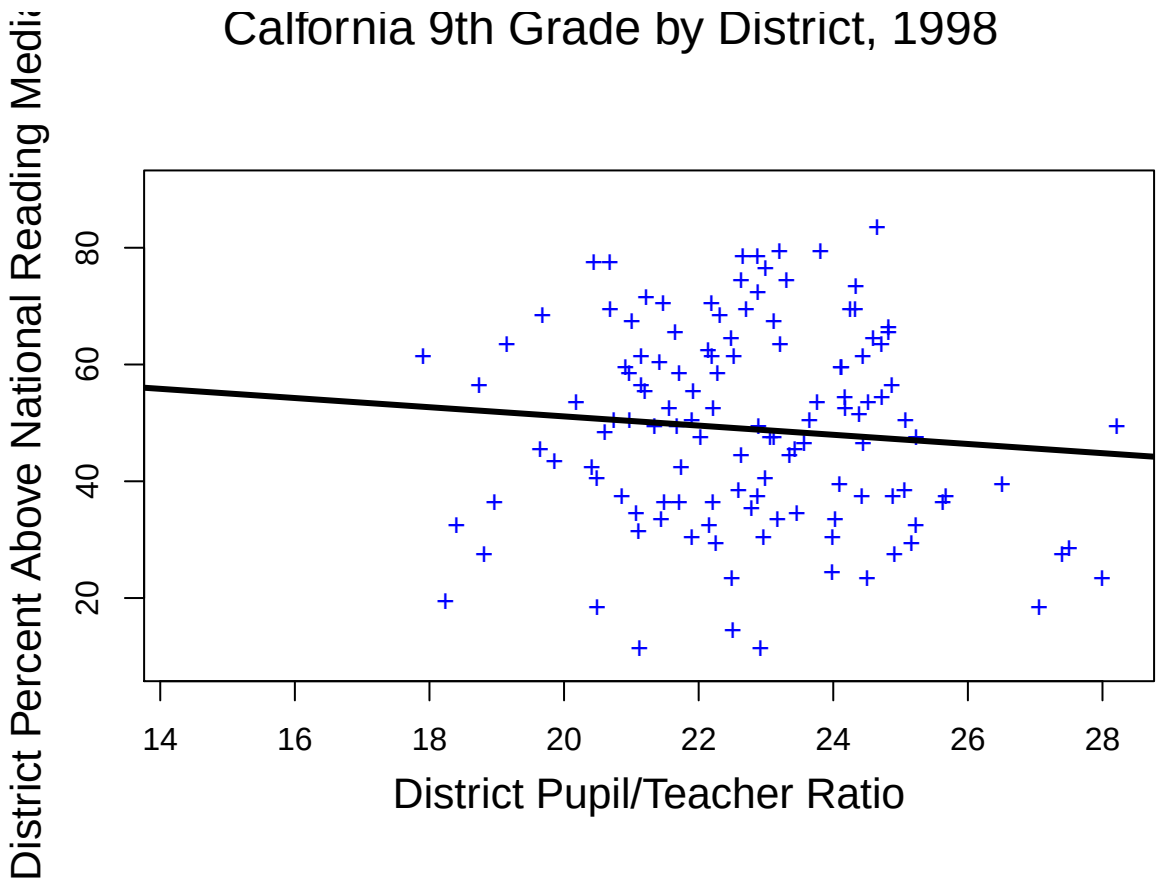
Q1.1

Determine how much missing data there is and if there is a discernable pattern. Now use mice to run a new model. Also run a model omitting cases with missing data. What differences do you observe? Which is better?

```
star98.missing <- read.table("star98.missing.dat",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", xlab = "", ylab
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", xlab = "", ylab = "")
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")
```




```
#Determine how much missing data there is:
library(mice)
```

```
## Loading required package: Rcpp
```

```
## mice 2.25 2015-11-09
```

```
sum(is.na(star98.missing))/prod(dim(star98.missing))
```

```
## [1] 0.330033
```

```
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
#hist(apply(apply(star98.missing, 2, is.na), 1, sum))
#Look at pattern of missingness, 0 implies missing 1 implies not missing
md.pattern(star98.missing)
```

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

```r
star98_omit <- na.omit(star98.missing)
lm1 <- lm(star98_omit$READING.ABOVE.50~star98_omit$SUBSIDIZED.LUNCH)
lm2 <- lm(star98_omit$READING.ABOVE.50~star98_omit$PTRATIO)
summary(lm1)
```

```
##
## Call:
## lm(formula = star98_omit$READING.ABOVE.50 ~ star98_omit$SUBSIDIZED.LUNCH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9576  -4.1594   0.4502   4.7060  26.7118
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   81.70887    1.90336   42.93   <2e-16 ***
## star98_omit$SUBSIDIZED.LUNCH -0.78843    0.04044  -19.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.569 on 87 degrees of freedom
## Multiple R-squared:  0.8138, Adjusted R-squared:  0.8116
## F-statistic: 380.1 on 1 and 87 DF,  p-value: < 2.2e-16
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = star98_omit$READING.ABOVE.50 ~ star98_omit$PTRATIO)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.294 -12.860   0.374  12.632  36.748
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          67.5097    20.8419   3.239   0.0017 **
```

```
## star98_omit$PTRATIO  -0.8623     0.9202  -0.937   0.3513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.45 on 87 degrees of freedom
## Multiple R-squared:  0.009993,   Adjusted R-squared:  -0.001387
## F-statistic: 0.8781 on 1 and 87 DF,  p-value: 0.3513
```

```
star98.mids1 <- lm.mids(star98.missing$READING.ABOVE.50~star98.missing$SUBSIDIZED.LUNCH,
                        data=star98.imp)
star98.mids2 <- lm.mids(star98.missing$READING.ABOVE.50~star98.missing$PTRATIO,
                        data=star98.imp)
summary(pool(star98.mids1))
```

```
##                                      est         se         t       df
## (Intercept)                     81.7062591 1.68392879  48.52121 134.0298
## star98.missing$SUBSIDIZED.LUNCH -0.7679826 0.03475518 -22.09693 134.0298
##                                 Pr(>|t|)      lo 95      hi 95 nmis
## (Intercept)                            0 78.3757481 85.0367701   NA
## star98.missing$SUBSIDIZED.LUNCH        0 -0.8367222 -0.6992431   NA
##                                       fmi lambda
## (Intercept)                     0.01459537      0
## star98.missing$SUBSIDIZED.LUNCH 0.01459537      0
```

```
summary(pool(star98.mids2))
```

```
##                               est        se         t      df     Pr(>|t|)
## (Intercept)            66.8507101 17.060768  3.918388 122.035 0.0001475341
## star98.missing$PTRATIO -0.7870485  0.747289 -1.053205 122.035 0.2943279505
##                            lo 95       hi 95 nmis        fmi lambda
## (Intercept)            33.077312 100.6241082   NA 0.01599552      0
## star98.missing$PTRATIO -2.266378   0.6922805   NA 0.01599552      0
```

A third of the cells have missing data. Only eight rows have all three columns not missing. The missing data appears to be pretty random across the columns with no immediately apparent pattern of missingness.

The inferences using the MICE data seem to be similar to the inferences from the data which omits missing data. This might be because our data is MCAR and we have enough observations after omiting. I do not observe a significant difference between the models with missing data and imputed data.

In the first model, the coefficient for SUBSIDIZED.LUNCH is significant and negative. This looks similar to what we'd expect from the plot. There's a decrease of -0.7679826 in reading units for every percent increase in subsidized lunch.

In the second model, the coefficient for PT ratio is not significant.

Q2

Explain what the following R does and why you would not want to do this.

```
mi <- function(data.mat) {
  for (i in 1:ncol(data.mat)) {
    if (sum(is.na(data.mat[,i])) > 0) {
```

```
      print(paste("column",i,"has missing data"))
      mean.col <- mean(data.mat[,i],na.rm=TRUE)
      for (j in 1:nrow(data.mat)) {
        if (is.na(data.mat[j,i]) ==TRUE) data.mat[j,i] <- mean.col
      }
    }
  }
  return(data.mat)
}
```

The code sets the missing values in a column to the mean value of the column. This underestimates the variance for that column. It also distorts relationships between variables by pulling the estimation of the correlation between them to zero.

Q3

Find an article in your literature that uses case-wise deletion. Discuss how you might replicate the model and improve the work.

Quantitative trait loci(QTL) are genetic loci that show association with a quantitative trait, for e.g gene expression. Studies that try to identify expression QTLs perform regressions between the genetic variation in an individual and the gene expression on a gene by gene basis. If the gene expression readout for a particular gene is unavailable or if genetic variation hasn't been typed at a nearby locus studies tend to exclude those individuals from the regression.

Genotype imputation can be used to infer the genetic variation for the samples which have missing genotypes. The patterns of genetic variation are shared in the population due to a phenomenon known as linkage disequilibrium. This affords us to make probabilistic guesses about missing genotypes. This can be used in the eQTL regressions and is pretty widely used now since we have a good idea of what normal genetic variation in humans looks like after large sequencing efforts.

One example of such a study in the early days of modern human genetics is http://www.nature.com/nature/journal/v430/n7001/full/nature02797.html where the authors state "We used PedStat21 to check for mendelian inconsistencies. This resulted in the removal of 815 genotypes at 237 distinct SNP markers." Imputing these genotypes could be a better approach though it might not affect the results significantly owing to the small fraction of missingness.