Missing Data Assignment - Marvin

Marvin Williams

12/11/2020

```
data <- read.csv("MissingDataAssignment.csv")
str(data)
```

```
## 'data.frame':    1000 obs. of  4 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ blues  : int  NA 1 1 0 1 0 1 1 1 1 ...
##  $ college: int  0 0 0 0 0 0 0 NA 0 NA ...
##  $ income : num  46449 46270 54615 26234 71102 ...
```

The first model is a linear regression which predicts income according to the respondent's group and whether or not the respondent graduated college. Our 2 predictor variables blues and college were both statistically significant predictors. With a coefficient of 10767, this tells us that for every unit increase, those with membership in the blue group is positively associated with a $10,767 increase in income. We also see a positive association with the college, with a coefficient of 8203.6. This tells us that those who attended college see an $8,203 increase in income for every unit increase.

```
model1<- glm(income ~ blues + college, data, family = "gaussian")
summary(model1)
```

```
##
## Call:
## glm(formula = income ~ blues + college, family = "gaussian",
##     data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -33677  -11015    -784   10081   52388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31755.9      729.4  43.537  < 2e-16 ***
## blues        10767.4     1061.4  10.145  < 2e-16 ***
## college       8203.6     1314.3   6.242 6.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 218192046)
##
##     Null deviance: 2.0567e+11  on 813  degrees of freedom
## Residual deviance: 1.7695e+11  on 811  degrees of freedom
##   (186 observations deleted due to missingness)
## AIC: 17945
```

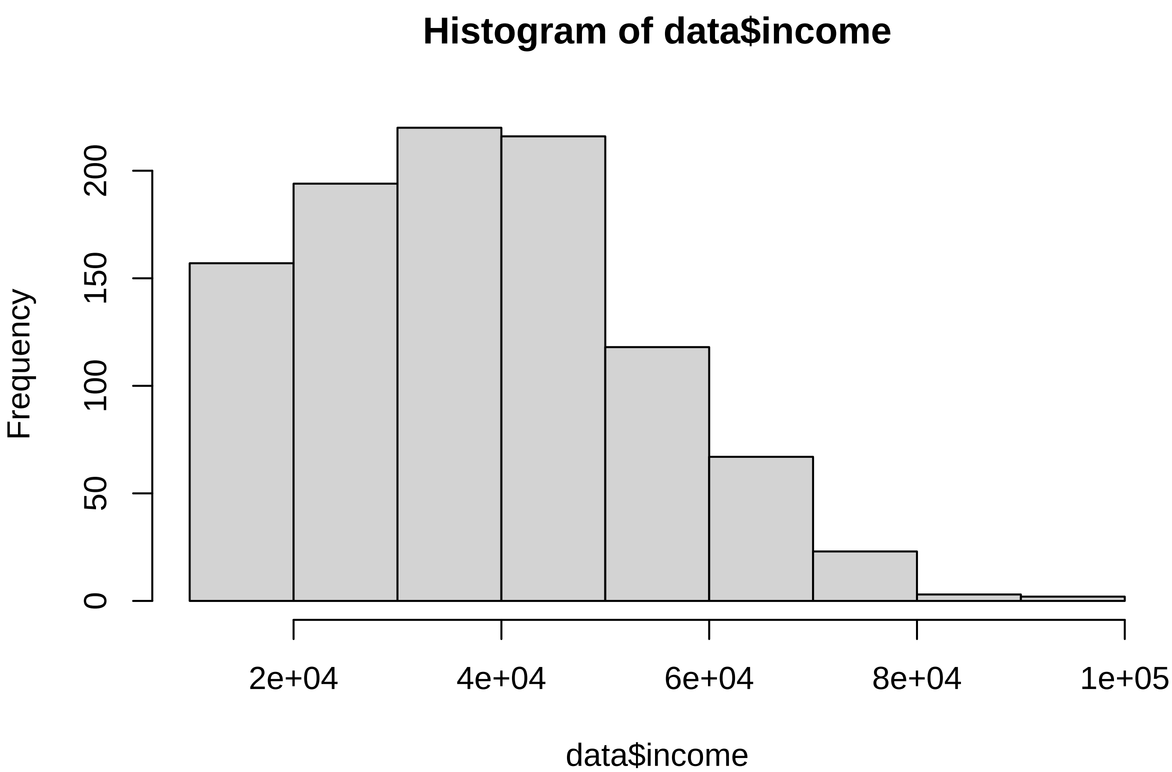
```
##
## Number of Fisher Scoring iterations: 2
```

Specifying the variables types

```
summary(data)
```

```
##        X              blues           college           income
##  Min.   :   1.0   Min.   :0.0000   Min.   :0.0000   Min.   :15000
##  1st Qu.: 250.8   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:25476
##  Median : 500.5   Median :0.0000   Median :0.0000   Median :36332
##  Mean   : 500.5   Mean   :0.3978   Mean   :0.1951   Mean   :37663
##  3rd Qu.: 750.2   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:48173
##  Max.   :1000.0   Max.   :1.0000   Max.   :1.0000   Max.   :93569
##                   NA's   :95       NA's   :98
```

```
id.vars <- c("X")
nom.vars <- names(data)[c(2,3)]
hist(data$income)
```



###

A histogram of the income variable shows an income distribution that is relatively similar to that of the actual income distribution. Logging wouldn't be necessary as this right-skew is expected.

```
data=apply_labels(data,
                  college= "Respondent a college Graduate?",
                  blues= "Is in Blue group?",
                  income= "Annual Household Income")
str(data)
```

```
## 'data.frame':    1000 obs. of  4 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ blues  :Class 'labelled' int  NA 1 1 0 1 0 1 1 1 1 ...
```

```
##    .. .. LABEL: Is in Blue group?
##  $ college:Class 'labelled' int  0 0 0 0 0 0 0 NA 0 NA ...
##    .. .. LABEL: Respondent a college Graduate?
##  $ income :Class 'labelled' num  46449 46270 54615 26234 71102 ...
##    .. .. LABEL: Annual Household Income
```

Creating labels for College variable

```
College = c(0,1)
var_lab(College)= "Is respondent a college Graduate?"
val_lab(College)= num_lab("
                          1 College Graduate
                          0 Not a College Graduate
                          ")
```

```
set.seed(914)
data.imp <-amelia(data,
                 m=5,
                 idvars = id.vars,
                 noms = nom.vars,
                 empr = 0,
                 emburn = c(25,100)
                 )
```

```
## -- Imputation 1 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  21 22 23 24 25
##
## -- Imputation 2 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  21 22 23 24 25
##
## -- Imputation 3 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  21 22 23 24 25
##
## -- Imputation 4 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  21 22 23 24 25
##
## -- Imputation 5 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  21 22 23 24 25
```

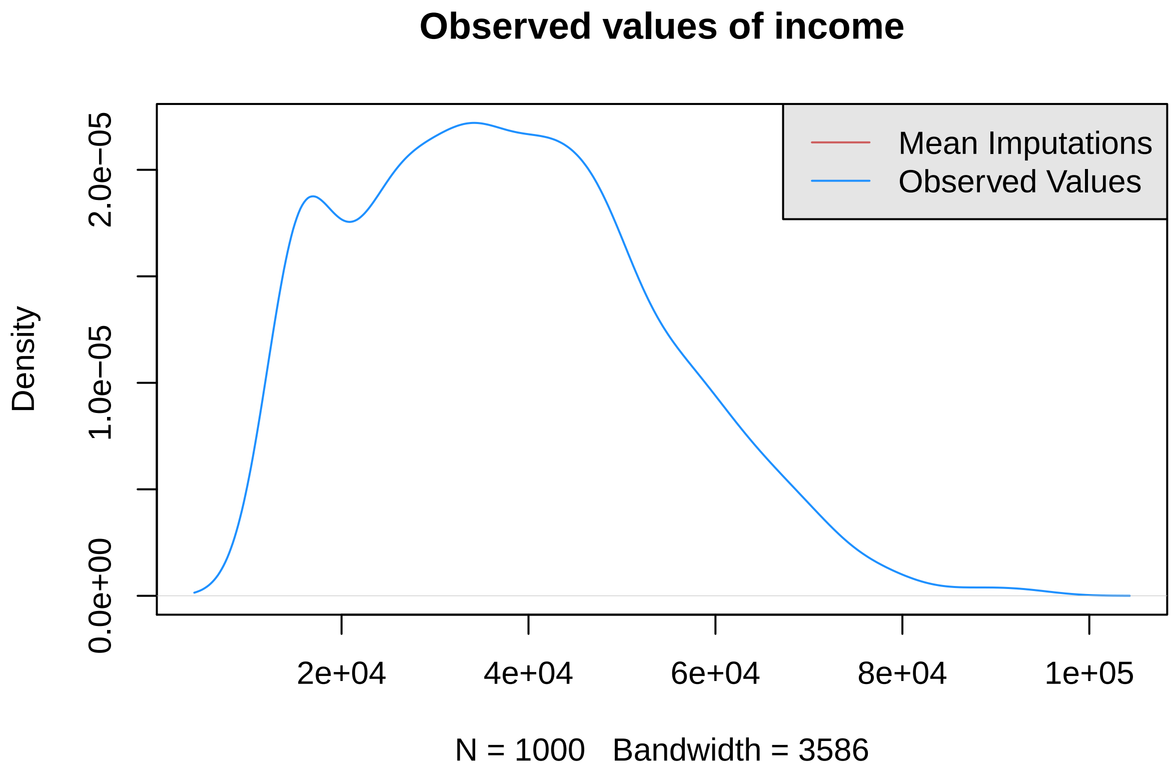
#NA's are removed from college and blues variables after using summary to check.

```
summary(data.imp$imputations$imp1)
```

```
##        X              blues         college           income
##  Min.   :   1.0   Min.   :0.0   Min.   :0.000   Min.   :15000
##  1st Qu.: 250.8   1st Qu.:0.0   1st Qu.:0.000   1st Qu.:25476
```

```
## Median : 500.5   Median :0.0   Median :0.000   Median :36332
## Mean   : 500.5   Mean   :0.4   Mean   :0.202   Mean   :37663
## 3rd Qu.: 750.2   3rd Qu.:1.0   3rd Qu.:0.000   3rd Qu.:48173
## Max.   :1000.0   Max.   :1.0   Max.   :1.000   Max.   :93569
```

```
compare.density(data.imp, var="income")
```

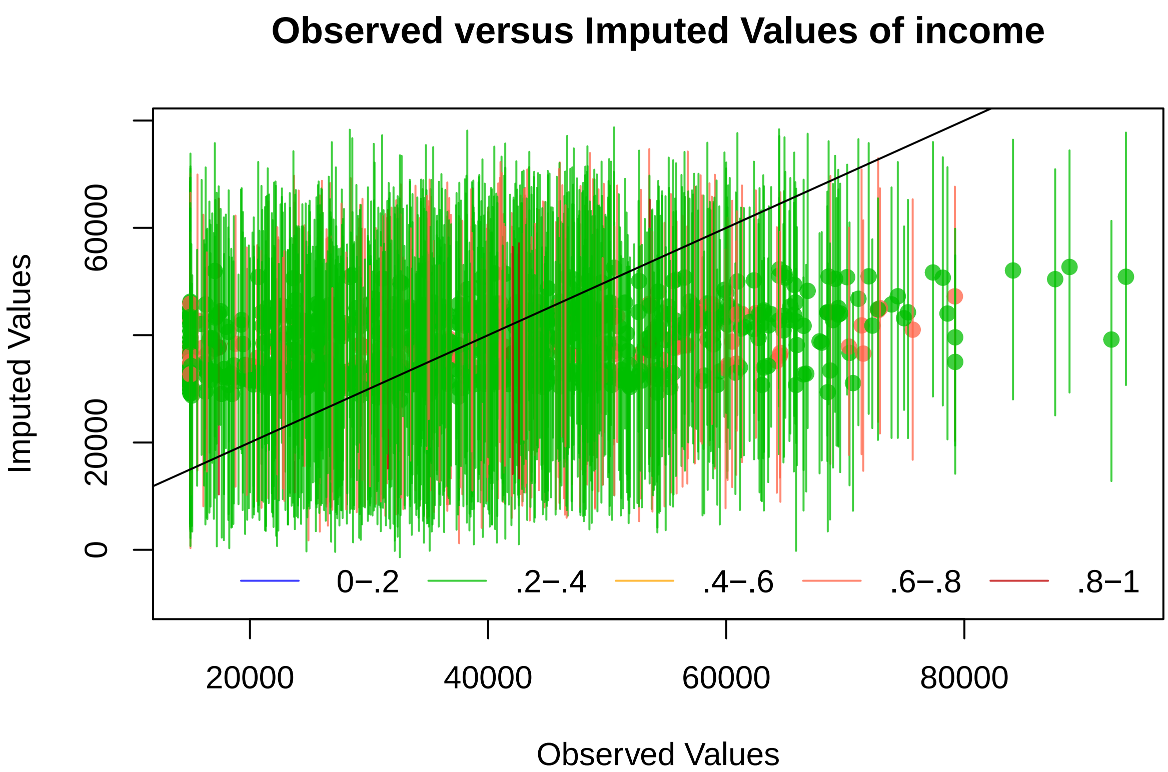



```
#Continuous Variable
```

Diagnosing Impuations

There doesn't show any mean imputations, with no missing data.

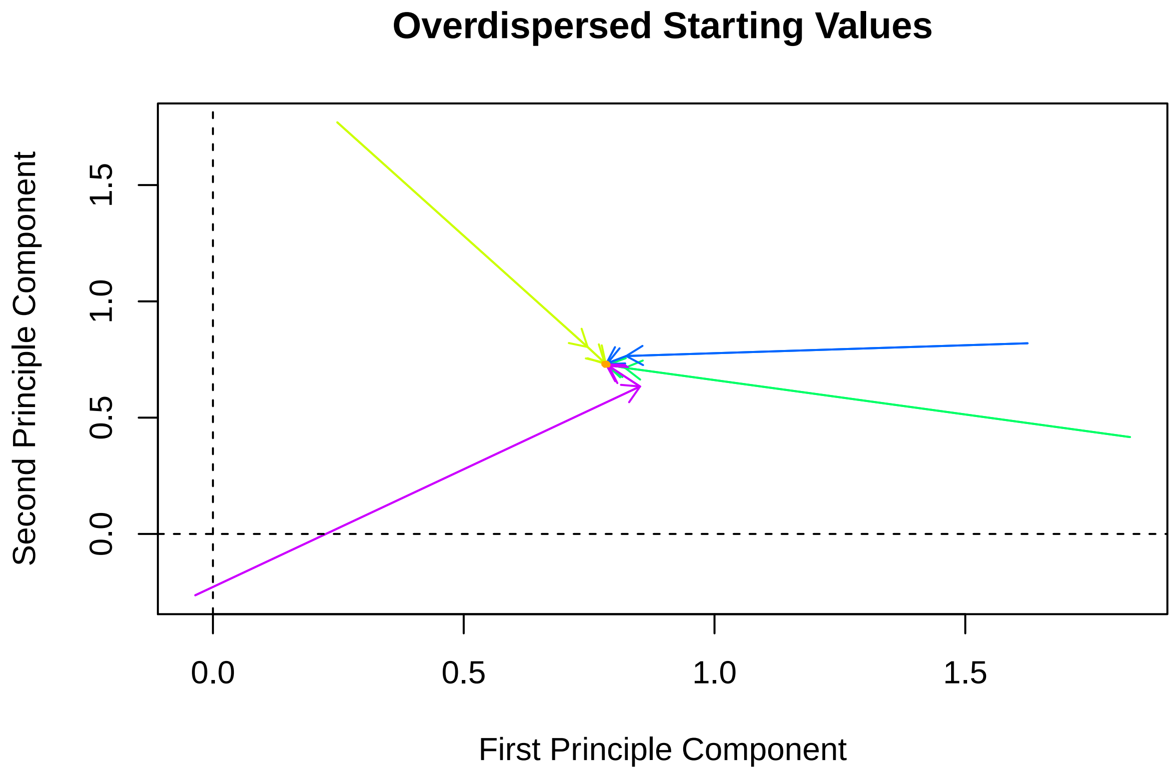
```
overimpute(data.imp, var = "income")
```




```
#Continuous Variable
```

The over imputation diagnostic chart shows us that the confidence intervals for the observed data, for the most part, fall within the y = x line. Although more accurate for at the lower-middle of they = x line, as the values get higher, we see accuracy begins to decrease as the y = x line no longer begins to fall within the confidence intervals.

```
disperse(data.imp, dims=2, m=5)
```

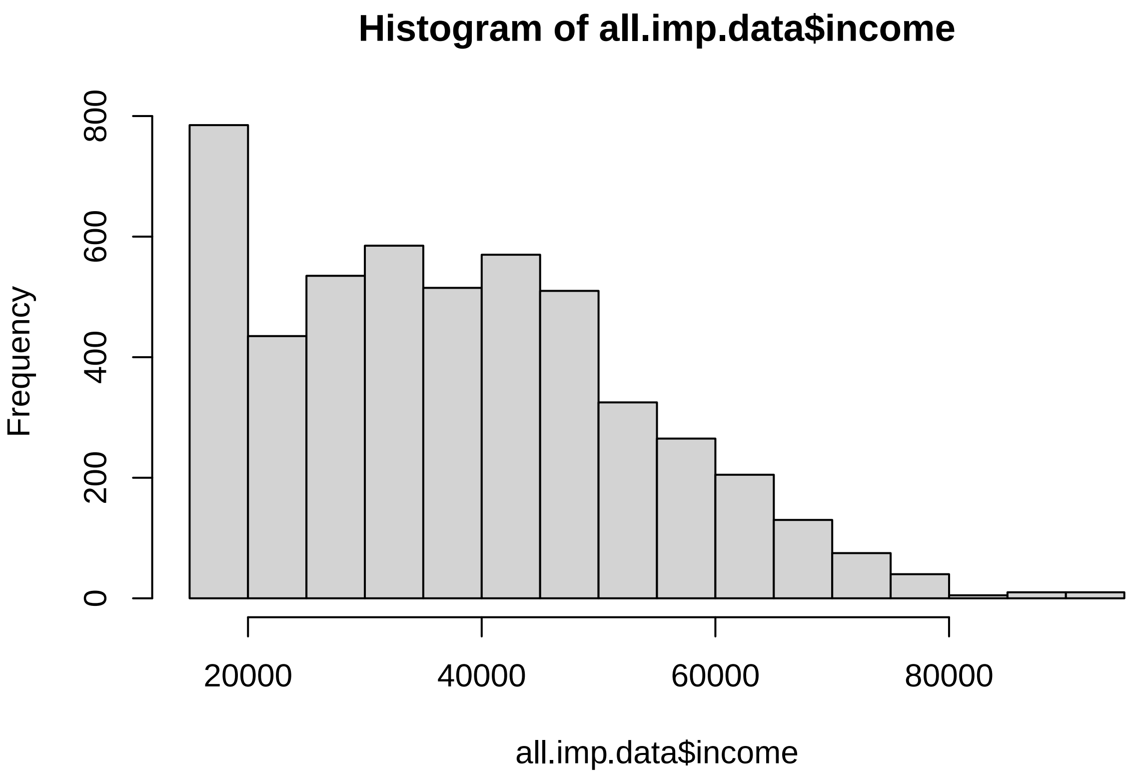


###

After running an Expectation–maximization algorithm, we have evidence that even with different starting values, each of the imputations arrives at the same path of predictions.

```
all.imp.data<- rbind(data.imp$imputations$imp1, data.imp$imputations$imp2, data.imp$imputations$imp3, da

#histogram
hist(all.imp.data$income)
```



The imputed data histogram shows an increase in the frequency of all the income groups, although specifically higher for those in the $0-$20,000 income bracket.

```
crosstab(all.imp.data$blues, all.imp.data$college, all.imp.data$income, prop.c = T, prop.r = T, plot =
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |-------------------------|
##
## =====================================================
##                      Respondent a college Graduate?
## Is in Blue group?             0           1       Total
## -----------------------------------------------------
## 0                     74418662    26980347   101399009
##                          73.4%       26.6%       53.8%
##                          51.4%       62.1%
## -----------------------------------------------------
## 1                     70463807    16450070    86913877
##                          81.1%       18.9%       46.2%
##                          48.6%       37.9%
## -----------------------------------------------------
## Total                144882469    43430417   188312886
##                          76.9%       23.1%
## =====================================================
```

```r
aggregate(all.imp.data$income ~ (all.imp.data$college + all.imp.data$blues), data=all.imp.data, na.rm=T
```

```
##   all.imp.data$college all.imp.data$blues all.imp.data$income
## 1                    0                  0            32271.75
## 2                    1                  0            38324.36
## 3                    0                  1            42042.84
## 4                    1                  1            52388.76
```

```r
wtd.cor(all.imp.data$income, as.numeric(as.character(all.imp.data$college)))
```

```
##   correlation    std.err  t.value      p.value
## Y   0.1594614 0.01396397 11.41949 7.810245e-30
```

```r
wtd.cor(all.imp.data$income, as.numeric(as.character(all.imp.data$blues)))
```

```
##   correlation    std.err  t.value       p.value
## Y   0.3083794 0.01345559 22.91831 1.288467e-110
```

```r
wtd.cor(all.imp.data$income, as.numeric(as.character(all.imp.data$blues + all.imp.data$college)))
```

```
##   correlation    std.err  t.value       p.value
## Y   0.3559973 0.01321829 26.93219 2.664366e-149
```

```r
model2 <- zelig(income~college + blues, model="normal", data=data.imp)
```

```
## Warning: `tbl_df()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Warning: `group_by_()` is deprecated as of dplyr 0.7.0.
## Please use `group_by()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## How to cite this model in Zelig:
##   R Core Team. 2008.
##   normal: Normal Regression for Continuous Dependent Variables
##   in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,
##   "Zelig: Everyone's Statistical Software," http://zeligproject.org/
```

```r
summary(model2)
```

```
## Model: Combined Imputations
##
##             Estimate Std.Error z value Pr(>|z|)
## (Intercept)    31937       736   43.37  < 2e-16
## college         7471      1238    6.04  1.6e-09
## blues          10560      1072    9.85  < 2e-16
##
## For results from individual imputed datasets, use summary(x, subset = i:j)
## Next step: Use 'setx' method
```

The results from the 2nd model show that both memberships in the blue group, and being a college graduate have a statistically significant effect on respondent income. With a coefficient of 7471, we see a $7,471 increase for those who are college graduates, per unit increase of income. With a coefficient of 10560, we see that membership in the blue group is associated with a $10,560 increase in income per every unit increase. We see a lower income increase for those in the blue group for imputed data, specifically a $207 decrease, and for college graduates, the income for the imputed data decreased by $732.