In Class Assignment 4

Lydia Strebe

February 18, 2019

Summary of Data

Below is a summary of the data set “big_missing.csv”. As you can see, each x column has missing data points.

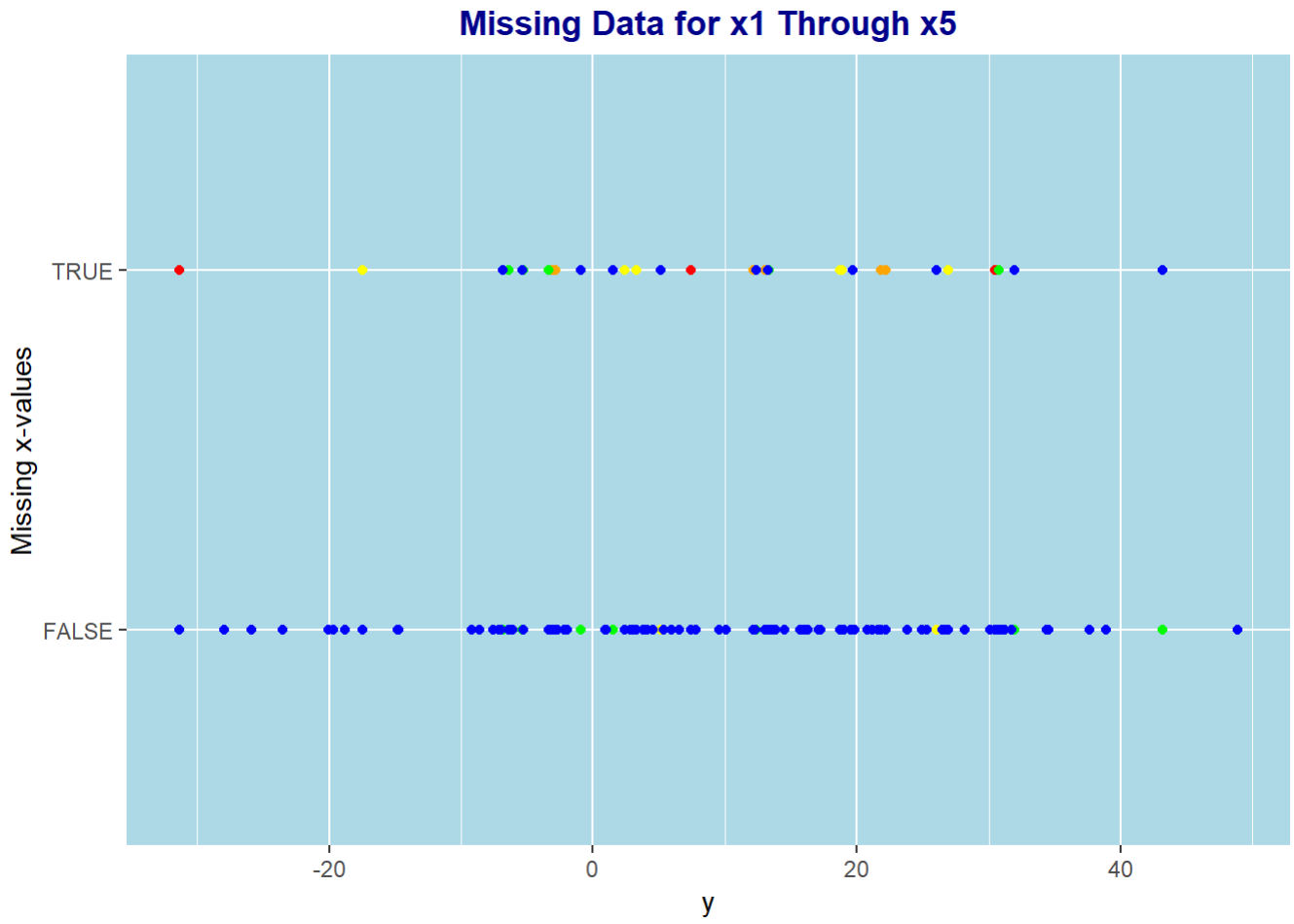
```
summary(data_set)
```

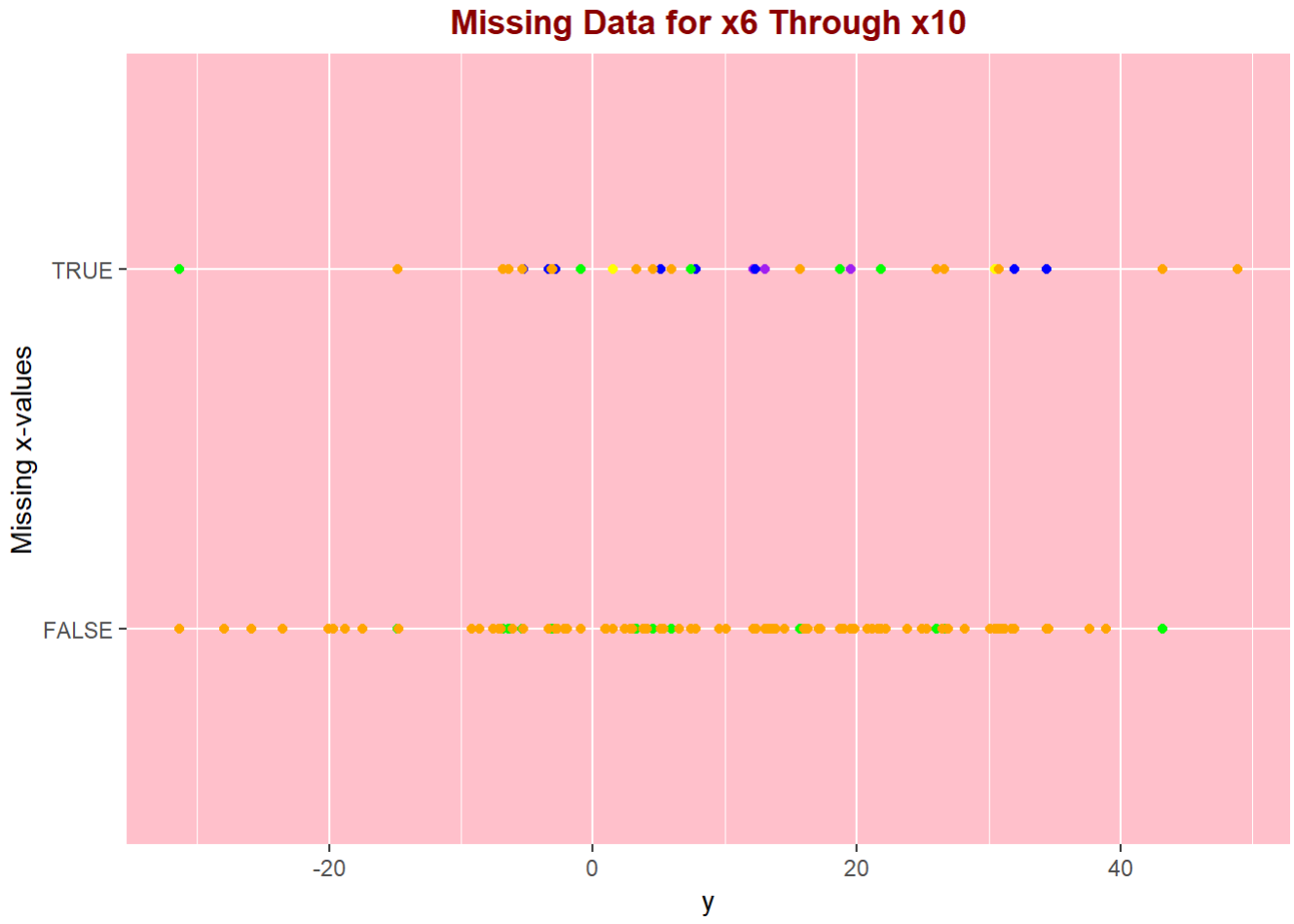
```
##        x1                 x2                 x3          
##  Min.   :-2.85696   Min.   :-2.40106   Min.   :-2.018833  
##  1st Qu.:-0.76317   1st Qu.:-0.57819   1st Qu.:-0.457711  
##  Median :-0.12731   Median : 0.06486   Median :-0.008622  
##  Mean   :-0.03669   Mean   : 0.03498   Mean   : 0.018659  
##  3rd Qu.: 0.86817   3rd Qu.: 0.64943   3rd Qu.: 0.453366  
##  Max.   : 2.66424   Max.   : 2.64657   Max.   : 1.784673  
##  NA's   :10         NA's   :12         NA's   :10         
##        x4                 x5                 x6          
##  Min.   :-1.1089   Min.   :-3.10974   Min.   :-2.78223  
##  1st Qu.:-0.4835   1st Qu.:-0.69247   1st Qu.:-0.72309  
##  Median :-0.1036   Median : 0.12553   Median :-0.23559  
##  Mean   :-0.0474   Mean   : 0.01172   Mean   :-0.08153  
##  3rd Qu.: 0.3514   3rd Qu.: 0.72377   3rd Qu.: 0.64835  
##  Max.   : 1.4986   Max.   : 2.15504   Max.   : 2.37292  
##  NA's   :8         NA's   :11         NA's   :13        
##        x7                 x8                 x9                x10          
##  Min.   :-2.52170   Min.   :-3.9179   Min.   :-2.2041   Min.   :-2.06412  
##  1st Qu.:-0.58542   1st Qu.:-0.8708   1st Qu.:-0.6859   1st Qu.:-0.79520  
##  Median : 0.04912   Median :-0.2553   Median : 0.1682   Median :-0.11243  
##  Mean   :-0.06950   Mean   :-0.1843   Mean   : 0.1111   Mean   : 0.03711  
##  3rd Qu.: 0.44330   3rd Qu.: 0.6912   3rd Qu.: 0.8439   3rd Qu.: 0.83159  
##  Max.   : 1.63941   Max.   : 2.8479   Max.   : 2.4595   Max.   : 2.30761  
##  NA's   :10         NA's   :15        NA's   :7         NA's   :15        
##        y          
##  Min.   :-31.360  
##  1st Qu.: -2.881  
##  Median : 12.345  
##  Mean   :  9.810  
##  3rd Qu.: 21.667  
##  Max.   : 48.838  
##
```

Plots of Missing Data

We can see that the data is missing completely at random by looking at the graphs below. The first graph shows the missing values of x1 through x5 for all the values of y and the second shows the missing values of x6 through x10 (these are split into two graphs for easier interpretation). Each color represents a different x. As you can see,

the missing data does not appear in any sort of pattern.


Missing Data for x1 Through x5
TRUE
FALSE
Missing x-values
-20
0
20
40
y


Missing Data for x6 Through x10
TRUE
FALSE
Missing x-values
-20
0
20
40
y

Regression Techniques

There are a few techniques to bypass missing data. First we will try a regression with line deletion:

```
data_set_delete = na.omit(data_set)
delete_model = lm(y~.,data=data_set_delete)
summary(delete_model)
```

```
##
## Call:
## lm(formula = y ~ ., data = data_set_delete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6805  -5.1175  -0.2035   6.8746  15.0412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.9932     1.4355   6.265 1.26e-07 ***
## x1            3.9131     1.9325   2.025  0.04883 *
## x2           -1.6148     2.3013  -0.702  0.48648
## x3            7.1822     3.1157   2.305  0.02583 *
## x4           -3.4870     2.6064  -1.338  0.18765
## x5           -0.2990     2.1386  -0.140  0.88943
## x6            5.3380     2.1329   2.503  0.01603 *
## x7            0.4504     2.9189   0.154  0.87806
## x8            5.0405     1.6074   3.136  0.00302 **
## x9           -0.6487     2.0402  -0.318  0.75198
## x10           2.9116     1.7447   1.669  0.10210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.551 on 45 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.6961
## F-statistic:  13.6 on 10 and 45 DF,  p-value: 1.372e-10
```

The intercept has a very statistically significant p-value (<0.001). There is one other coefficient with a p-value <0.01 (x8), three others that are less than 0.05 (x1, x3, and x6), but the rest are not statistically significant.

We do not have a feasible way to do single imputation, but we can use multiple imputation and pool the regressions.

```
data_set_mi = mice(data_set,m=15,print=FALSE)
mi_fit = with(data_set_mi,lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10))
mi_fit_pooled = pool(mi_fit)
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

```
summary(mi_fit_pooled)
```

```
##               estimate std.error  statistic       df      p.value
## (Intercept) 10.1471228  1.141787  8.8870566 61.67217 1.214140e-12
## x1           3.8705918  1.321033  2.9299738 56.20675 4.747239e-03
## x2          -2.0346881  1.869253 -1.0885033 50.67492 2.806084e-01
## x3           5.8819805  2.593535  2.2679391 49.12992 2.684639e-02
## x4          -4.0739424  2.269327 -1.7952209 59.17539 7.751847e-02
## x5          -0.4568413  1.810999 -0.2522593 46.78148 8.016795e-01
## x6           4.1941364  1.929256  2.1739657 44.90982 3.355477e-02
## x7          -0.6212299  2.766361 -0.2245657 25.42956 8.230595e-01
## x8           4.2653447  1.395200  3.0571569 50.90258 3.301505e-03
## x9          -2.3672672  1.582793 -1.4956268 58.78711 1.398509e-01
## x10          4.7848644  1.415433  3.3804951 51.78205 1.259825e-03
```

Again, the intercept has a very statistically significant p-value (<0.001). There are now three coefficients with p-values <0.01 (x1, x8, and x10) and one that is less than 0.05 (x3). This leads me to believe that x1 and x8 are definitely significant. x3 also seems fairly significant. x10 appears to be significant too, although it did not appear to be significant in the regression with line deletion. Finally, x6 is possibly significant since its first p-value was <0.05, and its second was <0.1.