Homework 5

Juntao Zhang

06/03/2021

#Read and inspect data set

```
data <-read.csv("C:/Users/joann/OneDrive/Desktop/missing data/week 2/aug_train.csv",
                 na.strings = "")
```

#Encode character variables

```
unique(data$relevent_experience )
```

```
## [1] "Has relevent experience" "No relevent experience"
```

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 4.0.3
```

```
data$relevent_experience <- revalue(data$relevent_experience,
                                    c("Has relevent experience"=1))
data$relevent_experience <- revalue(data$relevent_experience,
                                    c("No relevent experience"=0))
data$relevent_experience <-as.numeric(data$relevent_experience)

unique(data$last_new_job)
```

```
## [1] "1"     ">4"    "never" "4"     "3"     "2"     NA
```

```
data$last_new_job <- revalue(data$last_new_job, c("never"=0))
data$last_new_job <- revalue(data$last_new_job, c(">4"=5))
data$last_new_job <-as.numeric(data$last_new_job)

unique(data$enrolled_university )
```

```
## [1] "no_enrollment"    "Full time course" NA                 "Part time course"
```

```
data$enrolled_university <- revalue(data$enrolled_university,
                                    c("no_enrollment"=0))
data$enrolled_university <- revalue(data$enrolled_university,
                                    c("Part time course"=1))
data$enrolled_university <- revalue(data$enrolled_university,
```

```r
                                    c("Full time course" = 2))
data$enrolled_university <-as.numeric(data$enrolled_university)

unique(data$education_level)
```

```
## [1] "Graduate"       "Masters"         "High School"    NA
## [5] "Phd"            "Primary School"
```

```r
data$education_level <- as.numeric(factor(data$education_level,
                                          levels = c("Primary School",
                                                     "High School","Graduate",
                                                     "Masters","Phd")))
unique(data$gender)
```

```
## [1] "Male"   NA       "Female" "Other"
```

```r
data$gender <- as.factor(data$gender)
```

```r
#I will keep the variables that can be used for my analysis
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
data2 = select(data,'training_hours','gender','relevent_experience',
               'last_new_job','enrolled_university','education_level','target')
```

```r
#Generate missing values for training_hours depending on one variable
library(dplyr)
data_new = select(data,'city_development_index','training_hours')
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.0.3

##
## Attaching package: 'mice'
## The following object is masked from ’package:stats’:
##
##     filter
```

```
## The following objects are masked from ’package:base’:
##
##     cbind, rbind
```

```
cont_cat = ampute(data_new,prop = 0.2,patterns=c(1,0),mech = "MAR")$amp
data2[’training_hours’] = cont_cat[’training_hours’]


#check again for the generated missing values
sapply(data2, function(x) sum(is.na(x)))
```

```
##      training_hours              gender relevent_experience        last_new_job
##                3883                4508                   0                 423
## enrolled_university     education_level              target
##                 386                 460                   0
```

```
#Q1

library(mi)
```

```
## Warning: package ’mi’ was built under R version 4.0.3
```

```
## Loading required package: Matrix
```

```
## Loading required package: stats4
```

```
## mi (Version 1.0, packaged: 2015-04-16 14:03:10 UTC; goodrich)
```

```
## mi  Copyright (C) 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015 Trustees of Columbia University
```

```
## This program comes with ABSOLUTELY NO WARRANTY.
```

```
## This is free software, and you are welcome to redistribute it
```

```
## under the General Public License version 2 or later.
```

```
## Execute RShowDoc(’COPYING’) for details.
```

```
##
## Attaching package: ’mi’
```

```
## The following objects are masked from ’package:mice’:
##
##     complete, pool
```

```
#Run mi with 5 chains and 50 iterations on the dataset

# Create the missing data frame object
mdf = missing_data.frame(data2)

# Examine the default settings
show(mdf)
```

```
## Object of class missing_data.frame with 19158 observations on 7 variables
##
## There are 32 missing data patterns
##
## Append '@patterns' to this missing_data.frame to access the corresponding pattern for every observati
##
##                                      type missing method  model
## training_hours                  continuous    3883    ppd linear
## gender              unordered-categorical    4508    ppd mlogit
## relevent_experience                binary       0   <NA>   <NA>
## last_new_job                    continuous     423    ppd linear
## enrolled_university   ordered-categorical     386    ppd ologit
## education_level       ordered-categorical     460    ppd ologit
## target                             binary       0   <NA>   <NA>
##
##                           family     link transformation
## training_hours          gaussian identity    standardize
## gender               multinomial    logit           <NA>
## relevent_experience         <NA>     <NA>           <NA>
## last_new_job            gaussian identity    standardize
## enrolled_university multinomial    logit           <NA>
## education_level     multinomial    logit           <NA>
## target                      <NA>     <NA>           <NA>
```

```
# Running the chains
imputations <- mi(mdf, n.chains = 5, n.iter=50,max.minutes = 20)
```

```
#Q2

#Check convergence/diagnostics and make changes if necessary
Rhats(imputations)
```

```
##      mean_training_hours              mean_gender        mean_last_new_job
##                1.0005798                1.0476161                1.0099619
## mean_enrolled_university     mean_education_level        sd_training_hours
##                0.9934947                0.9904098                1.0169866
##                sd_gender          sd_last_new_job   sd_enrolled_university
##                1.0330956                0.9961615                0.9942598
##       sd_education_level
##                0.9903392
```

```
round(mipply(imputations, mean, to.matrix = TRUE), 3)
```

```
##                             chain:1 chain:2 chain:3 chain:4 chain:5
## training_hours                 0.001   0.002   0.000   0.000   0.002
## gender                         1.929   1.929   1.931   1.928   1.928
## relevent_experience            1.720   1.720   1.720   1.720   1.720
## last_new_job                  -0.003  -0.003  -0.004  -0.004  -0.003
## enrolled_university            1.467   1.468   1.469   1.468   1.468
## education_level                3.134   3.132   3.132   3.134   3.132
## target                         1.249   1.249   1.249   1.249   1.249
## missing_training_hours         0.203   0.203   0.203   0.203   0.203
## missing_gender                 0.235   0.235   0.235   0.235   0.235
## missing_last_new_job           0.022   0.022   0.022   0.022   0.022
## missing_enrolled_university    0.020   0.020   0.020   0.020   0.020
## missing_education_level        0.024   0.024   0.024   0.024   0.024
```

#make changes for last_new_job and training hours #since they have unequal means for each chain #the inspected problems are on the type of these two variables #training hours are always >0 and last new job is a ordered categorical variable

```r
mdf <- change(mdf, y = "last_new_job", what = "type",
              to = "ordered-categorical")
mdf <- change(mdf, y = "training_hours", what = "type", to = "pos")
show(mdf)
```

```
## Object of class missing_data.frame with 19158 observations on 7 variables
##
## There are 32 missing data patterns
##
## Append '@patterns' to this missing_data.frame to access the corresponding pattern for every observati
##
##                                      type missing method  model
## training_hours        positive-continuous    3883    ppd linear
## gender              unordered-categorical    4508    ppd mlogit
## relevent_experience                binary       0   <NA>   <NA>
## last_new_job          ordered-categorical     423    ppd ologit
## enrolled_university   ordered-categorical     386    ppd ologit
## education_level       ordered-categorical     460    ppd ologit
## target                             binary       0   <NA>   <NA>
##
##                          family     link transformation
## training_hours         gaussian identity            log
## gender              multinomial    logit           <NA>
## relevent_experience        <NA>     <NA>           <NA>
## last_new_job        multinomial    logit           <NA>
## enrolled_university multinomial    logit           <NA>
## education_level     multinomial    logit           <NA>
## target                     <NA>     <NA>           <NA>
```

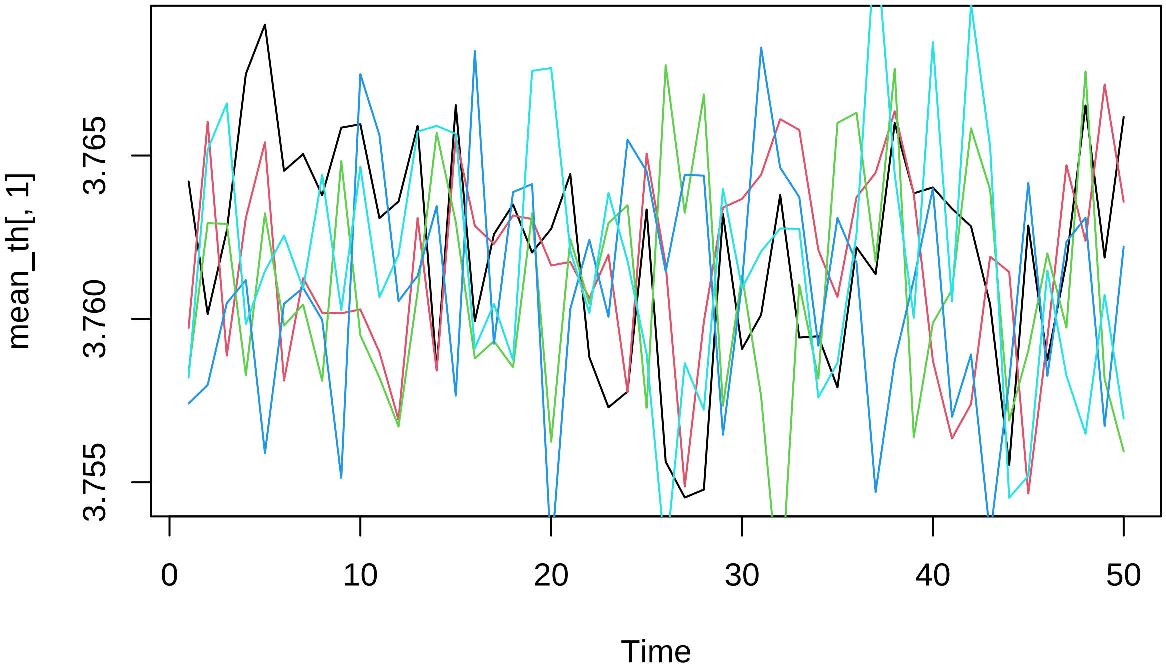
```r
# Rerunning the chains
imputations <- mi(mdf, n.chains = 5, n.iter=50)
round(mipply(imputations, mean, to.matrix = TRUE), 3)
```

```
##                             chain:1 chain:2 chain:3 chain:4 chain:5
## training_hours                3.766   3.764   3.756   3.762   3.757
## gender                        1.929   1.930   1.930   1.928   1.930
## relevent_experience           1.720   1.720   1.720   1.720   1.720
## last_new_job                  2.989   2.991   2.989   2.987   2.991
## enrolled_university           1.467   1.469   1.468   1.467   1.467
## education_level               3.132   3.131   3.132   3.132   3.132
## target                        1.249   1.249   1.249   1.249   1.249
## missing_training_hours        0.203   0.203   0.203   0.203   0.203
## missing_gender                0.235   0.235   0.235   0.235   0.235
## missing_last_new_job          0.022   0.022   0.022   0.022   0.022
## missing_enrolled_university   0.020   0.020   0.020   0.020   0.020
## missing_education_level       0.024   0.024   0.024   0.024   0.024
```

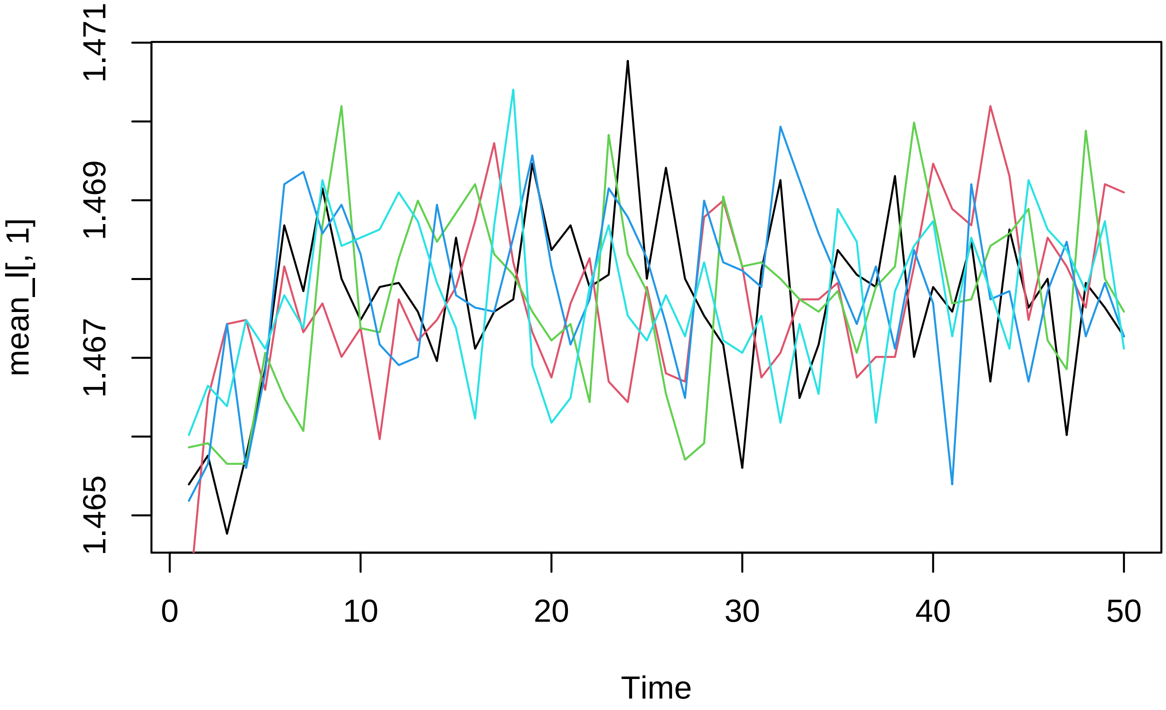
```r
converged <- mi2BUGS(imputations)
Rhats(imputations)
```

```
##      mean_training_hours              mean_gender        mean_last_new_job
##                1.0003277                1.0735958                1.0057900
## mean_enrolled_university     mean_education_level        sd_training_hours
##                0.9914568                0.9904133                1.0262518
##                sd_gender          sd_last_new_job   sd_enrolled_university
##                1.0466836                0.9999715                0.9910441
##       sd_education_level
##                0.9901035
```

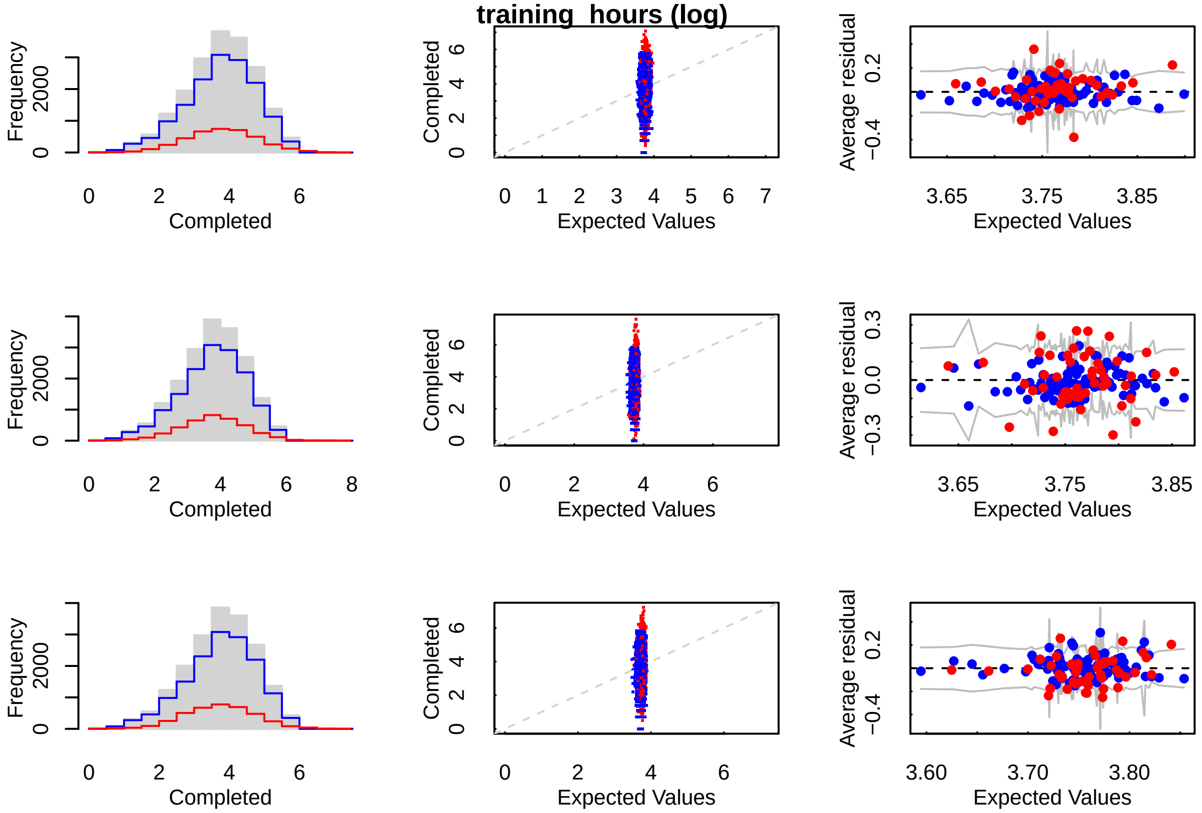
```r
mean_th = converged[, , 1]
# Traceplot of mean imputed training hours
ts.plot(mean_th[,1], col=1)
lines(mean_th[,2], col= 2)
lines(mean_th[,3], col= 3)
lines(mean_th [,4], col= 4)
lines(mean_th [,5], col= 5)
```

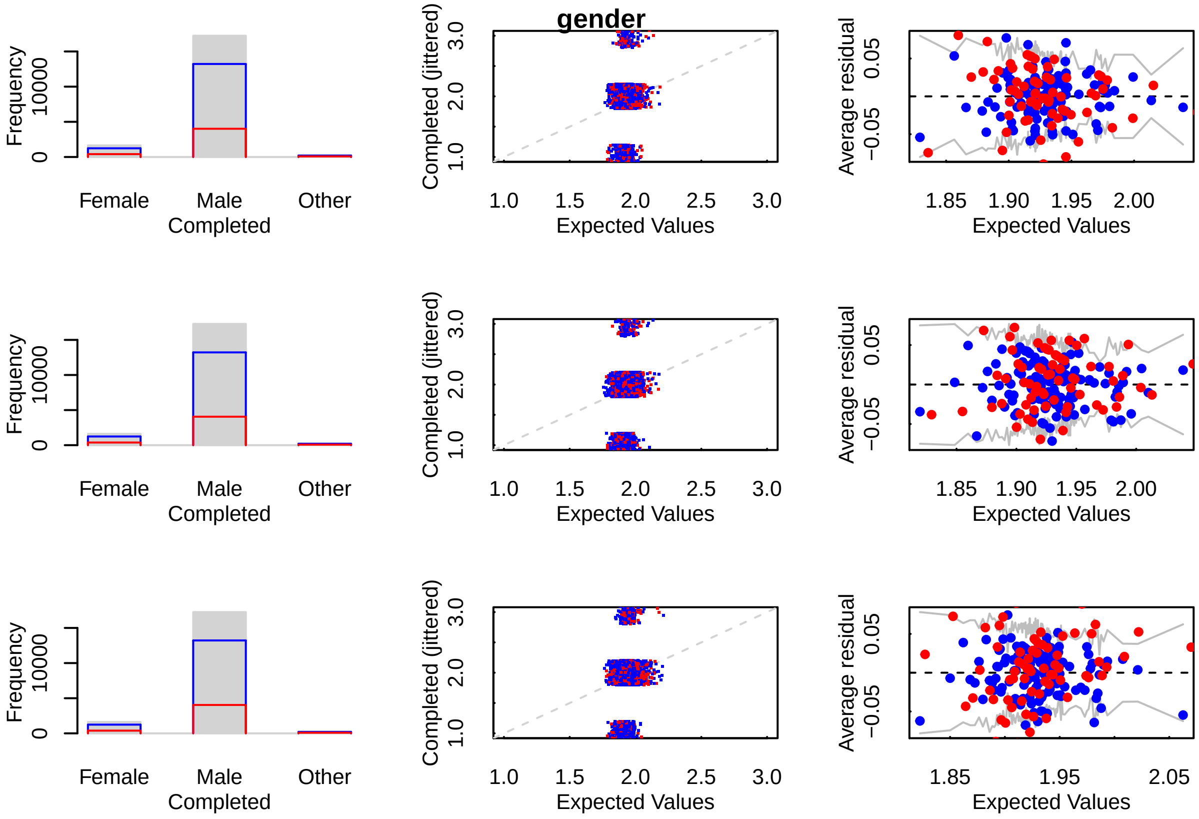


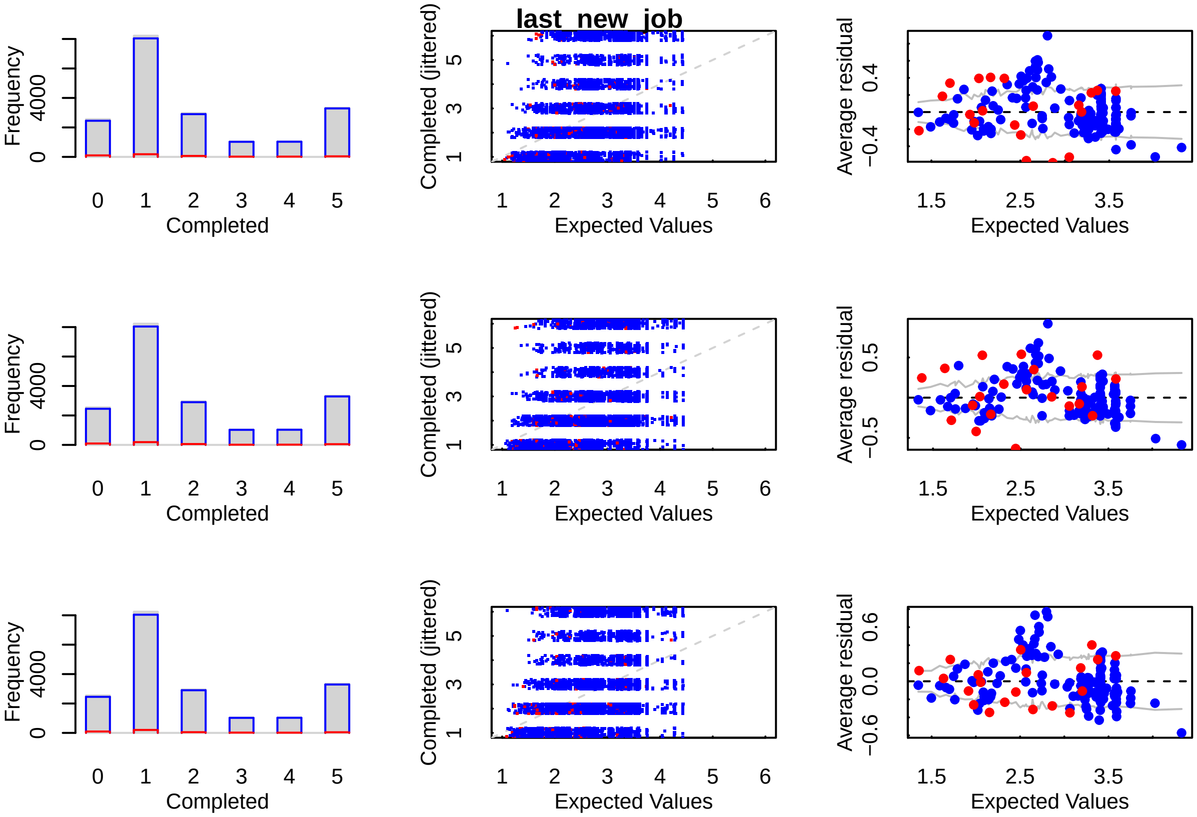
```
mean_l = converged[, , 4]
# Traceplot of mean imputed last new job
ts.plot(mean_l[,1], col=1)
lines(mean_l[,2], col= 2)
lines(mean_l[,3], col= 3)
lines(mean_l [,4], col= 4)
lines(mean_l [,5], col= 5)
```

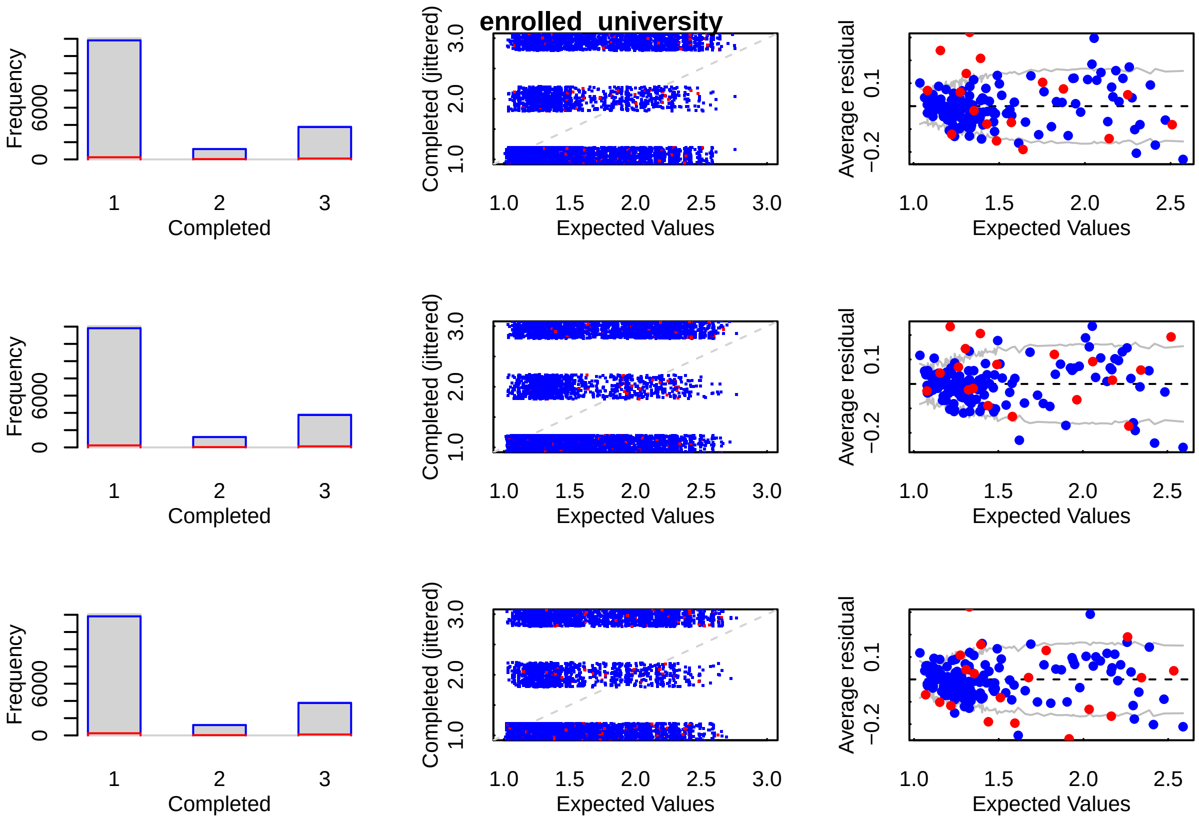


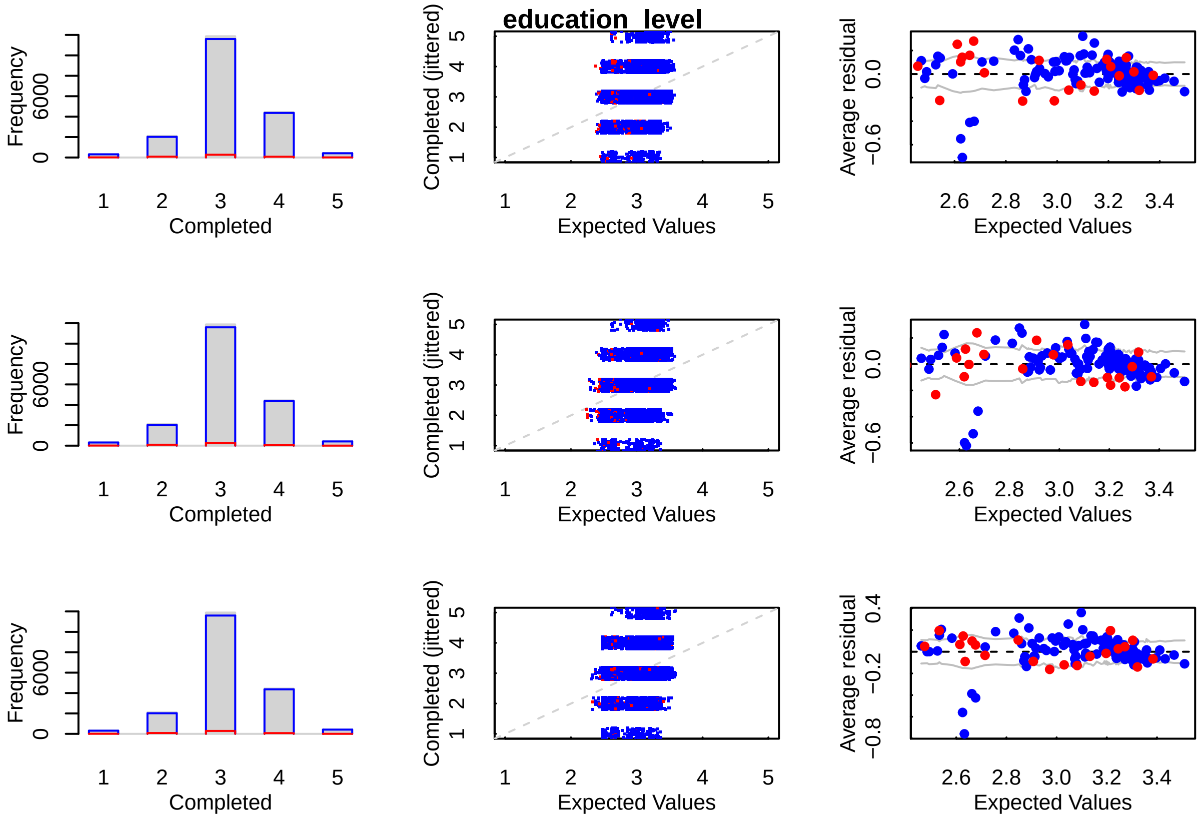
```
#check for the plots
plot(imputations)
```


training hours (log)
Frequency
0
2000
0
2
4
6
Completed
Completed
0
2
4
6
0
1
2
3
4
5
6
7
Expected Values
Average residual
−0.4
0.2
3.65
3.75
3.85
Expected Values
Frequency
0
2000
0
2
4
6
8
Completed
Completed
0
2
4
6
0
2
4
6
Expected Values
Average residual
−0.3
0.0
0.3
3.65
3.75
3.85
Expected Values
Frequency
0
2000
0
2
4
6
Completed
Completed
0
2
4
6
0
2
4
6
Expected Values
Average residual
−0.4
0.2
3.60
3.70
3.80
Expected Values


gender
Frequency
0
10000
Female
Male
Other
Completed
Completed (jittered)
1.0
2.0
3.0
1.0
1.5
2.0
2.5
3.0
Expected Values
Average residual
−0.05
0.05
1.85
1.90
1.95
2.00
2.05


last_new_job
Frequency
0
4000
Completed
0 1 2 3 4 5
Completed (jittered)
1
3
5
Expected Values
1 2 3 4 5 6
Average residual
−0.4
0.4
1.5 2.5 3.5
Expected Values
Frequency
0
4000
Completed
0 1 2 3 4 5
Completed (jittered)
1
3
5
Expected Values
1 2 3 4 5 6
Average residual
−0.5
0.5
1.5 2.5 3.5
Expected Values
Frequency
0
4000
Completed
0 1 2 3 4 5
Completed (jittered)
1
3
5
Expected Values
1 2 3 4 5 6
Average residual
−0.6
0.0
0.6
1.5 2.5 3.5
Expected Values


enrolled university
Frequency
0
6000
1
2
3
Completed
Completed (jittered)
1.0
2.0
3.0
1.0
1.5
2.0
2.5
3.0
Expected Values
Average residual
−0.2
0.1
1.0
1.5
2.0
2.5
Expected Values



```r
#the results converged
```

```r
#Q3

#Pool the results and report the estimated equation
analysis <- pool(target ~ training_hours+gender+relevent_experience+
                   last_new_job+enrolled_university+education_level,imputations)
display(analysis)
```

```
## bayesglm(formula = target ~ training_hours + gender + relevent_experience +
##     last_new_job + enrolled_university + education_level, data = imputations)
##                         coef.est coef.se
## (Intercept)             -0.94     0.09
## training_hours           0.00     0.00
## genderMale              -0.14     0.07
## genderOther             -0.06     0.17
## relevent_experience1    -0.54     0.04
## last_new_job.L          -0.28     0.05
## last_new_job.Q          -0.06     0.05
## last_new_job.C          -0.05     0.06
## last_new_job^4          -0.05     0.06
## last_new_job^5          -0.04     0.06
## enrolled_university.L    0.43     0.03
## enrolled_university.Q    0.14     0.06
## education_level.L        0.49     0.15
## education_level.Q     -0.77     0.12
## education_level.C     -0.27     0.08
## education_level^4      0.31     0.05
## n = 19142, k = 16
## residual deviance = 20624.8, null deviance = 21518.9 (difference = 894.1)
```

```
#the estimated equation of the pooled result is as shown above
```

#Q4 #compare pooled result with complete dataset result

```
#original complete data set
data_complete = na.omit(data)

#Linear regression analysis for the target variable
#0-not looking for a job change 1-looking for a job change


#model with the original complete data cases
model2 = lm(target ~ training_hours+gender+relevent_experience+last_new_job+
              enrolled_university+education_level, data = data_complete)
summary(model2)
```

```
##
## Call:
## lm(formula = target ~ training_hours + gender + relevent_experience +
##     last_new_job + enrolled_university + education_level, data = data_complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3144 -0.1779 -0.1532 -0.1169  0.9241
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.796e-01  3.151e-02   8.872  < 2e-16 ***
## training_hours      -8.833e-05  6.483e-05  -1.362  0.17308
## genderMale           5.328e-03  1.370e-02   0.389  0.69745
## genderOther         -2.146e-02  4.387e-02  -0.489  0.62471
## relevent_experience -2.577e-02  1.216e-02  -2.119  0.03409 *
## last_new_job        -1.397e-02  2.367e-03  -5.901 3.74e-09 ***
## enrolled_university  4.609e-02  6.608e-03   6.974 3.29e-12 ***
## education_level     -2.063e-02  7.458e-03  -2.767  0.00567 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3693 on 8947 degrees of freedom
## Multiple R-squared:  0.01365,    Adjusted R-squared:  0.01288
## F-statistic: 17.69 on 7 and 8947 DF,  p-value: < 2.2e-16
```

#Comparing the two results: #The estimated coefficients for the original complete data set are all very small but for the pooled result from the imputed data set, relevant experience, enrolled university and education level are showing a larger effect (estimated coefficient) for target.