# Data Imputation Practice

Sutianyi Wen

10/27/2020

---

## Lab report

### Load data here

#### Part1 - Question 1:

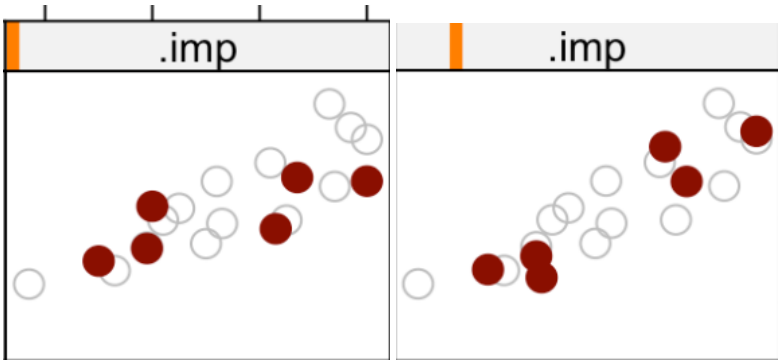
The R commands to generate missing values are printed below. There are 6 missing values in *age* column

```
tree_mcar <- delete_MCAR(ds=tree,p=0.3,cols_mis =c('age'))
tree_mcar
```

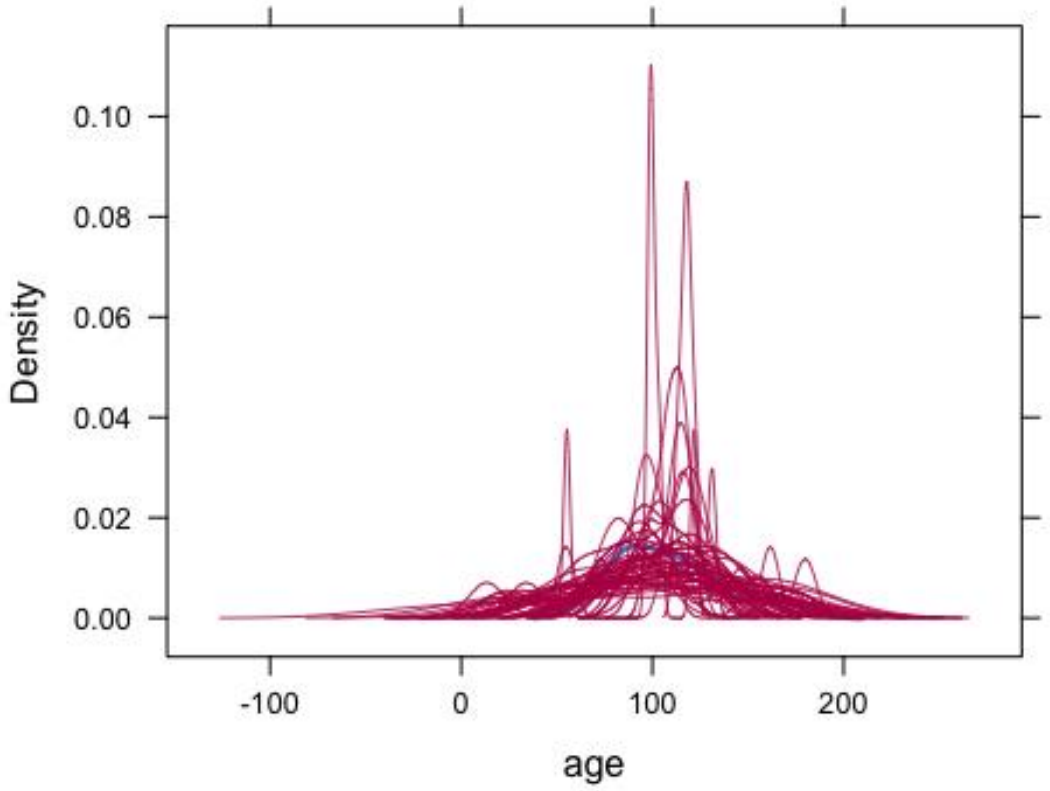
```
##    number diameter age
## 1       1     12.0 125
## 2       2     11.4 119
## 3       3      7.9  83
## 4       4      9.0  85
## 5       5     10.5  99
## 6       6      7.9 117
## 7       7      7.3  NA
## 8       8     10.2  NA
## 9       9     11.7 154
## 10     10     11.3 168
## 11     11      5.7  NA
## 12     12      8.0  80
## 13     13     10.3  NA
## 14     14     12.0  NA
## 15     15      9.2 122
## 16     16      8.5 106
## 17     17      7.0  82
## 18     18     10.7  88
## 19     19      9.3  97
## 20     20      8.2  NA
```

#### Part1 - Question 2:

Let's first look at the scatter plots. I pulled out 2 ideal imputation from xyplot's result. As you can see from the plots below, red dots are generated and they're really close to the observed values.




Then Let's take a look at the density plot. The density plots for 50 imputed data(red line) share the same shape as the observed density plot(blue line). Most of the red density line have the same center as observed density line and a good portion of the red lines have a really similar look as blue lines.




Based on the marginal distribution of age(density plot) and the scatter plot of age versus diameter, I think the imputation quality is pretty good since the generated red dots are close to observed data in scatter plot and the red density lines are pretty good match with observed blue density line.

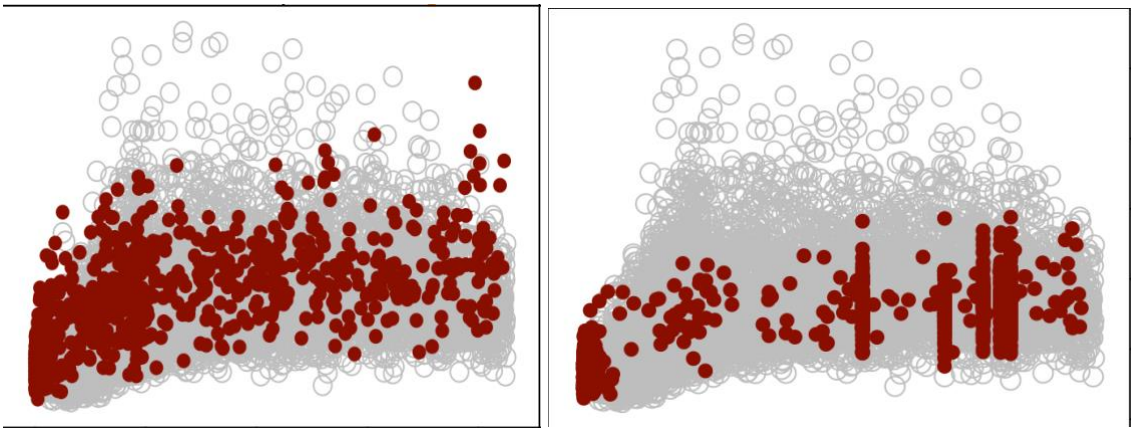
### Part1 - Question 3:

From the summary of linear regression, the age of a tree with 0 diameter is 10.12. With everything staying the same, one unit increase in tree diameter will result in 10.35 years increase in age. Therefore, diameter and age have a positive association and diameter is a significant predictor because the p-value is less than 0.05.

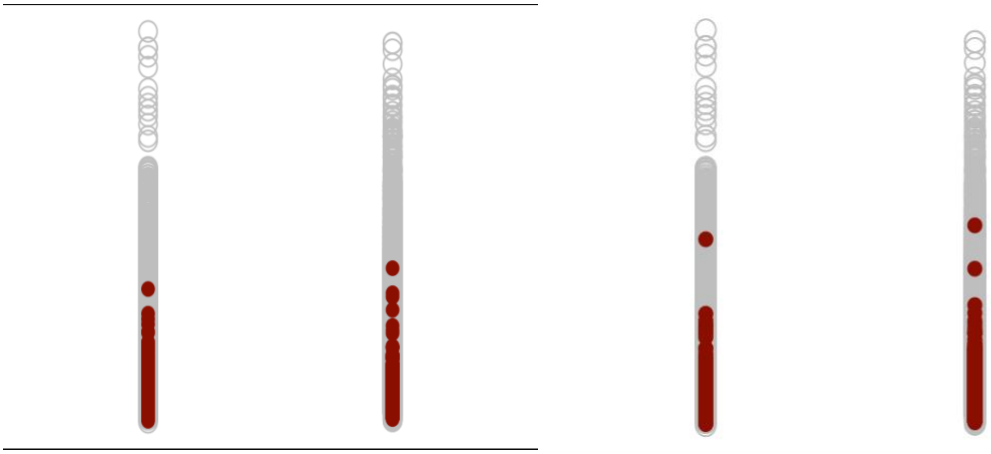
```
lm_imp <- with(data=tree_imp,lm(age ~ diameter))
lm_model <- pool(lm_imp)
summary(lm_model)

##          term estimate std.error statistic       df    p.value
## 1 (Intercept) 10.12073 39.198620  0.258191 7.584723 0.80312650
## 2    diameter 10.34845  3.999022  2.587746 7.967767 0.03233297
```

### Part2 - Question 1:

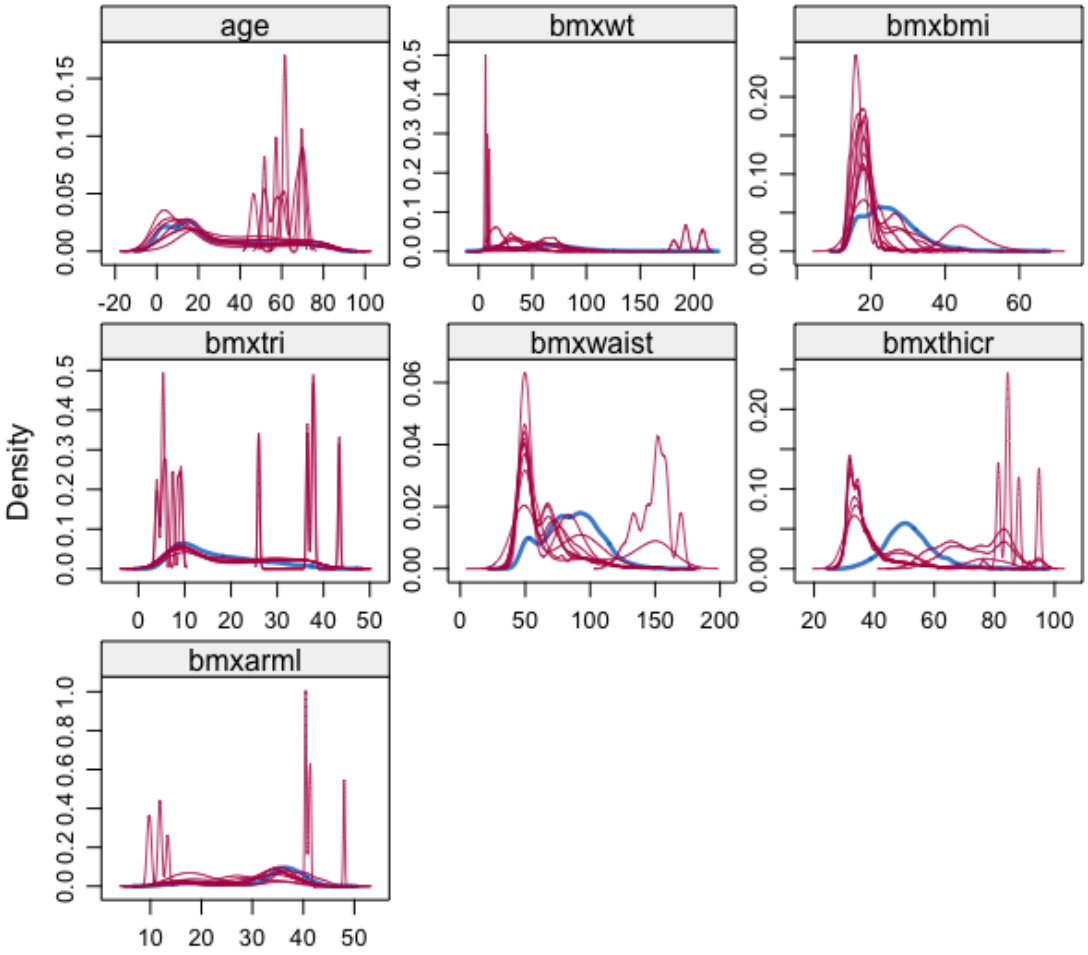
First, let's look at the imputed scatter plots for bmi by age. The imputed data points are mostly within the observed data as you can see.



Then, let's take a look at imputed scatter plot for bmi by gender.Because gender is a binary variable, imputed variables belongs to either one of the category. The imputed points are also among observed data so the imputation is good.



Finally, let's check out the marginal distribution by looking at density plot. For age, as you can see there are multiple red curves(imputed) which have roughly the same shape and range as the blue curve(observed). It indicated the imputed data is good. After eyeballing the 7 graphs, I think the imputation for age,bmxwt, bmxtri and bmxarml are good because the density curves for imputed data share approximately same shapes and positions as original density curve.



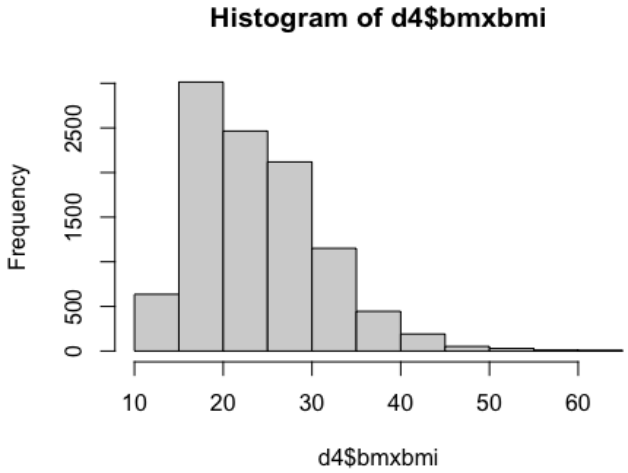

Based on the result of scatter plot and marginal distribution, I think the quality of the imputation is good. Although for some predictors the imputed marginal distribution is off compared to observed marginal distribution, most of the predictors are in good shapes.

### Part2 - Question 2:

I select the 4th imputed dataset, d4, to perform model selection. After performing EDA, I found the response variable is skewed so I took ***log*** on *bmxbmi* and there are no interesting potential interactions from EDA . Then I use AIC to do model selection and the final model is $log(bmxbmi) \sim age + dmdeduc + ridreth2 + riagendr + indfminc$. All the predictors are significant(for some categorical variables, they are at least significant in some levels) in

the model and linearity, normality, Independence and equal variance are hold. Please refer to codes and outputs below.

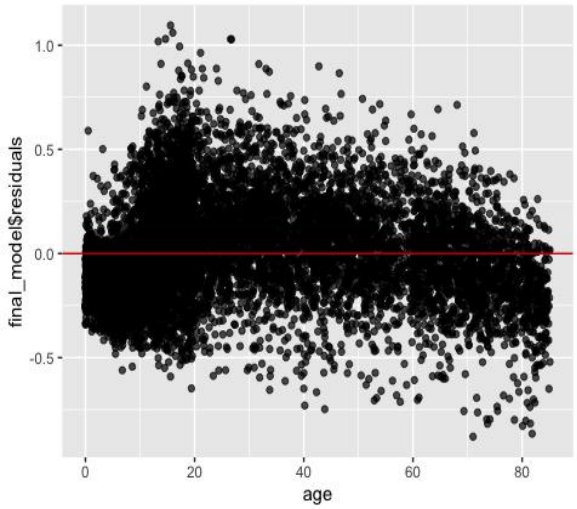
```
# Enter your code for question 5 here
nhanes_imp <- mice(nhanes,m=10,
                   defaultMethod = c("pmm", "logreg", "polyreg", "polr"),prin
t=F)

## Warning: Number of logged events: 512

d4 <- complete(nhanes_imp,4)
# Response variable distribution
hist(d4$bmxbmi)
```



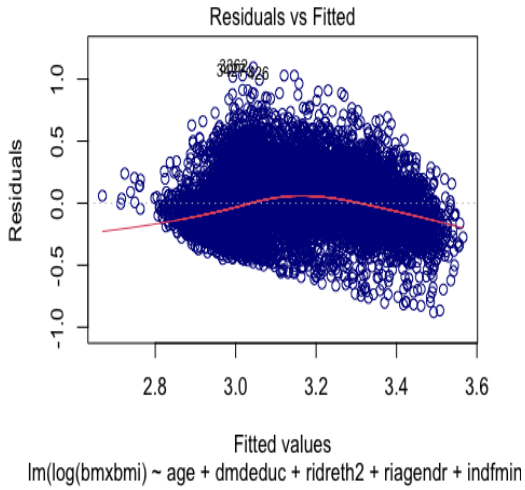

```
form1 <- log(bmxbmi) ~ 1
form2 <- log(bmxbmi) ~ age+riagendr+ridreth2+dmdeduc+indfminc
null_model <- lm(form1,data=d4)
full_model <- lm(form2,data=d4)
final_model <- step(null_model,scope = formula(full_model),direction='forward
',trace = 0)
summary(final_model)

##
## Call:
## lm(formula = log(bmxbmi) ~ age + dmdeduc + ridreth2 + riagendr +
##     indfminc, data = d4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87940 -0.16113 -0.02671  0.13922  1.09493
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.8694962  0.0108253 265.074  < 2e-16 ***
```

```
## age          0.0060436  0.0001166  51.842  < 2e-16 ***
## dmdeduc2     0.1017580  0.0075434  13.490  < 2e-16 ***
## dmdeduc3     0.0996093  0.0067012  14.864  < 2e-16 ***
## dmdeduc7    -0.1579807  0.0799087  -1.977 0.048067 *
## dmdeduc9    -0.1968879  0.0524879  -3.751 0.000177 ***
## ridreth22    0.0719717  0.0061689  11.667  < 2e-16 ***
## ridreth23    0.0671522  0.0064424  10.424  < 2e-16 ***
## ridreth24   -0.0563303  0.0136280  -4.133  3.6e-05 ***
## ridreth25    0.0403979  0.0135946   2.972 0.002969 **
## riagendr2    0.0184437  0.0047543   3.879 0.000105 ***
## indfminc10   0.0160614  0.0155182   1.035 0.300692
## indfminc11   0.0037883  0.0114195   0.332 0.740092
## indfminc12  -0.0247786  0.0229163  -1.081 0.279605
## indfminc13   0.0055878  0.0224283   0.249 0.803257
## indfminc2   -0.0050947  0.0128925  -0.395 0.692728
## indfminc3   -0.0009586  0.0117870  -0.081 0.935184
## indfminc4   -0.0127857  0.0123504  -1.035 0.300579
## indfminc5    0.0082621  0.0122994   0.672 0.501758
## indfminc6   -0.0130887  0.0116462  -1.124 0.261097
## indfminc7    0.0098483  0.0123099   0.800 0.423712
## indfminc77  -0.0254648  0.0273615  -0.931 0.352041
## indfminc8    0.0007329  0.0128011   0.057 0.954342
## indfminc9    0.0427241  0.0145005   2.946 0.003222 **
## indfminc99   0.0296015  0.0279327   1.060 0.289287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2386 on 10097 degrees of freedom
## Multiple R-squared:  0.3349, Adjusted R-squared:  0.3334
## F-statistic: 211.9 on 24 and 10097 DF,  p-value: < 2.2e-16
```

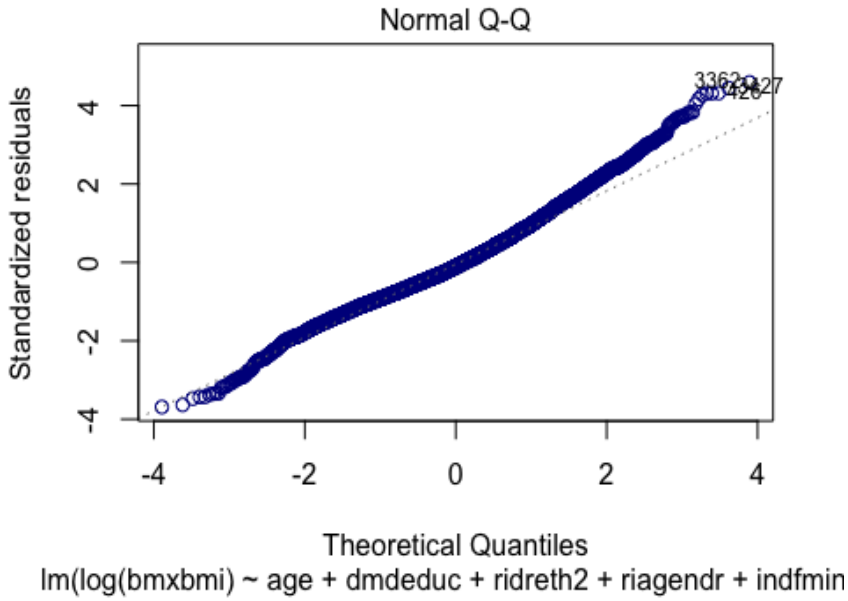
```
ggplot(d4,aes(x=age,y=final_model$residuals))+geom_point(alpha=.7)+
  geom_hline(yintercept=0,col="red3")
```



```r
plot(final_model,which=1,col=c("blue4"))
```




```r
plot(final_model,which=2,col=c("blue4"))
```




Because most of predictors are categorical variables and most of the levels are significant, I'm going to interpret predictors with largest absolute value of effect.

The bmi for a zero age,non-hispanic white male with less than high school education and less than $5000 annual family income is 19.13. Holding other predictors unchanged, as a person's age increase by 1 year, the bmi will increase by 0.5%. Holding other predictors unchanged, the bmi index will decrease by 2% if a person is female. Holding other predictors unchanged, if a person's race is non-Hispanic Black the bmi will increase by 6%

but if a person's race is other race the bmi will decrease by 4.9%. Holding other predictors unchanged, the bmi will increase by 11.6% if a person's education level is high school but the bmi will decrease by 7.7% if a person's education level is 'refused'. The income predictor in the pooled model is not significant so it's not necessary to interpret it.

```
bmireg_imp <- with(data=nhanes_imp,lm(log(bmxbmi) ~ age+riagendr+ridreth2+dmd
educ+indfminc))
lm_model <- pool(bmireg_imp)
summary(lm_model)
```

```
##           term      estimate   std.error   statistic         df      p.value
## 1  (Intercept)  2.9512237150 0.146719463 20.11473910   8.017445 3.791693e-08
## 2          age  0.0051007609 0.002106813  2.42107939   7.207660 4.503492e-02
## 3    riagendr2  0.0206132931 0.006956929  2.96298744  45.293292 4.841381e-03
## 4    ridreth22  0.0584648903 0.030853286  1.89493238   9.725285 8.818062e-02
## 5    ridreth23  0.0545637426 0.019608791  2.78261638  11.757561 1.685055e-02
## 6    ridreth24 -0.0464261512 0.020738940 -2.23859803  37.854060 3.114139e-02
## 7    ridreth25  0.0354590820 0.018403440  1.92676381  72.346226 5.793701e-02
## 8     dmdeduc2  0.1059509573 0.025529365  4.15016034  11.013956 1.611689e-03
## 9     dmdeduc3  0.1013954250 0.022991549  4.41011723  10.962635 1.054037e-03
## 10    dmdeduc7 -0.0846078300 0.176221991 -0.48012073  15.701659 6.377617e-01
## 11    dmdeduc9 -0.0461314012 0.175373385 -0.26304676  11.056299 7.973520e-01
## 12  indfminc10 -0.0021988367 0.038564646 -0.05701690  13.703543 9.553548e-01
## 13  indfminc11 -0.0099257580 0.033527015 -0.29605254  12.029252 7.722387e-01
## 14  indfminc12 -0.0042383058 0.035897961 -0.11806536  34.190507 9.067068e-01
## 15  indfminc13  0.0311528660 0.032925865  0.94615177  44.045387 3.492332e-01
## 16   indfminc2 -0.0034015012 0.014912938 -0.22809061 732.621283 8.196394e-01
## 17   indfminc3 -0.0110693448 0.019581591 -0.56529344  27.937041 5.763863e-01
## 18   indfminc4 -0.0169230760 0.021840910 -0.77483383  23.426615 4.461887e-01
## 19   indfminc5  0.0008716967 0.026334159  0.03310137  16.377576 9.739938e-01
## 20   indfminc6 -0.0220595538 0.019018239 -1.15991566  29.426306 2.554082e-01
## 21   indfminc7  0.0048044071 0.023717052  0.20257185  19.480555 8.415704e-01
## 22  indfminc77 -0.0168032713 0.033190289 -0.50627071 255.703420 6.131028e-01
## 23   indfminc8 -0.0126218223 0.026485635 -0.47655350  17.292318 6.396463e-01
## 24   indfminc9  0.0222724778 0.038093469  0.58467969  13.051141 5.687334e-01
## 25  indfminc99  0.0096947927 0.035640626  0.27201522 124.393046 7.860613e-01
```