Missing Data

Contents

1 Getting Set Up

1.1 Setting chunk options and generating R script

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE)
knitr::purl("missing-data.Rmd")
```

```
##
##
## processing file: missing-data.Rmd

## output file: missing-data.R
```

1.2 Installing Packages

```
install.packages('mice')
install.packages('lattice')
install.packages('VIM')
install.packages('aod')
install.packages('BaM')

library(mice)
library(lattice)
library(VIM)
library(aod)
library(BaM)
```

1.3 Loading in data

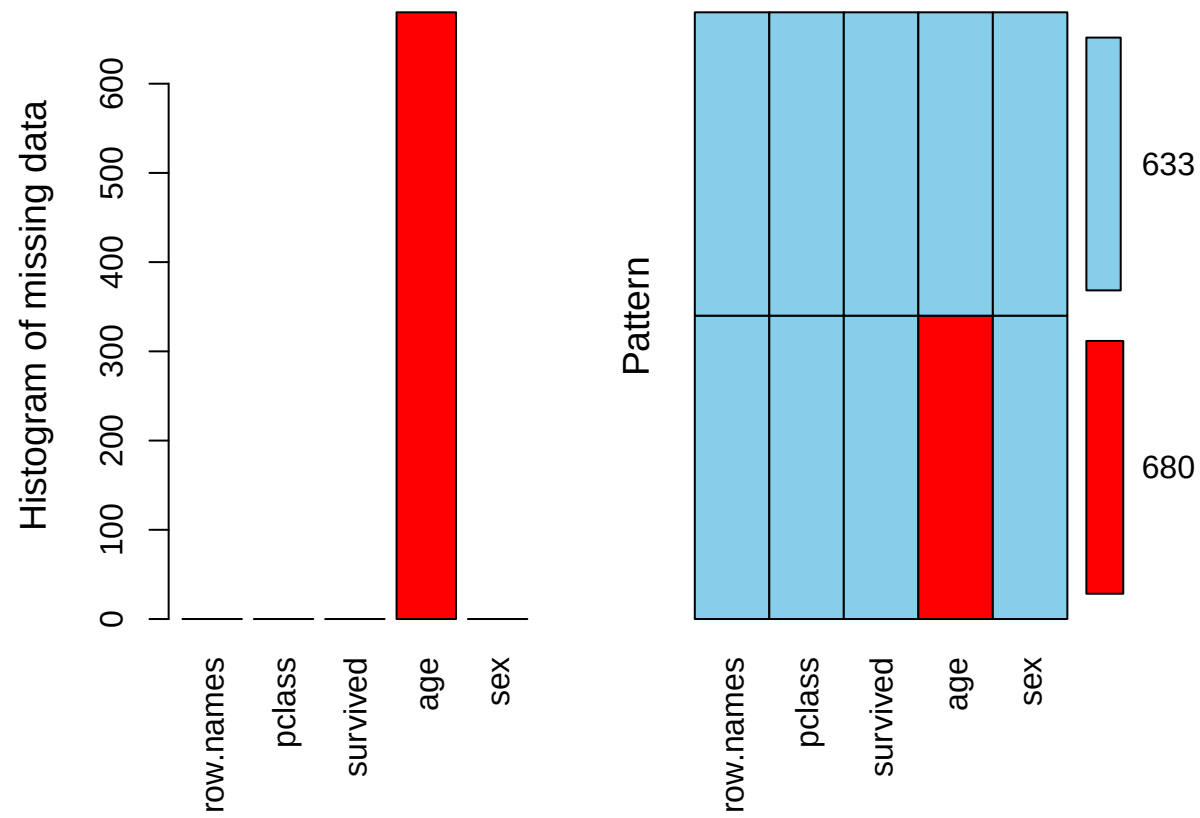
```
titanic = read.table("titanic.txt", header=T, sep=",")
titanic.missing = titanic[,c(1,2,3,5,11)]
head(titanic.missing, 10)
```

```
##    row.names pclass survived     age    sex
## 1          1    1st        1 29.0000 female
## 2          2    1st        0  2.0000 female
## 3          3    1st        0 30.0000   male
## 4          4    1st        0 25.0000 female
## 5          5    1st        1  0.9167   male
## 6          6    1st        1 47.0000   male
## 7          7    1st        1 63.0000 female
## 8          8    1st        0 39.0000   male
## 9          9    1st        1 58.0000 female
## 10        10    1st        0 71.0000   male
```

2 Exploratory Analysis

2.1 Exploring the missingness using VIM

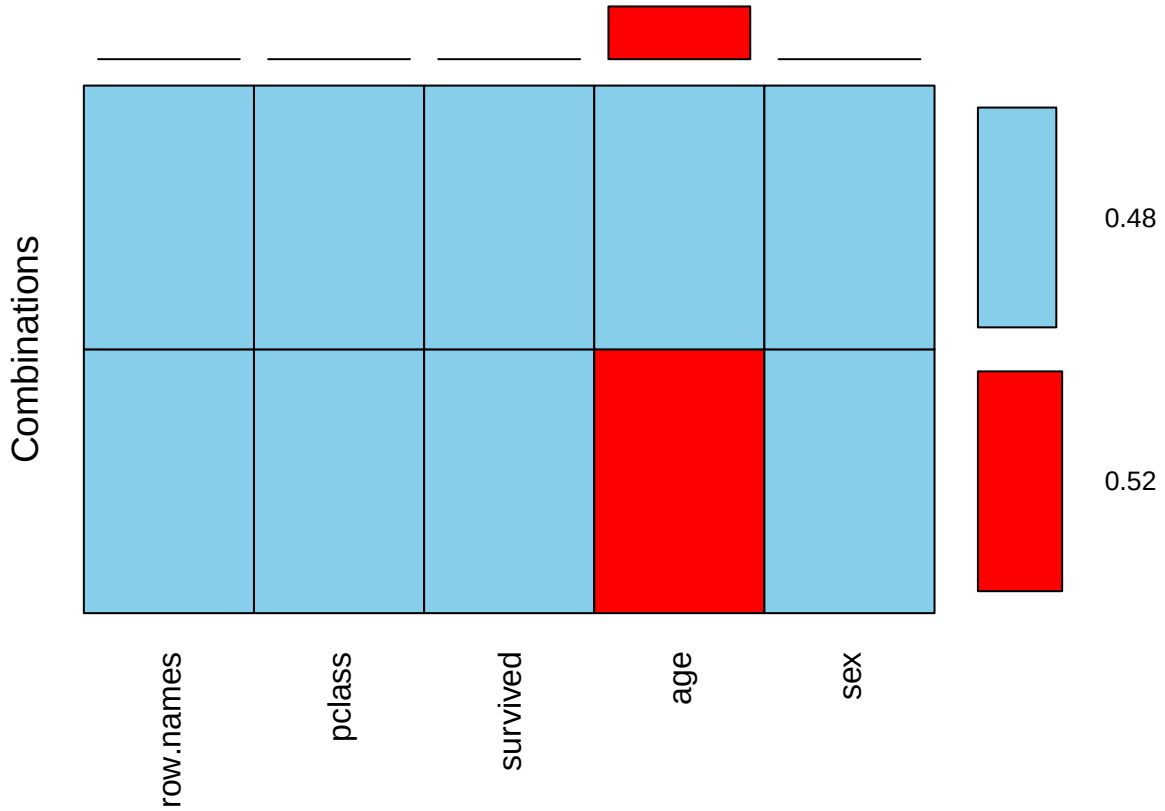
```
titanic.missing.aggr = aggr(titanic.missing, numbers=TRUE, prop=FALSE, ylab=c("Histogram of missing data
```




```
titanic.missing.aggr
```

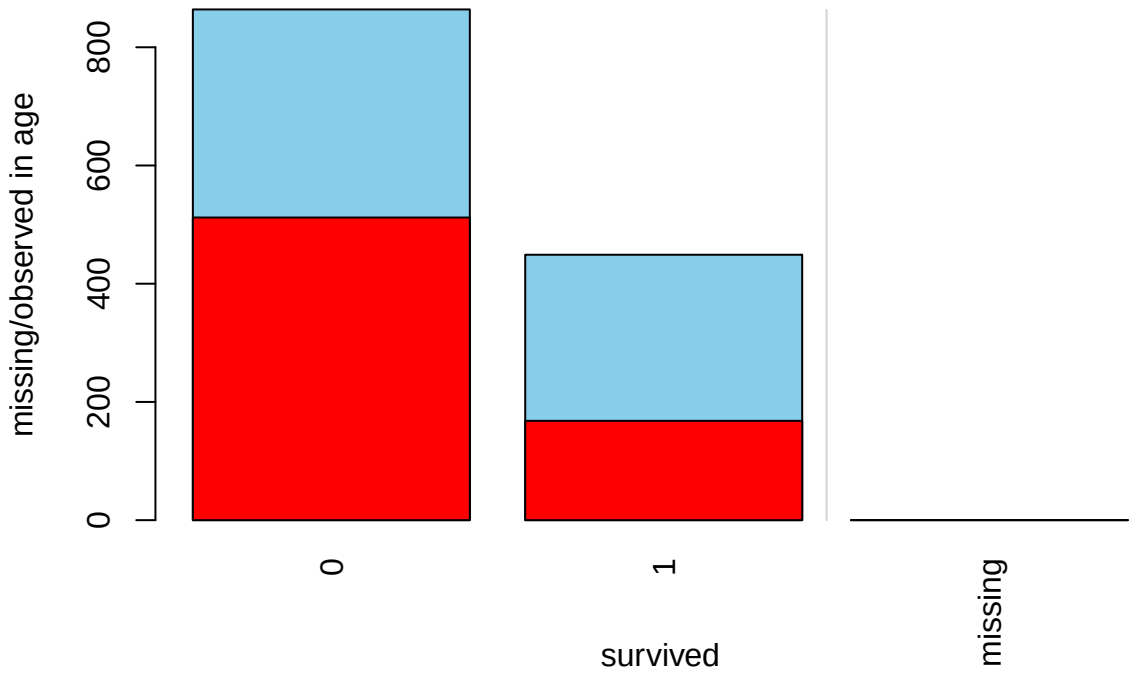
```
##
##  Missings in variables:
##  Variable Count
##       age   680
```

```
aggr(titanic.missing, combined=TRUE, numbers=TRUE, prop=TRUE, cex.numbers=0.87, varheight=FALSE)
```



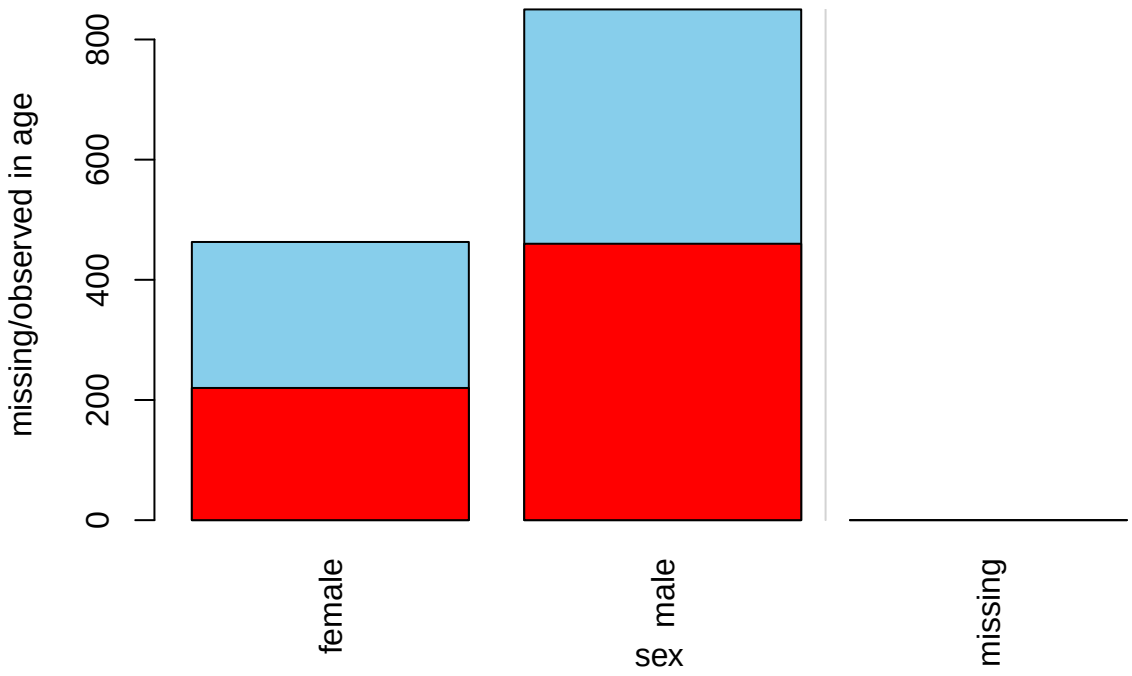
Amount of missingness in age for each survived group.

```
barMiss(titanic.missing[,c("survived","age")])
```

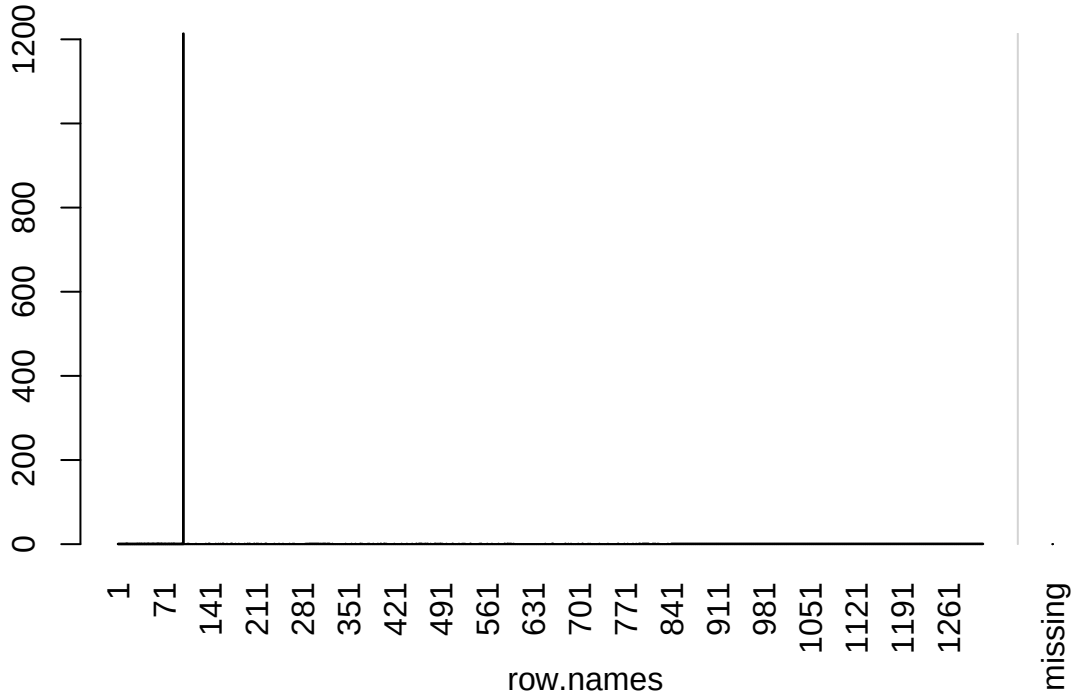


Amount of missigness in age for each sex group

```
barMiss(titanic.missing[,c("sex","age")])
```




```
histMiss(titanic.missing)
```



3 Complete Cases

3.1 Fitting a logistic regression model for the complete cases

```
titanic.logistic.omit = glm(survived ~ pclass+sex+age, family=binomial, data=titanic.missing)
summary(titanic.logistic.omit)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age, family = binomial,
##     data = titanic.missing)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9784  -0.6520  -0.3142   0.5894   2.7022
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.522163   0.471008   9.601  < 2e-16 ***
## pclass2nd   -1.495229   0.281986  -5.302 1.14e-07 ***
## pclass3rd   -2.841271   0.338897  -8.384  < 2e-16 ***
## sexmale     -3.086709   0.241063 -12.805  < 2e-16 ***
```

```
## age         -0.049309   0.008732  -5.647 1.63e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 869.54  on 632  degrees of freedom
## Residual deviance: 539.71  on 628  degrees of freedom
##   (680 observations deleted due to missingness)
## AIC: 549.71
##
## Number of Fisher Scoring iterations: 5
```

3.2 Global effect of class

```
wald.test(b=coef(titanic.logistic.omit), Sigma=vcov(titanic.logistic.omit), Terms=2:3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 70.6, df = 2, P(> X2) = 4.4e-16
```

3.3 Odds ratios

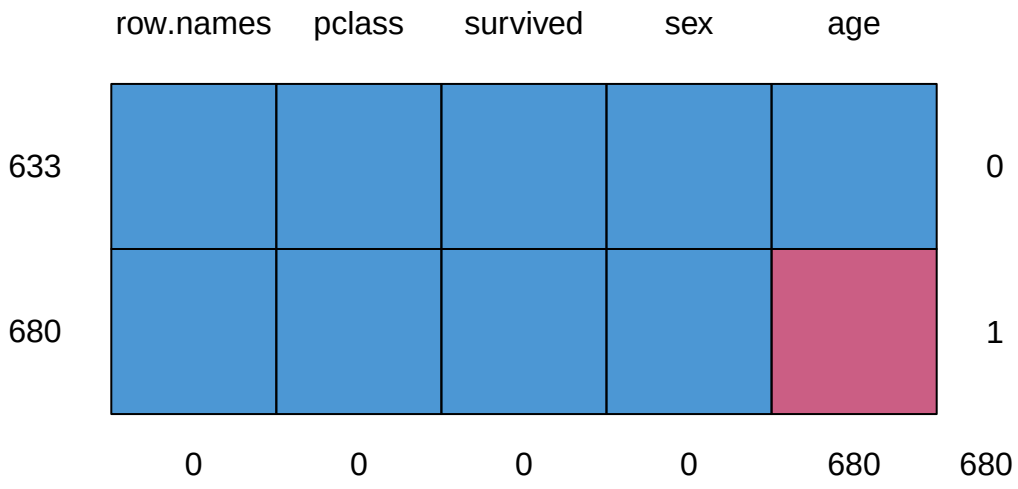
```
exp(cbind(OR=titanic.logistic.omit$coefficients, confint(titanic.logistic.omit)))
```

```
##                       OR       2.5 %       97.5 %
## (Intercept) 92.03444459 37.77003177 239.96152852
## pclass2nd    0.22419723  0.12754708   0.38595627
## pclass3rd    0.05835143  0.02938342   0.11121310
## sexmale      0.04565195  0.02800873   0.07219361
## age          0.95188735  0.93534032   0.96796827
```

4 Multiple Imputation

4.1 Studying the patterns of missingness

```
pattern = md.pattern(titanic.missing)
```




```
pattern
```

```
##     row.names pclass survived sex age
## 633         1      1        1   1   1   0
## 680         1      1        1   1   0   1
##             0      0        0   0 680 680
```

```
pairs = md.pairs(titanic.missing)
pairs
```

```
## $rr
##           row.names pclass survived age  sex
## row.names      1313   1313     1313 633 1313
## pclass         1313   1313     1313 633 1313
## survived       1313   1313     1313 633 1313
## age             633    633      633 633  633
## sex            1313   1313     1313 633 1313
##
## $rm
##           row.names pclass survived age sex
## row.names         0      0        0 680   0
## pclass            0      0        0 680   0
## survived          0      0        0 680   0
## age               0      0        0   0   0
## sex               0      0        0 680   0
```

```
##
## $mr
##           row.names pclass survived age sex
## row.names         0      0        0   0   0
## pclass            0      0        0   0   0
## survived          0      0        0   0   0
## age             680    680      680   0 680
## sex               0      0        0   0   0
##
## $mm
##           row.names pclass survived age sex
## row.names         0      0        0   0   0
## pclass            0      0        0   0   0
## survived          0      0        0   0   0
## age               0      0        0 680   0
## sex               0      0        0   0   0
```

4.2 Imputing the missing values

```
imp = mice(titanic.missing, m=100)
imp
```

4.3 Imputed values for age.

Each row corresponds to a missing entry in age. The columns contain the multiple imputations.

```
imp$imp$age[1:10,1:5]
```

```
##     1  2  3  4  5
## 13 13 19 22 23 50
## 14 49 27 31 38 52
## 15 48 58 24 45 29
## 30 36 58 47 33 30
## 33 49 26 31 26 36
## 36 36 46 47 49 46
## 41 49 47 49 18 52
## 46 70 46 36 46 28
## 47 70 58 45 46 28
## 53 47 33 19 24 21
```

4.4 Combine observed and imputed data.

Only first ten passengers shown.

```
complete(imp,1)[1:10,]
```

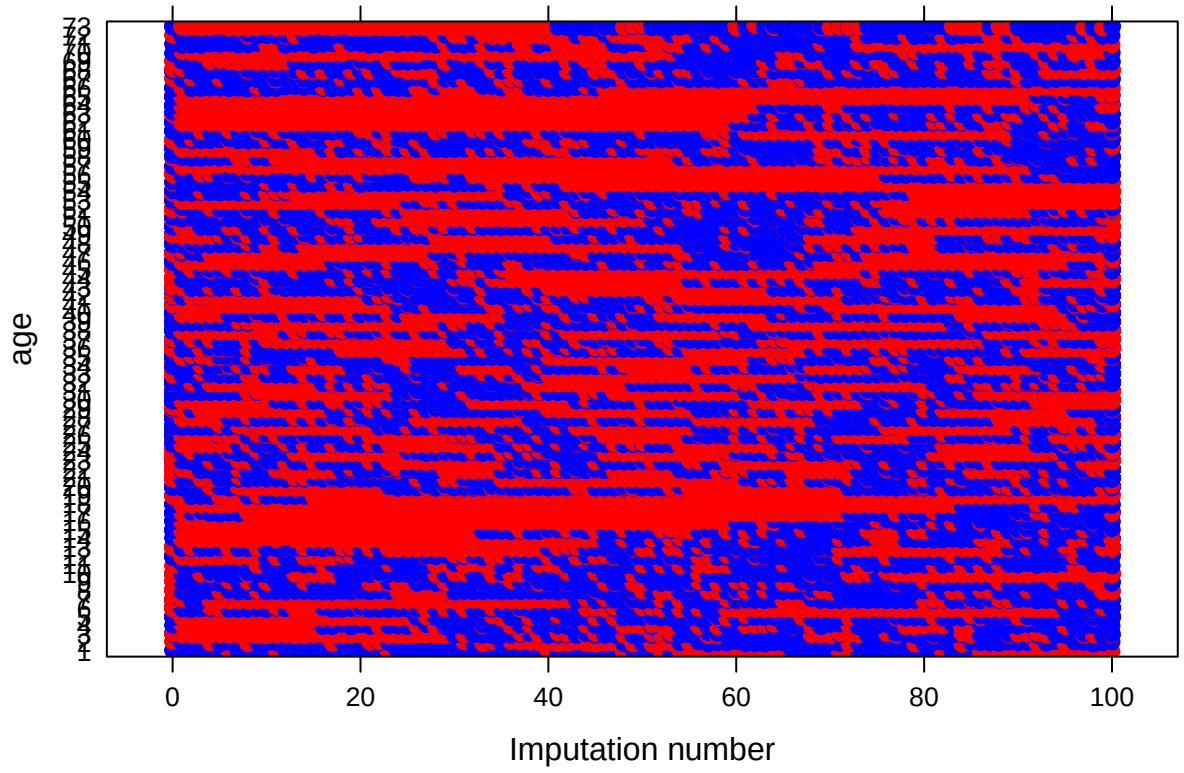
```
##    row.names pclass survived     age    sex
## 1          1    1st        1 29.0000 female
## 2          2    1st        0  2.0000 female
```

```
## 3          3    1st        0 30.0000   male
## 4          4    1st        0 25.0000 female
## 5          5    1st        1  0.9167   male
## 6          6    1st        1 47.0000   male
## 7          7    1st        1 63.0000 female
## 8          8    1st        0 39.0000   male
## 9          9    1st        1 58.0000 female
## 10        10    1st        0 71.0000   male
```

4.5 Diagnostic Checking

It is often useful to inspect the distributions of original and the imputed data. The complete() function extracts the original and the imputed data sets from the imp object as a long (row-stacked) matrix. The col vector separates the observed (blue) and imputed (red) data for age.

```
com = complete(imp, "long", inc=T)
col = rep(c("blue","red")[1+as.numeric(is.na(imp$data$age))],101)
stripplot(age~.imp, data=com, jit=TRUE, fac=0.8, col=col, pch=20, cex=1.4, xlab="Imputation number")
```



5 Analyses of Imputed Data

```
# Analyzing the imputed data sets
fit = with(data=imp, exp=glm(survived ~ pclass + sex + age, family=binomial))

# Creating a data set with the results of all the analysis
MI.matrix<-matrix(0,100,5)
for(k in 1:100) {
  MI.matrix[k,] = coefficients(fit$analyses[[k]])
  MI.results = data.frame(Intercept=MI.matrix[,1], pclass2=MI.matrix[,2], pclass3=MI.matrix[,3], sex=MI
}

MI.results[1:10,]
```

```
##    Intercept   pclass2   pclass3       sex         age
## 1   4.424648 -1.615221 -3.118535 -2.575062 -0.05842351
## 2   4.638783 -1.632769 -3.186411 -2.640705 -0.06267951
## 3   3.511878 -1.293492 -2.733150 -2.422880 -0.03908010
## 4   4.035999 -1.448908 -2.913417 -2.438252 -0.05124997
## 5   4.019575 -1.432094 -2.971920 -2.489937 -0.04967263
## 6   4.414237 -1.586121 -3.050886 -2.521151 -0.05747052
## 7   4.035444 -1.420229 -2.630619 -2.586655 -0.04920925
## 8   4.453767 -1.580213 -3.031338 -2.564475 -0.05941595
## 9   3.538763 -1.266417 -2.579937 -2.499612 -0.03816236
## 10  3.944591 -1.411176 -3.052917 -2.527078 -0.04753117
```

5.1 Combining the results using Rubin's rule

The column fmi contains the fraction of missing information, i.e. the proportion of the variability that is attributable to the uncertainty caused by the missing data.

```
est = pool(fit)
summary(est)
```

```
##          term    estimate   std.error  statistic       df      p.value
## 1 (Intercept)  4.06113601 0.445787373   9.110029 281.6264 0.000000e+00
## 2   pclass2nd -1.44106632 0.241718943  -5.961743 841.1775 3.669940e-09
## 3   pclass3rd -2.88717091 0.254027101 -11.365602 592.5140 0.000000e+00
## 4     sexmale -2.53512473 0.171418630 -14.789085 937.8208 0.000000e+00
## 5         age -0.05061867 0.009089018  -5.569213 218.2134 7.484449e-08
```

6 Inverse Probability Weighting

6.1 Creating the missing data indicator variable r

```
titanic.missing$r = as.numeric(!is.na(titanic.missing$age))*as.numeric(!is.na(titanic.missing$sex))
head(titanic.missing,15)
```

```
##    row.names pclass survived     age    sex r
## 1          1    1st        1 29.0000 female 1
```

```
## 2          2    1st        0  2.0000 female 1
## 3          3    1st        0 30.0000   male 1
## 4          4    1st        0 25.0000 female 1
## 5          5    1st        1  0.9167   male 1
## 6          6    1st        1 47.0000   male 1
## 7          7    1st        1 63.0000 female 1
## 8          8    1st        0 39.0000   male 1
## 9          9    1st        1 58.0000 female 1
## 10        10    1st        0 71.0000   male 1
## 11        11    1st        0 47.0000   male 1
## 12        12    1st        1 19.0000 female 1
## 13        13    1st        1      NA female 0
## 14        14    1st        1      NA   male 0
## 15        15    1st        0      NA   male 0
```

6.2 Fitting the logistic regression model to calculate the probabilities of being complete

```
titanic.ipw.glm<-glm(r ~ pclass + survived, data=titanic.missing,family=binomial)
summary(titanic.ipw.glm)
```

```
##
## Call:
## glm(formula = r ~ pclass + survived, family = binomial, data = titanic.missing)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7750  -0.7768  -0.7768   0.7923   1.6404
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6547     0.1451   4.511 6.45e-06 ***
## pclass2nd     0.3430     0.1874   1.830   0.0672 .
## pclass3rd    -1.6985     0.1561 -10.878  < 2e-16 ***
## survived      0.3457     0.1382   2.502   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1818.5  on 1312  degrees of freedom
## Residual deviance: 1532.0  on 1309  degrees of freedom
## AIC: 1540
##
## Number of Fisher Scoring iterations: 4
```

6.3 Calculating the weights: Inverse Probabilities

```
titanic.missing$w<-1/fitted(titanic.ipw.glm)
head(titanic.missing, 15)
```

```
##    row.names pclass survived     age    sex r        w
## 1          1    1st        1 29.0000 female 1 1.367731
## 2          2    1st        0  2.0000 female 1 1.519607
## 3          3    1st        0 30.0000   male 1 1.519607
## 4          4    1st        0 25.0000 female 1 1.519607
## 5          5    1st        1  0.9167   male 1 1.367731
## 6          6    1st        1 47.0000   male 1 1.367731
## 7          7    1st        1 63.0000 female 1 1.367731
## 8          8    1st        0 39.0000   male 1 1.519607
## 9          9    1st        1 58.0000 female 1 1.367731
## 10        10    1st        0 71.0000   male 1 1.519607
## 11        11    1st        0 47.0000   male 1 1.519607
## 12        12    1st        1 19.0000 female 1 1.367731
## 13        13    1st        1      NA female 0 1.367731
## 14        14    1st        1      NA   male 0 1.367731
## 15        15    1st        0      NA   male 0 1.519607
```

6.4 Generating Model

```
titanic.results.ipw = glm(survived ~ pclass + sex + age, data=titanic.missing, weights=titanic.missing$w
summary(titanic.results.ipw)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age, family = binomial,
##     data = titanic.missing, weights = titanic.missing$w)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3821  -0.8499  -0.5710   0.7741   4.6167
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.825513   0.323628  11.821  < 2e-16 ***
## pclass2nd   -1.314875   0.221127  -5.946 2.74e-09 ***
## pclass3rd   -2.758631   0.221916 -12.431  < 2e-16 ***
## sexmale     -2.657877   0.159649 -16.648  < 2e-16 ***
## age         -0.042672   0.006179  -6.906 4.99e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1687.4  on 632  degrees of freedom
## Residual deviance: 1110.7  on 628  degrees of freedom
##   (680 observations deleted due to missingness)
## AIC: 1086.7
##
## Number of Fisher Scoring iterations: 4
```