Impute missing data

Thuy Nguyen

5/25/2022

Explore the missingness

```
##        id           resp             age             smok             time
##  Min.   :  1   Min.   :0.0000   Min.   :-2.00   Min.   :0.0000   Min.   :1.00
##  1st Qu.:135   1st Qu.:0.0000   1st Qu.:-1.25   1st Qu.:0.0000   1st Qu.:1.75
##  Median :269   Median :0.0000   Median :-0.50   Median :0.0000   Median :2.50
##  Mean   :269   Mean   :0.1553   Mean   :-0.50   Mean   :0.3482   Mean   :2.50
##  3rd Qu.:403   3rd Qu.:0.0000   3rd Qu.: 0.25   3rd Qu.:1.0000   3rd Qu.:3.25
##  Max.   :537   Max.   :1.0000   Max.   : 1.00   Max.   :1.0000   Max.   :4.00
##                NA's   :229
```

Above is the summary of the data sixcity.Only resp column has 229 missning data.

```
## `summarise()` has grouped output by 'time'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 8 x 5
## # Groups:   time [4]
##    time  smok     N n_missing percent_missing
##   <int> <int> <int>     <int>           <dbl>
## 1     1     0   350        16           0.046
## 2     1     1   187         4           0.021
## 3     2     0   350        20           0.057
## 4     2     1   187        11           0.059
## 5     3     0   350        50           0.143
## 6     3     1   187        26           0.139
## 7     4     0   350        55           0.157
## 8     4     1   187        47           0.251
```

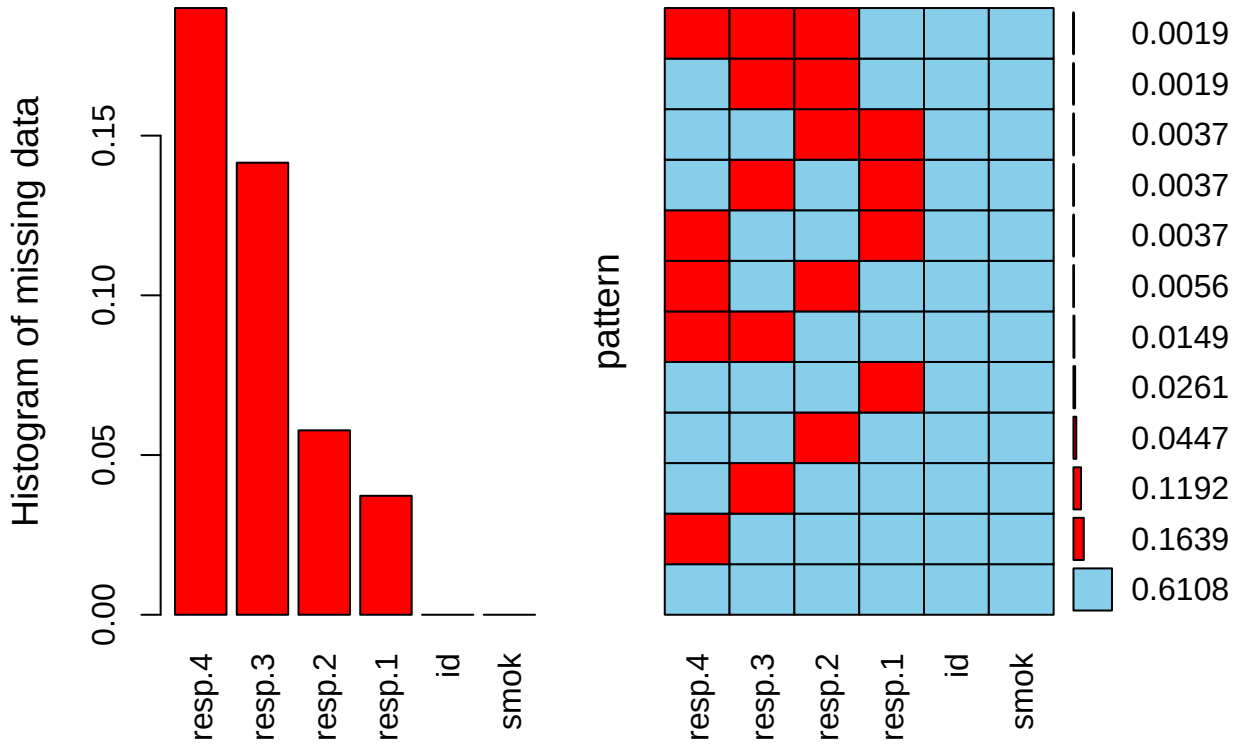
Above is the proportion of missing data by smoke and time visit

```
## `summarise()` has grouped output by 'smok'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'smok'. You can override using the
## `.groups` argument.
```

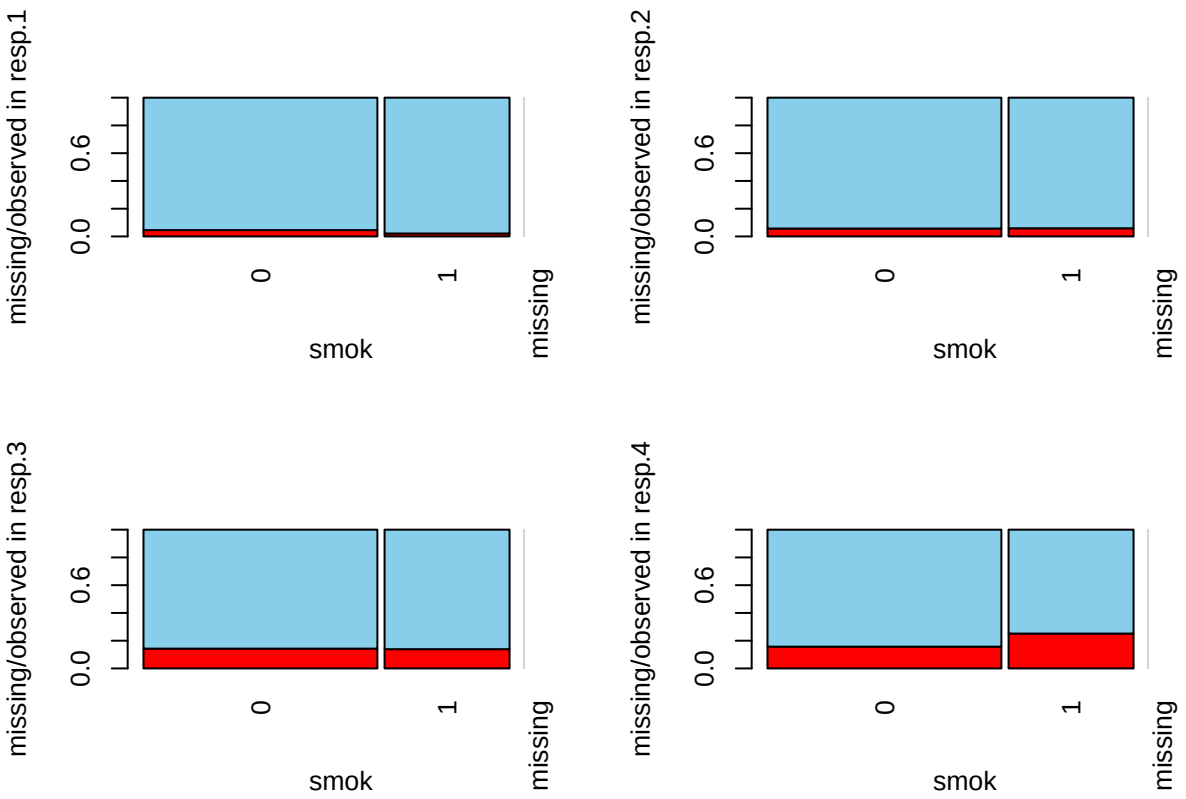
```
## # A tibble: 7 x 3
## # Groups:   smok [2]
##    smok nvisit     N
```

```
##   <int>  <int> <int>
## 1     0      2    11
## 2     0      3   119
## 3     0      4   220
## 4     1      1     1
## 5     1      2     7
## 6     1      3    71
## 7     1      4   108
```

The table above shows the total of times people showing up and total of people in each categorical time by smoke status.




```
##
##  Variables sorted by number of missings:
##  Variable      Count
##    resp.4 0.18994413
##    resp.3 0.14152700
##    resp.2 0.05772812
##    resp.1 0.03724395
##        id 0.00000000
##      smok 0.00000000
```



The figure on the left shows a histogram of missing data by column in our dataset. When we reshape to wide form, resp4 has 18% missing data, resp3 has about 14% missing data and so on. The completed observations are 61%. The figure on the right show that the pattern of missing data is non-monotone. Some people missed one follow-up but came back to the next follow-ups.

1. Fitting random effects model for respiratory function with only the main effects of smoking status and age using complete data

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]
##  Family: binomial  ( logit )
## Formula: resp ~ age * smok + (1 | id)
##    Data: sixcity_comp
##
##      AIC      BIC   logLik deviance df.resid
##   1604.7   1633.1   -797.4   1594.7     2143
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3660 -0.1988 -0.1784 -0.1437  2.5930
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 4.693    2.166
## Number of obs: 2148, groups:  id, 537
##
```

```
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.12822    0.22284 -14.038   <2e-16 ***
## age         -0.21638    0.08656  -2.500   0.0124 *
## smok         0.46197    0.28556   1.618   0.1057
## age:smok     0.10533    0.13850   0.761   0.4469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) age    smok
## age       0.301
## smok     -0.517 -0.199
## age:smok -0.171 -0.623  0.290
```

2: Using available data to fit a random effects model for respiratory function with only the main effects of smoking status and age.

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]
##  Family: binomial  ( logit )
## Formula: resp ~ age * smok + (1 | id)
##    Data: sixcity
##
##      AIC      BIC   logLik deviance df.resid
##   1468.7   1496.5   -729.4   1458.7     1914
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3146 -0.2053 -0.1930 -0.1585  2.4846
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 4.676    2.162
## Number of obs: 1919, groups:  id, 537
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.08476    0.23204 -13.294   <2e-16 ***
## age         -0.16554    0.09191  -1.801   0.0717 .
## smok         0.56106    0.29500   1.902   0.0572 .
## age:smok     0.13224    0.14883   0.889   0.3742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) age    smok
## age       0.311
## smok     -0.526 -0.214
## age:smok -0.197 -0.618  0.336
```

3. Imputations using long format

```
## Warning: Number of logged events: 1
```

```
##                   term   estimate  std.error   statistic        df   p.value
## 1          (Intercept) 0.05485132 0.20929513 -13.8709818  766.5492 0.0000000
## 2                  age 0.84904656 0.08684976  -1.8841877 1367.2590 0.0597516
## 3     as.factor(smok)1 1.53429961 0.26988152   1.5861552  867.4336 0.1130685
## 4 age:as.factor(smok)1 1.02131026 0.14013569   0.1504711 1034.8783 0.8804223
##        2.5 %     97.5 %
## 1 0.03637081 0.08272203
## 2 0.71604342 1.00675469
## 3 0.90336991 2.60588189
## 4 0.77577236 1.34456277
```

4. Imputations using wide format

Fit glmer model to each completed dataset and pool the result

```
##                            pooled.estimates  pooled.se    pooled.pv
## coef.(Intercept)                -3.0500091 0.23764069 1.050239e-37
## coef.age                        -0.1658666 0.09570334 8.307196e-02
## coef.as.factor(smok)1            0.5501984 0.29493966 6.211683e-02
## coef.age:as.factor(smok)1        0.1263982 0.14603794 3.867556e-01
```

Comparasons:

```
##         ...          complete          imp_long         imp_wide
## 1 Intercept -3.0848 (0.2320) 0.05485 (0.20930) -3.0500 (0.2376)
## 2       age -0.1655 (0.0919) 0.84905 (0.08685) -0.1659 (0.0957)
## 3     smoke  0.5611 (0.2950) 1.53430 (0.26988) 0.5502  (0.2949)
## 4 age:smoke  0.1322 (0.1488) 1.02131 (0.14014)  0.1264 (0.1460)
```

Our result from fitting a random effect model are fairly consistent across three methods, complete data, imputations using data in long format,and wide format, though there are some differences. The complete data model and imputations using wide format model are the most similar in terms of results. The table above shows point estimates for the regression coefficients(standard errors).