Missing data and Multiple Imputation

Henrik Støvring & Morten Frydenberg

stovring@ph.au.dk – morten@ph.au.dk

December 12, 2016 – Aarhus University

AARHUS UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH

Overview

- The ordinary analysis
- A first example of an MI-based analysis
- Causes for the missing values – the types of missingness
- Imputation – by hand and automated using MICE
- Analysis of imputed datasets
- Stata's commands for analyzing MI-data
- Skeleton of an analytic strategy

Case I: Wood dust, Jacobsen (2008)

Eur Respir J 2008; 31: 334–342
DOI: 10.1183/09031936.00146806

Longitudinal lung function decline and wood dust exposure in the furniture industry

G. Jacobsen*,#, V. Schlünssen#, I. Schaumburg*, E. Taudorf¶ and T. Sigsgaard#

ABSTRACT: The aim of the present study was to investigate the relationship between change in lung function and cumulative exposure to wood dust.

In total, 1,112 woodworkers (927 males, 185 females) and 235 reference workers (104 males, 185 females) participated in a 6-yr longitudinal study. Forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), height and weight were measured, and questionnaire data on respiratory symptoms, wood dust exposure and smoking habits were collected. Cumulative inhalable wood dust exposure was assessed using a study-specific job exposure matrix and exposure time.

The median (range) for cumulative wood dust exposure was 3.75 (0–7.55) mg·year·m^{-3}. A dose–response relationship between cumulative wood dust exposure and percent annual decrease in FEV_1 was suggested for female workers. This was confirmed in a linear regression model adjusted

AFFILIATIONS
*Dept of Occupational Medicine, Region Hospital Skive, Skive,
#Dept of Occupational and Environmental Medicine, Institute of Public Health, Århus University, Århus, and
¶Dept of Respiratory Medicine, Hilleroed Hospital, Hilleroed, Denmark.

CORRESPONDENCE
G. Jacobsen
Dept of Occupational Medicine

Exercise 1

- Identify the main outcome variable in the paper – and in the dataset *Wooddata1.dta*
- Identify the main exposure variable in the paper – and in the dataset
- Identify the main confounders in the paper and in the dataset
- Do the variables have missing values?
- Change all values coded '.a', '.b', etc to '.'

Case II: Compliance, Larsen (2009)

BMC Public Health

Research article

Open Access

Can differences in medical drug compliance between European countries be explained by social factors: analyses based on data from the European Social Survey, round 2

John Larsen*, Henrik Stovring, Jakob Kragstrup and Dorte G Hansen

Address: Research Unit of General Practice, Institute of Public Health, University of Southern Denmark, Odense, Denmark

Email: John Larsen* - jlarsen@health.sdu.dk; Henrik Stovring - hstovring@health.sdu.dk ; Jakob Kragstrup - jkragstrup@health.sdu.dk; Dorte G Hansen - dgilsaa@health.sdu.dk

* Corresponding author

Published: 16 May 2009

Received: 4 July 2008
Accepted: 16 May 2009

BMC Public Health 2009, **9**:145 doi:10.1186/1471-2458-9-145

This article is available from: http://www.biomedcentral.com/1471-2458/9/145

European Social Survey, Round 2

- Questionnaire on income, political opinion, employment, religion, health, etc
- 24 participating countries scattered over Europe
- Data publicly available at http://ess.nsd.uib.no, ESS Round 2 as well as round 1-7
- The primary focus of Larsen (2009) was on compliance:
  - Primary non-compliance (did not collect prescription)
  - Secondary non-compliance (did not take prescription as prescribed)
  - Non-compliance: primary and/or secondary non-compliance
- Here we focus on the Scandinavian countries, *ess2e03_scand.dta*

AARHUS UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH

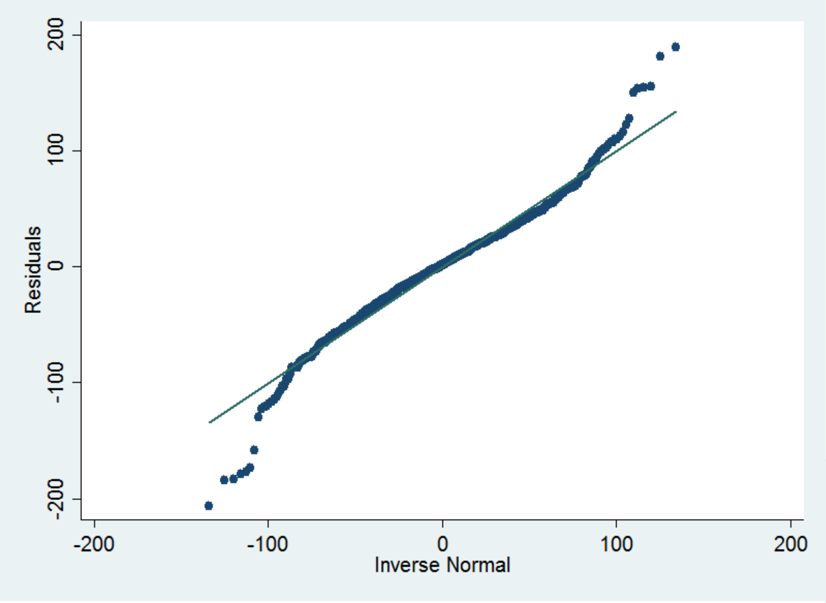
Wood dust: Complete case analysis

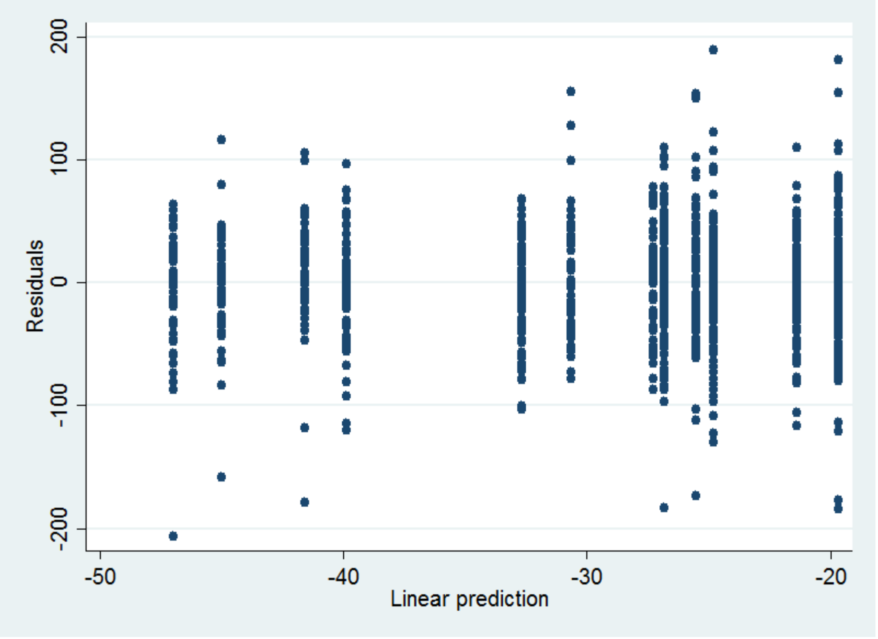
```
. regress fev1aendaar i.wooddustgrp i.packryg

      Source |       SS           df       MS      Number of obs   =     1,216
-------------+----------------------------------   F(5, 1210)      =      8.76
       Model |  79776.3293         5  15955.2659   Prob > F        =    0.0000
    Residual |  2203100.28     1,210  1820.74404   R-squared       =    0.0349
-------------+----------------------------------   Adj R-squared   =    0.0310
       Total |  2282876.61     1,215  1878.91079   Root MSE        =     42.67

------------------------------------------------------------------------------
 fev1aendaar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 wooddustgrp |
      2.97-  |  -5.126245   3.349532    -1.53   0.126    -11.69778     1.44529
      3.75-  |  -7.119821   3.298705    -2.16   0.031    -13.59164   -.6480041
      4.72-  |   -1.72048   3.397087    -0.51   0.613    -8.385315    4.944354
             |
     packryg |
< 7 packy..  |  -5.856692   2.923619    -2.00   0.045    -11.59262   -.1207655
>=7 packy..  |  -20.17586   3.220615    -6.26   0.000    -26.49447   -13.85725
             |
       _cons |  -19.68595   2.319777    -8.49   0.000    -24.23718   -15.13472
------------------------------------------------------------------------------
```

Complete case analysis - diagnostics





AARHUS UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH

Exercise 2

- Refer to the flow chart of Jacobsen et al (2008)
- Is it complete? If not, what is missing?
- Open the dataset – why are only 1,216 workers included in the regression analysis when 1,347 workers were included in the study?

Exercise 2 – sensitivity analyses

- 12 sub-groups
- Each group imputes a combination of wood dust and smoking (packryg) according to this table:

Wood dust exposure	Non-smoker	<7 pack years	>7 pack years
0-	1	2	3
2.97-	4	5	6
3.75-	7	8	9
4.72	10	11	12

- Report results on the ~~black board~~ https://docs.google.com/spreadsheets/d/11GxmvTPvH2_xbj93lGqgrBuwh13rnTR3JKzuyrv_PLY/edit?usp=sharing

MI analysis - preview

```
. replace packryg = . if packryg == .a
(95 real changes made, 95 to missing)

. mi set flong

. mi register imputed wooddustgrp packryg
(131 m=0 obs. now marked as incomplete)

. mi impute chained (mlogit) wooddustgrp packryg = fev1aendaar,
add(10)

Conditional models:
       wooddustgrp: mlogit wooddustgrp i.packryg fev1aendaar
           packryg: mlogit packryg i.wooddustgrp fev1aendaar

Performing chained iterations ...

Multivariate imputation                     Imputations =       10
Chained equations                                 added =       10
Imputed: m=1 through m=10                       updated =        0
```

MI analysis - preview

(Continued…)

```
Initialization: monotone                             Iterations =       100
                                                        burn-in =        10

       wooddustgrp: multinomial logistic regression
           packryg: multinomial logistic regression

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
       wooddustgrp |       1263           84        84 |      1347
           packryg |       1252           95        95 |      1347
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

AARHUS UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH

MI analysis - preview

```
. mi estimate: regress fev1aendaar i.wooddustgrp i.packryg

Multiple-imputation estimates                   Imputations       =         10
Linear regression                               Number of obs     =      1,347
                                                Average RVI       =     0.1274
                                                Largest FMI       =     0.1748
                                                Complete DF       =       1341
DF adjustment:   Small sample                   DF:     min       =     244.84
                                                        avg       =     477.61
                                                        max       =     691.20
Model F test:       Equal FMI                   F(   5,   825.8)  =       8.25
Within VCE type:          OLS                   Prob > F          =     0.0000

--------------------------------------------------------------------------------
   fev1aendaar |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
   wooddustgrp |
        2.97-  |  -6.106555   3.506857    -1.74   0.083    -13.00995    .7968414
        3.75-  |  -6.807146   3.492523    -1.95   0.052    -13.68637    .0720782
        4.72-  |  -2.624091   3.406981    -0.77   0.441    -9.313767    4.065585
               |
       packryg |
< 7 packyears  |  -6.559535   3.022964    -2.17   0.031    -12.50421   -.6148639
>=7 packyears  |  -20.27876   3.236952    -6.26   0.000     -26.6342   -13.92333
               |
         _cons |  -19.43084   2.328201    -8.35   0.000    -24.00296   -14.85872
--------------------------------------------------------------------------------
```

What causes the missing data?

- Known as the *missing data mechanism*
- Was lung function not measured due to a defect in the measuring device?
- Is smoking unrecorded because the interviewer did not recognize the question?
- Was exposure not measured on a subset of factories?
- More interesting:
  - Do heavy smokers not report smoking?
  - Are those most exposed those for whom exposure is unknown?
- How it can be examined:
  - Step 1: Define a 0/1 variable indicating missingness for smoking and wood dust exposure
  - Step 2: Analyze the 0/1 variables as outcome variables

Exercise 3

- Examine why information on smoking and wood dust exposure are missing using the algorithm on the previous slide.

Types of missing data – a taxonomy

- **MCAR: Missing Completely At Random**
  Whether a value is missing has no relation with its value or any other of the values in the dataset.
- **MAR: Missing At Random**
  Whether a value is missing depends on the other observed values for the person, but once we know those, the value being missing does not depend on being missing or not (it has the same distribution as those observed, given the other observed values)
- **MNAR: Missing Not At Random**
  - Whether a value is missing depends on the value that would have been observed

Exercise 4

1. Make groups of three
2. For each of the three missingness types:
   Construct a story describing how a variable in the dataset have come to have missing values (criterion is not whether it is true, but how well it represents an instance of the missingness type)
3. Designate one person to carry each story
4. All with an MCAR-story gather toghether, all with an MAR and all with a MNAR
5. Each person tell their story in the group
6. Rank stories after how well they represent an instance of the missingness type

MI analysis - imputation

- Assume data is MAR
- Sex of the person was predictive for whether wood exposure was missing
- Sex also appears to be associated with wood dust exposure among those with observed values:

```
. tab sex wooddustgrp, row

           |            wooddust i 4 grupper
       sex |        0-      2.97-      3.75-      4.72- |     Total
-----------+--------------------------------------------+----------
    female |       190         52         38         19 |       299
           |     63.55      17.39      12.71       6.35 |    100.00
-----------+--------------------------------------------+----------
      male |       291        209        234        230 |       964
           |     30.19      21.68      24.27      23.86 |    100.00
-----------+--------------------------------------------+----------
     Total |       481        261        272        249 |     1,263

           |     38.08      20.67      21.54      19.71 |    100.00
```

Prediction of wood dust exposure according to sex

- Among women (I=0 - 2.96; II = 2.97 - 3.74; III = 3.75 - 4.71; IV = 4.72+)
  I: 64%
  II: 17%
  III: 13%
  IV: 6%
- Among men
  I: 30%
  II: 22%
  III: 24%
  IV: 24%
- I.e., if a woman lacks information on wood dust, we should give her the value 1 with a 64% chance, the value 2 with a 17% chance, and so on

Exercise 5: Imputation of wood dust exposure

- If a woman lacks information on wood dust, we should give her the value 1 with a 64% chance, the value 2 with a 17% chance, and so on
- Implement in Stata (first set the seed to your cpr-number):

```
. generate tmpwd = runiform() /* random number between 0 and 1 */
. generate wdimp = wooddustgrp
. replace wdimp = 0 if sex == 0 & tmpwd < .64 & missing(wooddustgrp)
. replace wdimp = 1 if sex == 0 & tmpwd > .64 & tmpwd < .81 &
missing(wooddustgrp)
. replace wdimp = 2 if sex == 0 & tmpwd > .81 & tmpwd < .94 &
missing(wooddustgrp)
. replace wdimp = 3 if sex == 0 & tmpwd > .94 & missing(wooddustgrp)
```

- Similarly for males
- Implement the above, conduct an analysis of the change in FEV1 with respect to wood dust exposure – report your estimates here:

https://docs.google.com/spreadsheets/d/1CwqP7mRs2-rrZfazrsNdL3tN9o5LdGcSnsgBIkJb410/edit?usp=sharing

MICE

- How can we deal with a variable having missing values, when the variable we predict from also have missing values?
- Consider wood dust and smoking (pack years):

```
. tab wooddustgrp packryg, missing

            | +- smokers incl. packyears, ex-smoker<2 ye
wooddust i  |              baseline smokers
 4 grupper  | nonsmoker  < 7 packy  >=7 packy         .a |     Total
------------+--------------------------------------------+----------
         0- |       237        122        101         21 |       481
      2.97- |       138         62         51         10 |       261
      3.75- |       146         74         44          8 |       272
      4.72- |       140         57         44          8 |       249
          . |        27          7          2         48 |        84
------------+--------------------------------------------+----------
      Total |       688        322        242         95 |     1,347
```

MICE

- An *iterative* procedure, i.e. we repeat the following an appropriate number of times

1. Estimate the association between the variable with fewest missings (V1) and the other explanatory variables
2. Impute from this model the missing values of V1
3. Estimate the association of the variable with 2nd fewest missings (V2) and the other explanatory variables including the imputed V1
4. Impute from this model the missing values of V2
5. Repeat steps 3 and 4 for V3, V4, …, VK
6. Repeat the above 20 times, say
7. Impute all variables from the K estimated models to create one complete dataset
8. Repeat all of the above *m* times, 100 say, to create *m* complete datasets

MICE in Stata

- Three steps

1. Declare data to be MI data:
```
mi set flong
```
2. Declare the variable to be imputed
```
mi register imputed V1 V2 V3
```
3. Do the prediction and imputation in a single command:
```
mi impute chained (regress) V1 ///
        (logit) V2 ///
        (mlogit) V3 = var1 var2, add(100)
```

- Yields 100 complete datasets, where the variables V1, V2, V3 no longer have missing values

Analysis of imputed data

Rubin's rule

- For each of the m imputed datasets we get the estimates $\hat{\theta}_j$ and $SE(\hat{\theta}_j)$
- As overall estimate we use the average estimate:

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$$

- As uncertainty estimate we use the combined SE:

$$SE(\hat{\theta}) = \sqrt{\overline{SE}^2(\hat{\theta}_j) + \frac{m+1}{m}\left(\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \hat{\theta})^2\right)}$$

- Note: Can be implemented in any spreadsheet
- Is automated in Statas –mi estimate–

AARHUS UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH

Exercise 6

We want to replicate the analysis reported by Jakobsen (2008) in Table 4, but using multiple imputation. Do this in pairs by completing the following steps:

1. Identify all relevant variables. Use –mi misspattern– to investigate the missingness pattern and amount
2. Choose a relevant regression model for each variable to be imputed (regress, logit, mlogit, etc...)
3. Declare the data to be of mi-type and register the variables to be imputed
4. Use –mi impute chained– to make a ”black box” imputation model (see slide 23) with 100 imputed datasets
5. Use –mi estimate– to obtain the final estimates of the analysis
6. Compare your estimates with those of Jakobsen (2008) and write a short conclusion

Thanks for your attention – questions welcome!

(Aarhus University, March 2016 – H Støvring)

AARHUS
UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH