## Applied Statistical Analysis with Missing Data
## Sensitivity analysis and other methods

Morten Frydenberg ©
Section of Biostatisics, Aarhus Univ, Denmark

Sensitivity analyses

"Robust" substantive models

Adjusting for missing observation by weights

Full information likelihood method.

1

## Sensitivity analysis

In general, a sensitivity analysis is an **alternative analysis** of the data, where you try to see how sensitive you results are to minor or large deviations from the assumptions behind the statistical analysis you have made.

There are at least **two types** of sensitivity analysis that you should make, when you have missing data (and use MI).

A: Try different plausible **imputation models**.
If these give very different results, then you probably have not really understood the structure of the data and the missing process correctly.

B: Try modelling the missing data in a **MNAR way** and se how this influences the complete data analysis.
A **unrealistic** model could be the "worst-case scenario", where all missing data are set to extremes: highest/lowest birth weight, mother non-smoker/ 20 cigs per day.

2

## Sensitivity analysis – type B

A worst-case sensitivity analysis is **seldom interesting** as it is not plausible.

Often authors try to make **"mental"** sensitivity analyses based on what they learned long ago for a **two by two** table:

> *If small children have a higher risk of not being recorded, then we have underestimated the effect of smoking during pregnancy.*

But often the setting is **too complicated** for the logic based on a two by two table to be correct.

With modern computer programs **you can do much better**.

You can **simulate** the missing data from any distribution you want, e.g. assuming that the missing birth weight are **on average** 150g lower than the observed birth weights.
And analysis this new data set to see how things change.

3

## Sensitivity analysis – type B

- mis. birth weight 150 g **higher/lower**, than observed (given sex, age parity....?)
- mis. birth weight had larger **variance** than observed (given sex, age parity....?)
- missing data is mainly among "unhealthy – strong", where there is **no relationship** between smoking and birth weight
- mis. income is 10% lower/higher, than what is found **based on the imputation model**
- mis health score is 20% higher (you do not turn up if **well**) than we see in the rest of the data.
- mis health score is 20% lower (you do not turn up if **not well**) than we see in the rest of the data.
- mis health scores is 20% higher among women on drugE.

4

## An old example!

Case-control study of squamous cell cancer of the oral cavity in Denmark.
Bundgaard, Wildt, Frydenberg, Elbrond, & Nielsen; *Cancer Causes and Control* **1995**,6 57-67

161 cancer cases – each matched to **four** controls.
Primary exposures were: number of teeth, alcohol and smoking habits – data collected by a mailed questionnaire.
All cases, but only 400 controls **(85%)** returned the questionnaire.

**Complete data analysis** showed that cancer was highly associated with alcohol, smoking and having few teeth.

Did we have MCAR? – **no!**
missing was definitely higher among alcoholics.

**The complete data analysis was biased!!**

5

## An old example!

Case-control study of squamous cell cancer of the oral cavity in Denmark.
Bundgaard, Wildt, Frydenberg, Elbrond, & Nielsen; *Cancer Causes and Control* **1995**,6 57-67

Did we have MAR? – **no!**
We do not believe that participation is independent of alcohol habits given just age and sex,
**which was all we knew about the non-responders.**

As we expect that the use of alcohol (and smoking) was higher among missing controls compared to those who participated – the study (probably) **overestimated** the association between cancer and alcohol – **but how much**?

We made **two sensitivity studies**
-A: The missing controls were sampled among **cases**
-B: The missing controls were sampled among **controls** but the use of alcohol and smoking were **doubled**.

6

## An old example!

**Table 8.** Results of two simulation studies evaluating the effect of nonrespondents

Comp case

| | | Actual estimates OR[c] | A: Simulation (i)[a] 1,000 simulations geometric average OR[c] | B: Simulation (ii)[b] 1,000 simulations geometric average OR[c] |
|---|---|---|---|---|
| Current alcohol | 1-5 drinks/day | 1.1 | 1.2 | 0.9 |
| | 6+ drinks/day | 9.7[d] | 4.5[d] | 7.2[d] |
| Current tobacco | 1-20 g/day | 2.0[d] | 1.8[d] | 1.2 |
| | 21+ g/day | 6.3[d] | 3.7[d] | 4.4[d] |
| Number of teeth | 5-14 | 1.5 | 1.4 | 1.3 |
| | 0-4 | 2.1[d] | 1.9[d] | 1.6[d] |

[a]Pseudo-controls taken as random cases.
[b]Pseudo-controls taken as random control, but with double the tobacco and alcohol consumption.
[c]OR = odds ratio (reference: 0 drinks/day, 0 g tobacco/day, 15-32 teeth).
[d]$P < 0.05$.

64 Cancer Causes and Control. Vol 6. 1995

Result of the sensitivity studies:
The association with alcohol and smoking were reduced, but still high.

7

## Robust Analysis/substantive models

Some analysis/substantive models are "robust" to some types of MAR.

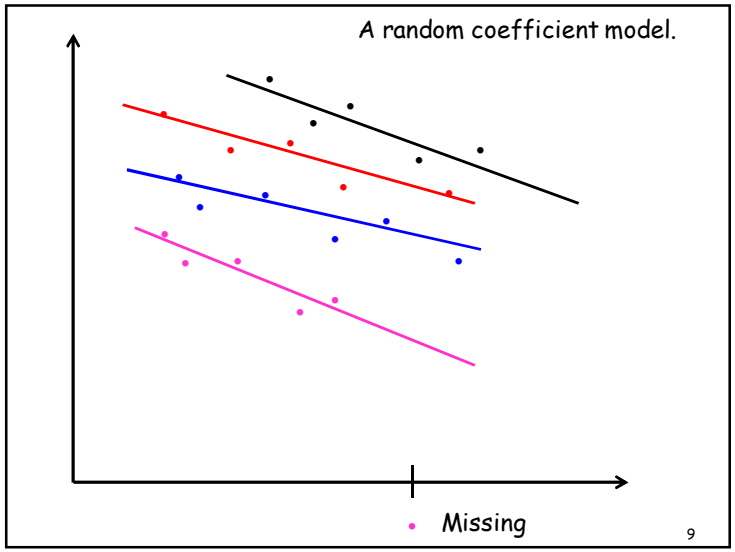
That is analysing the observed data is the most efficient way to analysis the data – using multiple imputation will not be more efficient!

Some, but not all, Mixed Models is robust for missing values response, **Y**.

One requirement is that the Ys is MAR given the variables included in the model and that the model is true.

Typical you need to consult a statistician in order to find out if have to worry about missing responses if you use a mixed model to analyse your data.

8




## Adjusting for missing by using sampling weights

An extreme type of missing data is, when you use **stratified sampling**.
E.g. you sample 100 random persons within each five year age and sex strata in the population.

Such a sample is clearly not **MCAR**.

But the data is **MAR**, as you know the age and sex of each of the persons you have not observed.

We could analyse the data with MI using models within each age and sex strata.

Another, valid and simple way is to analyse the data by **weighting** each observed data by **w=(strata size)/100**.

I.e. give a high weight to observation from large strata and a low weight for observations from small strata.

10

## Adjusting for missing by using sampling weights

Using weights **w=(strata size)/100**.

This correspond to
**w=1/"chance of being sampled"**
or in terms of observed/missing
**w=1/"chance of being observed"**

The method of using weights can be used in general.

If we have a data set with **only missing outcome** Y and MAR, then we can

1. Model the "chance of being observed" by a logistic regression analysis using the relevant covariates.
2. Use 1/(estimated probability) as weight in the analysis of the observed/complete data set.
   **Inverse Probability Weights**.

11

## "Full information likelihood methods"

Remember that you data consists of $Z_{obs}$ and $R$

Full modelling require that you make a completely specified statistical model for this data.

This is typically done in two ways:

- One specify a model for $R$, i.e. whether or not a variable is recorded (observed) and a model $Z_{obs}$ given $R$
- One specify a model for $Z$, i.e. whether or not a variable is recorded (observed) and a model $R$ given $Z$.

This can be complicated to do this and can involve numerical problems.
But it will make the assumptions transparent and if the models are true, the then estimates will be unbiased and efficient.

12