

SOS3003  
**Applied data analysis for  
social science**  
Lecture note 10-2010

Erling Berge  
Department of sociology and political  
science  
NTNU

Spring 2010

© Erling Berge 2010

1

## Literature

- Missing data  
Allison, Paul D 2002 "Missing Data", Sage  
University Paper: QASS 136, London, Sage,

Spring 2010

© Erling Berge 2010

2

## There is a missing case in the sample

- If one person
  - Refuses to answer
  - Are not at home
  - Has moved away
  - Etc.
- The problem of missing data belong to the study of biased samples.
- In general biased samples is a more severe problem than the fact that we are missing answers for a few variables on some cases (see Breen 1996 "Regression Models: Censored, Sample Selected, or Truncated Data", QASS Paper 111, London, Sage)
- But the problems are related

Spring 2010

© Erling Berge 2010

3

## There are missing answers for a few variables if

- Persons refuse to answer certain questions
- Persons forget, or do not notice some question, or the interviewer does it
- Persons do not know any answer to the question: "Do not know" are often a valid answer category. But the result is a missing answer
- The question is irrelevant (for the person)
- In administrative registers some documents may have been lost
- In research designs where variables with measurement problems may have been measured only for a minority of the sample

Spring 2010

© Erling Berge 2010

4

## Missing data entail problems

- There are practical problems due to the fact that all statistical procedures assume complete data matrices
- It is an analytical problem since missing data as a rule produce biased parameter estimates
- It is important to distinguish between data missing for random causes and those missing from systematic causes

Spring 2010

© Erling Berge 2010

5

## The simple solution: remove all cases with missing data

- Listwise/ casewise removal of missing data means to remove all cases with missing data on one or more variables included in the model
- The method has good properties, but may in some cases remove most of the cases in the sample
- Alternatives like pairwise removal or replacement with average variable value has proved not to have good properties
- More recently developed methods like "maximum likelihood" and "multiple imputation" have better properties but are more demanding
- In general it pays to do good work in the data collection stage

Spring 2010

© Erling Berge 2010

6

## Types of randomly missing

- **MCAR: missing completely at random**
  - Means that missing data for one person on the variable  $y$  is uncorrelated with the value on  $y$  and with the value on any other variable in the data set (however, internal case by case the value of missing may of course correlate with the value missing on other variables)
- **MAR: missing at random**
  - Means that missing data for person  $i$  on the variable  $y$  do not correlate with the value on  $y$  if one control for the variation of other variables in the model
  - More formally:  
$$\Pr(Y_i = \text{missing} \mid Y_i, X_i) = \Pr(Y_i = \text{missing} \mid X_i)$$

Spring 2010

© Erling Berge 2010

7

## Process resulting in missing

- **Is ignorable if**
  - The result is MAR and the parameters governing the process are unrelated to the parameters that are to be estimated
- **Is non-ignorable if**
  - The result is not MAR. Estimation of the model will then require a separate model of the missing process
  - See Breen 1996 "Regression Models: Censored, Sample Selected, or Truncated Data", QASS Paper 111, London, Sage
- **Here the situation with MAR will be discussed**

Spring 2010

© Erling Berge 2010

8

## Conventional methods

Common methods in cases with MAR data:

- Listwise deletion
- Pairwise deletion
- Dummy variable correction
- Imputation (guessing a value for the missing)

Of the conventional methods listwise deletion is the best

## Listwise deletion (1)

- Can always be used
- If data are MCAR we have a simple random subsample of the original sample
- Smaller  $n$  entails larger variance estimates
- In the case of MAR data and the missing values on an  $x$ -variable are independent of the value on  $y$ , listwise deletion will produce unbiased estimates

## Listwise deletion (2)

- In logistic regression listwise deletion may cause problems only if missing is related both to dependent and independent variables
- If missing depends only on the values of the independent variable listwise deletion is better than replacement of missing values by maximum likelihood and multiple imputation

Spring 2010

© Erling Berge 2010

11

## Pairwise deletion

- Means that all computations are based on all available information seen pairwise for all pairs of variables included in the analysis
- One consequence is that different parameters will be estimated on different samples (we see variation in n from statistic to statistic)
- Then all variance estimates are biased
- Common test statistics provides biased estimates (e.g. t-values and F-values)
- **DO NOT USE PAIRWISE DELETION !!**

Spring 2010

© Erling Berge 2010

12

## Dummy variable correction

If data is missing for the independent variable  $x$

- Let  $x^*_i = x_i$  if  $x_i$  is not missing and  
 $x^*_i = c$  (an arbitrary constant) if  $x_i$  is missing
- Define  $D_i=1$  if  $x_i$  is missing, 0 otherwise
- Use  $x^*_i$  and  $D_i$  in the regression instead of  $x_i$
- In nominal scale variables missing can get its own dummy

Investigations reveal that even if we have MCAR data parameter estimates will be biased

**Do not use dummy variable correction!**

## Imputation

- The goal is to replace missing values with reasonable guesses about what the value might have been before one does an analysis as if this were real values; e.g.
  - Average of valid values
  - Regression estimates based on many variables and cases with valid observations
- Parameter estimates are consistent, but estimates of variances are biased (consistently too small), and the test statistics are too big
- Avoid if possible the simple forms of imputation

## Concluding on conventional methods for missing data

- Conventional methods of correcting for missing data make problems of inference worse
- Be careful in the data collection so that the missing data are as few as possible
- Make an effort to collect data that may help in modelling the process resulting in missing
- If data are missing use listwise deletion if not maximum likelihood or multiple imputation is available

Spring 2010

© Erling Berge 2010

15

## New methods for ignorable missing data (MAR data): Maximum Likelihood (ML)

- Conclusions
  - Based on the probability for observing just those values found in the sample
  - ML provides optimal parameter estimates in large samples in the case of MAR data
  - But ML require a model for the joint distribution of all variables in the sample that are missing data, and it is difficult to use for many types of models

Spring 2010

© Erling Berge 2010

16



## ML-method: example (1)

- Observing  $y$  and  $x$  for 200 cases
- 150 distributed as shown
- For 19 cases with  $Y=1$   $x$  is missing and for 31 cases with  $Y=2$   $x$  is missing
- We want to find the probabilities  $p_{ij}$  in the population

	Y=1	Y=2
X=1	52	21
X=2	34	43

	Y=1	Y=2
X=1	$p_{11}$	$p_{12}$
X=2	$p_{21}$	$p_{22}$

Spring 2010

© Erling Berge 2010

17

## ML-method: example (2)

- In a table with  $I$  rows and  $J$  columns, complete information on all cases and with  $n_{ij}$  cases in cell  $ij$  the Likelihood is

$$\mathcal{L} = \prod_{i, j} \left( p_{ij} \right)^{n_{ij}}$$

That is the product of all probabilities for every table cell taken to the power of the cell frequency

Spring 2010

© Erling Berge 2010

18

## ML-method: example (3)

For a fourfold table the Likelihood will be

$$\mathcal{L} = (p_{11})^{n_{11}} (p_{12})^{n_{12}} (p_{21})^{n_{21}} (p_{22})^{n_{22}}$$

For the 150 cases in the table above where we have all observations the Likelihood will be

$$\mathcal{L} = (p_{11})^{52} (p_{12})^{21} (p_{21})^{34} (p_{22})^{43}$$

## ML-method: example (4)

- For tables the ML estimator is  $p_{ij} = n_{ij}/n$
- This provides good estimates for the table where we do not have missing data (listwise deletion)
- How can one use the information about  $y$  for the 50 cases where  $x$  is missing?
- Since MAR is assumed to be the case, the 50 extra cases with known  $y$  should follow the marginal distribution of  $y$
- $\Pr(Y=1) = (p_{11} + p_{21})$  and  $\Pr(Y=2) = (p_{12} + p_{22})$

## ML-method: example (5)

- Taking into account all that is known about the 200 cases the Likelihood becomes

$$\mathcal{L} = (p_{11})^{52} (p_{12})^{21} (p_{21})^{34} (p_{22})^{43} (p_{11} + p_{21})^{19} (p_{11} + p_{21})^{31}$$

- The ML-estimators will now be

$$\hat{p}_{ij} = \hat{p}(x = i | y = j) \hat{p}(y = j)$$

## ML-method: example (6)

- Taking into account the information we have about cases with missing data, parameter estimates change

Estimate of	Missing deleted	Missing included
$p_{11}$	0.346	0.317
$p_{21}$	0.227	0.208
$p_{12}$	0.140	0.156
$p_{22}$	0.287	0.319

## The ML-method in practice

- For the general case with missing data there are two approaches
  - The expectation-maximization (EM) method, a two stage method where one starts out with the expected value of the missing data and use these to obtain parameter estimates that again will be used to provide better estimates of the missing values and so on ...  
(this method provides biased estimates of standard errors)
  - Direct ML estimates are better but can be provided only for linear and log-linear models

Spring 2010

© Erling Berge 2010

23

## New methods for ignorable missing data (MAR data): Multiple Imputation (MI)

- Conclusions
  - MI is based on a random component added to estimates of the missing data values
  - Has as good properties as the ML method and is easier to implement for all kinds of models
  - But it gives different results every time it is used

Spring 2010

© Erling Berge 2010

24

## Multiple Imputation (1)

- MI have the same optimal properties as the ML method. It can be used on all kinds of data and with all kind of models. In principle it can be done with the ordinary analytical tools
- The use of MI can be rather convoluted. This makes it rather easy to commit errors. And even if it is done correctly one will never have the same result twice due to the random component in the imputed variable value

## Multiple Imputation (2)

- Use of data from a simple imputation (with or without a random component) will underestimate the variance of parameters. Conventional techniques are unable to adjust for the fact that data have been generated by imputation
- The best way of doing imputation with a random component is to repeat the process many times and use the observed variation of parameter estimates to adjust the estimates of the parameter variances
- Allison, p.30-32, explains how this can be done

## Multiple Imputation (3)

- MI requires a model that can be used to predict values of missing data. Usually there is an assumption of normally distributed variables and linear relationships. But models can be tailored to each problem
- MI can not handle interactions
- MI model should contain all variables of the analysis model
- (including the dependent variable)
- MI works only for interval scale variables. If nominal scale variables are used special programs are needed
- Testing of several coefficients in one test is complicated

Spring 2010

© Erling Berge 2010

27

## When data are missing systematically

- Will usually require a model of how the missing cases came about
- ML and MI approaches can still be used, but with much stronger restrictions and the results are very sensitive for deviations from the assumptions

Spring 2010

© Erling Berge 2010

28

## Summary

- If listwise deletion leaves enough data this is the simplest solution
- If listwise deletion do not work one should test out multiple imputation
- If there is a suspicion that data are not MAR one needs to create a model of the process creating missing. This can then be used together with ML or MI. Good results require that the model for missing is correct

Spring 2010

© Erling Berge 2010

29

## Important types of biased samples

### References:

- Breen, Richard. 1996. Regression Models: Censored, Sample Selected, or Truncated Data. QASS Paper 111, London: Sage
- Winship, Christopher, and Robert D. Mare. 1992. Models for sample selection bias. Annual Review of Sociology, 18:327-350

Spring 2010

© Erling Berge 2010

30

## Types of biased samples

- Censored
- Truncated
- Selected
- Such samples arise because society works “selectively”, and because we do not get complete answers to questions asked
- Which variables and how they are truncated determine the type of bias

Spring 2010

© Erling Berge 2010

31

## Causal analysis in biased samples

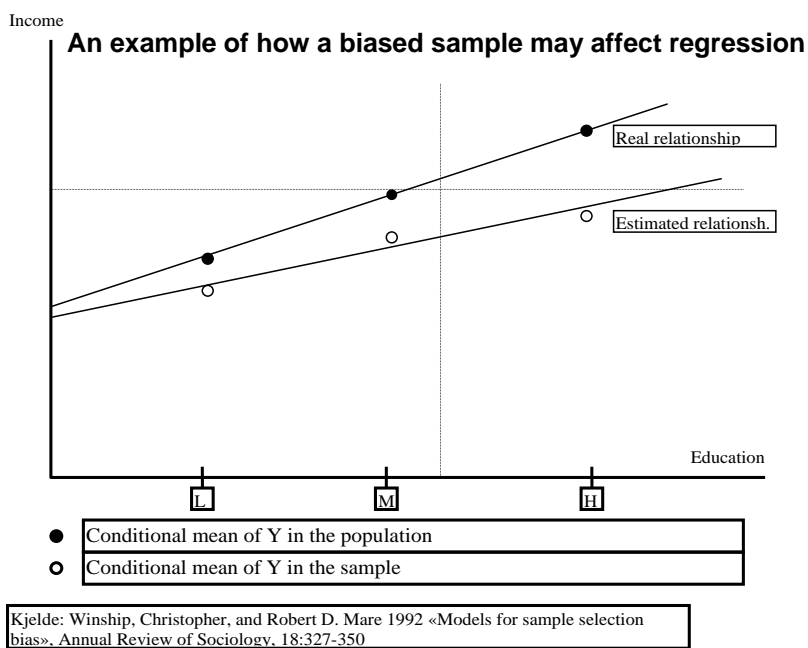
- Regression analysis
  - Will (as a rule) have severe problems if the sample is biased
- Hence
  - The process of selection needs to be included in the model or analysis

Spring 2010

© Erling Berge 2010

32





Spring 2010

© Erling Berge 2010

33

## Comments to the figure

- Only persons with incomes below 15000 USD are included in the sample
- Result is erroneous estimate of the (real) impact of education
- Errors in reporting income creates a selected sample
- Large errors in the original sample leads to exclusion
- Large values on the independent variable leads to large (negative) errors
- The errors in the sample will be correlated with x (violating the OLS assumptions)

Spring 2010

© Erling Berge 2010

34

## Truncation of variables

- A variable,  $X$ , is called truncated if we for  $X < c$  or for  $X > c$  do not know more than that  $X < c$  or  $X > c$
- This is known as left or right side truncation respectively
- We may have multiple truncation such as simultaneous left side and right side truncation

Spring 2010

© Erling Berge 2010

35

## Biased samples and missing data I

- Censored samples (explicit selection on  $Y$ )
  - $Y$  is unknown for cases where  $Y$  has value above or below  $c$
  - $X$  is known for all cases in the sample
- Selected samples (unsystematic selection)
  - $Y$  is unknown for cases where f. e.  $z=1$  and known if  $z=0$
  - $X$  is known for all cases in the sample

Spring 2010

© Erling Berge 2010

36

## Selected or censored sample?

- The terminology is not very clear
- In general the distinction is a question of interpretation and theoretical meaning
  - If the missing observations on  $Y$  are caused by the measurement method or data collection method the sample is called censored
  - If the missing observations of  $Y$  are caused by the behaviour of the individuals the sample is called selected

Spring 2010

© Erling Berge 2010

37

## Biased samples and missing data II

- Truncated samples (explicit selection on  $Y$ )
  - $Y$  is unknown for cases where  $Y$  has value above or below  $c$
  - $X$  is known when  $Y$  is known
- Selection on the independent variable
  - $Y$  is known for cases where  $X$  has a value above or below  $c$
  - $X$  is known when  $Y$  is known

Spring 2010

© Erling Berge 2010

38

## Consequences of biased samples

- **Selection on the independent variable do not cause problems**
- Truncated, selected, and censored samples cause the residual to be correlated with the independent variables. Both external and internal validity is compromised

## Causes of biased samples

- Data collection procedures and missing answers may lead to truncated, selected or censored samples
  - For example: "missing" on a dependent variable give a selected sample based on the variable Z, answer or no answer
- In every non-random sample there is a potential for erroneous conclusions due to biased sample

## How to handle biased samples

- The analysis should at the outset acknowledge the problem and use models that are able to correct for bias in the sample unless there are good reasons to believe the problem is small
- The solution then is to
  - 1) construct a model that predicts selection
  - 2) use this model to construct a model that predicts  $y$  conditional on the person having been selected

Spring 2010

© Erling Berge 2010

41

## A basic model for censored samples

$$E[Y | X] = \Pr[Y > c | X] * E[Y | Y > c \& X] + \Pr[Y \leq c | X] * E[Y | Y \leq c \& X]$$

Left side truncation of  $Y$  at  $c$  gives

$$E[Y | Y \leq c \& X] = c$$

It is always possible to transform  $Y$  so that  $c=0$ , hence the real regression,  $E[Y | X]$ , can be written

- **$E[Y | X] = \Pr[Y > 0 | X] * E[Y | Y > 0 \& X]$**

Spring 2010

© Erling Berge 2010

42

## The model in a truncated sample

- $Y_i = E[Y_i | Y_i < a \ \& \ X'_i] + e_i$

It can be shown that this is equivalent to

- $Y_i = E[Y_i | X_i] - \sigma \lambda'_i(m) + e_i$   
where  $\lambda'_i(m)$  is an estimate of the Hazard rate at point  $m = (a - E[Y_i | X_i]) / \sigma$
- The hazard rate (from event history analysis) is # events ( $Y=1$ ) per time interval as the time interval goes towards 0

The parameters of  $E[Y_i | X_i]$  are overestimated

The model can be estimated by the ML method

## Two step model on censored samples

- The selection model,  $\Pr[Y > c | X]$ , can be modelled by probit regression on the censored sample
- The model of the outcome,  $E[Y | Y > c \ \& \ X]$ , can then be estimated on the censored sample
- The results are trustworthy only in large samples

## Problems in the two step model

- Results are sensitive for assumptions about the distribution of the residual
  - Homoscedasticity: deviation for this assumption is more serious than in OLS since estimates in a censored model are neither consistent, nor efficient
  - Normal distributionBoth assumptions have to be properly tested
- There are also problems of identification of parameters due to multicollinearity between the hazard rate and the explanatory variables (see Breen 1996:16 (equation 2.7))

Spring 2010

© Erling Berge 2010

45

## Two step model in OLS

is sensitive for

- Correlations between errors in the selection equation ( $u$ ) and errors in the outcome equation ( $e$ )
- Correlations between variables in the selection and outcome equations
- Degree of censoring in the sample (how large a fraction of the cases have missing  $y$  values?)

Conclusion: use ML-estimation

Spring 2010

© Erling Berge 2010

46