

Workshop on methods for studying cancer patient survival with application in Stata

Karolinska Institute, 6th September 2007



Modeling relative survival in the presence of incomplete data

Ula Nur

Cancer Research UK Cancer Survival Group
London School of Hygiene and Tropical Medicine

Outline

- ❖ Types of missing data
- ❖ Single imputation
- ❖ Multiple imputation
- ❖ Examples
- ❖ Multiple Imputation By Chained Equations
- ❖ Conclusion

Missing data is a problem

- ❖ Loss of information, efficiency or power due to loss of data
- ❖ Problems in data handling, computation and analysis due to irregularities in the data patterns and non-applicability of standard software
- ❖ Serious bias if there are systematic differences between the observed and the unobserved data.

Types of missing data

Missing completely at random (MCAR)

When there are no systematic differences between complete and incomplete records.

For example, a cancer registry has incomplete records with the variable **stage** missing, because two hospitals failed to collect this information.

Missing at random (MAR)

Incomplete data differ from cases with complete data, but the pattern of data missingness is traceable from other observed variables in the dataset.

❖ A typical example of MAR, was presented by (Van Buuren, 1999), in a study on imputation of missing blood pressure in survival analysis.

❖ He found that probability of missing BP depends on survival.

❖ Short term survivors have more missing BP data.

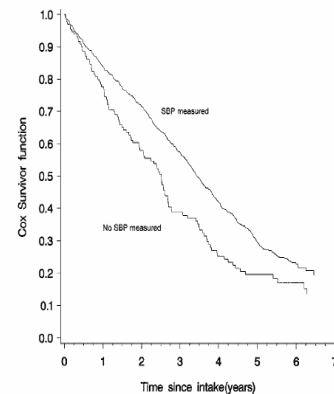


Figure 1. Survival curves obtained by fitting a proportional hazards model adjusted for age, sex and type of residence and stratified by the presence of a systolic blood pressure measurement ($n_{SBP} = 835$, $n_{noSBP} = 121$).

Missing not at random (MNAR)

When the pattern of data missingness is non-random, and is not predictable from other variables in the data set.

An example of MNAR could be that the variable *stage* was unobserved (not recorded) for patients with late cancer stage, i.e. probability of missingness is related to the incomplete variable *stage*.

What to do about Missing Data

- ❖ Analyse complete records (complete case analysis)
- ❖ Impute a number
 - run complete case analysis
- ❖ Multiple imputation
 - Generate m imputations
 - run complete case analysis
 - combine estimates and standard errors

Complete case analysis

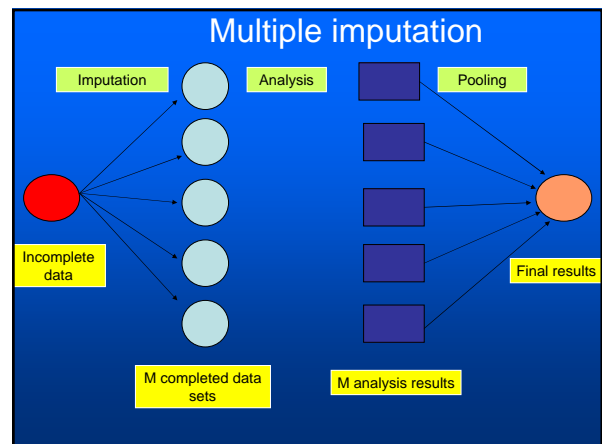
- ❖ The easiest solution of handling missing data is to exclude all records (cases) that are incomplete.
- ❖ This method can be a reasonable solution when the incomplete cases comprise only a small fraction (5% or less) of all cases.
- ❖ The main advantage of this method is simplicity
- ❖ Doesn't compensate for the sampling bias due to the loss of data
- ❖ Loss of substantial amount of data – thus loss of statistical power

Single imputation

- ❖ Mean substitution
- ❖ Indicator method
- ❖ Regression methods
- ❖ Hotdeck imputation

Multiple Imputation

- ❖ A simulation based approach to the analysis of incomplete data
- ❖ Assumes the mechanism of missingness to be at least MAR
- ❖ Replace each missing observation with $m > 1$ simulated values
- ❖ Analyze each of the m datasets in an identical fashion
- ❖ Combine results (Rubin's Rules)



Imputation model

- ❖ The most difficult part of this method is how to generate the values to be imputed.
- ❖ A very simple model, might not reflect the data well.
- ❖ A too complex model, can be extremely difficult to implement and program.

How many imputations are needed ?

Rubin (1987, p. 114) shows that the efficiency of an estimate based on m imputations is approximately

$$\left(1 + \frac{\lambda}{m}\right)^{-1}$$

Where λ is the fraction of missing information

m	λ				
	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Pooling of results

Let us assume that following the analysis of m completed datasets, there are now m estimates $\hat{P}_j, j = 1, \dots, m$ with sampling variance $\hat{s}_j^2, j = 1, \dots, m$

- The mean of \hat{P}_j is then given by $\bar{P} = \frac{\sum_{j=1}^m \hat{P}_j}{m}$

- The variability of \hat{P}_j is divided into two components

$$\text{Within imputation variance } \bar{U} = \frac{1}{m} \sum_{j=1}^m s_j^2$$

$$\text{Between imputation variance } B = \frac{1}{m-1} \sum_{j=1}^m (\hat{P}_j - \bar{P})^2$$

$$\text{Total variance } T = \bar{U} + (1 + m^{-1})B$$

What if there are more than one incomplete variables in the dataset?

- ❖ Condition on one variable
- ❖ Multiple imputation by chained equations

Multiple imputation by chained equations (Van Buuren, 1999).

- ❖ Assume a multivariate distribution exist without specifying its form.
- ❖ Assume Missing at random (MAR).
- ❖ Imputation model include all variables in the analysis model.

- ❖ Fill in each missing value with a starting value (mode for categorical variables, mean for continuous variables).
- ❖ Discard the filled-in values from the first variable.
- ❖ Regress x_1 on x_2, \dots, x_n .
- ❖ Replace missing values on x_1 .
- ❖ Repeat for x_2, \dots, x_n on the other x 's (1 iteration).
- ❖ The same procedure is repeated for several (in this case 10) iterations. This generates one completed dataset.
- ❖ For m completed datasets, repeat the procedure m times independently.

Imputation models

- ❖ Logistic regression for binary variables.
- ❖ Linear regression for continuous variables.
- ❖ polytomous logistic regression for categorical variables.

- ❖ The chain of the Gibbs sampling should be iterated until it reaches convergence.
- ❖ There is no need for burn-in stage (the initial iterations of the Markov Chain that are discarded because they are usually influenced by the starting distribution).
- ❖ There is no definite method to assess that the algorithm has converged.
- ❖ The main aim would be to choose sufficient number of iterations to stabilize the distribution of the parameters.
- ❖ Usually 10 iterations are sufficient.

Software

- ❖ ICE: STATA implementation of multiple imputation by Chained equations (By Patrick Royston)
- ❖ MICE: S-PLUS software for flexible generation of multivariate imputations (By Stef van Buuren).
- ❖ IVEWARE: SAS-based application for creating multiple imputations (By Raghunathan, Solenberger and John Van Hoewyk).
- ❖ SOLAS: a commercial software for multiple imputation by Statistical Solutions Limited.

Multiple imputation in STATA

Finnish colon cancer data

- localised colon carcinoma (15,564 patients).
- The data sets contain all cases diagnosed in Finland (population 5.1 million)
- during 1975–94 with follow-up to the end of 1995.
- Information on sex, age, stage, subsite, year of diagnosis, survival time, vital status.

variable name	storage type	display format	value label	variable label
sex	byte	%8.0g		Sex
age	byte	%8.0g		Age at diagnosis
stage	byte	%8.0g		Clinical stage at diagnosis
subsite	byte	%8.0g		Anatomical subsite of tumour
year8594	byte	%8.0g		Year of diagnosis 1985-94
agegrp	byte	%8.0g		Age in 4 categories
surv_mm	float	%8.0g		Survival time in months
surv_yy	float	%8.0g		Survival time in years
status	byte	%8.0g		Vital status at last contact
id	float	%8.0g		Unique ID

- We forced missing values on the two variables *stage* and *subsite*.
- For *stage* missing values were generated depending on age at diagnosis
- For *subsite* missing values were generated depending on year of diagnosis.
- Use *mvpatterns* command to see the pattern of missing data

```

. gen r=runiform()
. gen stage_miss=stage
. replace stage_miss=. if agegrp==0 & r>=.9
(10 real changes made, 70 to missing)
. replace stage_miss=. if agegrp==1 & r>=.8
(485 real changes made, 485 to missing)
. replace stage_miss=. if agegrp==2 & r>=.7
(1968 real changes made, 1968 to missing)
. replace stage_miss=. if agegrp==3 & r>=.4
(3494 real changes made, 3494 to missing)
. gen subsite_miss=subsite
. replace subsite_miss=. if year8594==0 & r>=.6
(2605 real changes made, 2605 to missing)
. replace subsite_miss=. if year8594==1 & r>=.8
(1821 real changes made, 1821 to missing)
. end
. save "colon_miss.dta"

```

- ### Multiple imputation in stata using *ICE*
- The *cmd* option specifies the type of prediction model
 - Our incomplete variable *stage* has 4 ordered categories.
 - By default *ice* will treat this variable as unordered, therefore *mlogit* (is used in the prediction model).
 - We can change this using *cmd* to use ordinal logistic regression instead

```

. ice sex age stage_miss surv_yy status subsite_miss year8594, dryrun

```

Missing values	Freq.	Percent	Cum.
0	8,884	57.08	57.08
1	2,919	18.75	75.84
2	3,761	24.16	100.00
Total	15,564	100.00	

Variable	Command	Prediction equation
sex		[No missing data in estimation sample]
age		[No missing data in estimation sample]
stage_miss	mlogit	sex age surv_yy status subsite_miss year8594
surv_yy		[No missing data in estimation sample]
status		[No missing data in estimation sample]
subsite_miss	mlogit	sex age stage_miss surv_yy status year8594
year8594		[No missing data in estimation sample]

```

. ice sex age stage_miss surv_yy status subsite_miss year8594, cmd(stage_miss:ologit) dryrun

```

Missing values	Freq.	Percent	Cum.
0	8,884	57.08	57.08
1	2,919	18.75	75.84
2	3,761	24.16	100.00
Total	15,564	100.00	

Variable	Command	Prediction equation
sex		[No missing data in estimation sample]
age		[No missing data in estimation sample]
stage_miss	ologit	sex age surv_yy status subsite_miss year8594
surv_yy		[No missing data in estimation sample]
status		[No missing data in estimation sample]
subsite_miss	mlogit	sex age stage_miss surv_yy status year8594
year8594		[No missing data in estimation sample]

- We choose m=10 to create ten completed datasets
 - Each completed dataset was generated using 10 iterations
- ```

ice sex age surv_yy status subsite_miss stage_miss
year8594 using "colon_imp", cmd(stage_miss:ologit)m(10)

```

- All completed data sets are saved in one file specified by the variable `_j`
- This was then split into 10 separate completed datasets
- Ran `strs` on each completed dataset to estimate relative survival using actuarial methods.

```
strs using popmort, br(0(1)10) mergeby(_year sex _age)
by(sex year8594 agegrp stage_miss subsite_miss)
save(replace)
```

- `Strs` creates two output data files:
  1. `individ.dta` contains one observation for each patient and each life table interval
  2. `grouped.dta` contains one observation for each life table interval
- Ten versions of each was created.
- This was then appended to one large completed `individ.dta` and one large completed `grouped.dta` dataset

```
sex
age
stage
mmdx
yydx
surv_mm
surv_yy
status
subsite
year8594
agegrp
dx
exit
id
stage_miss
subsite_miss

use grouped, clear
gen _j=1
local i=1
while '1' < 11 {
 append using grouped`i'
 replace _j=`i' if _j==.
 local i=`i'+1
}
save grouped.dta

use individ, clear
gen _j=1
local i=1
while '1' < 11 {
 append using individ`i'
 replace _j=`i' if _j==.
 local i=`i'+1
}
save individ.dta
```

```
xt: mcombine gln d end sex age year8594 i.agegrp i.stage_miss i.subsite_miss, eform family(pois)
> tar) lnoffset(y)
year8594 _i.agegrp_0-3 (naturally coded; _i.agegrp_0 omitted)
stage_miss _i.stage_miss_0-3 (naturally coded; _i.stage_miss_0 omitted)
subsite_miss _i.subsite_miss_1-4 (naturally coded; _i.subsite_miss_1 omitted)

Multiple imputation parameter estimates (10 imputations)

 d exp(b) std. err. z P>|z| [95% Conf. Interval]
-----+-----
d end .6391558 .0081487 -35.11 0.000 .6231834 .6553273
 sex .9998722 .0281301 -0.01 0.991 .9460311 1.053535
 age 1.023571 .0032687 7.30 0.000 1.017171 1.029996
 year8594 .8154708 .023626 -6.36 0.000 .7692425 .8610841
 i.agegrp_1 .7508854 .0612056 -3.52 0.000 .6398187 .8807629
 i.agegrp_2 .6064815 .0668924 -4.53 0.000 .4881286 .7510814
 i.agegrp_3 .604182 .0815147 -3.98 0.000 .4714082 .7478351
 i.stage_miss_1 .7024537 .042257 -5.87 0.000 .6243277 .7803561
 i.stage_miss_2 1.453333 .0951464 5.71 0.000 1.273118 1.652309
 i.stage_miss_3 1.377226 .1702189 24.14 0.000 1.059417 1.724828
 i.subsite_1 1.118924 .0487373 2.58 0.010 1.027365 1.218644
 i.subsite_2 .9945772 .0314883 -6.16 0.000 .9316623 1.064247
 i.subsite_3 1.064612 .0694746 0.96 0.337 .9357928 1.209872
 i.subsite_4 (exposure)
 _j
```

- We can now fit a Poisson regression on `individ.dta` using `mcombine`

- `Individ.dta` is large (595,520 records)
- `Grouped` data could not be used, as each completed data set had a different size

|        |       |        |       |
|--------|-------|--------|-------|
| group1 | 2,438 | group5 | 2,404 |
| group2 | 2,394 | group6 | 2,419 |
| group3 | 2,445 | group7 | 2,427 |
| group4 | 2,381 | group8 | 2,382 |

The STATA command `mim` can also combine multiply imputed data set. Results below are very similar to `mcombine`

```
year8594
agegrp
stage_miss
subsite_miss

xt: mim: gln d end sex age year8594 i.agegrp i.stage_miss i.subsite_miss, eform family(pois) 10
> lnoffset(y)
year8594 _i.agegrp_0-3 (naturally coded; _i.agegrp_0 omitted)
stage_miss _i.stage_miss_0-3 (naturally coded; _i.stage_miss_0 omitted)
subsite_miss _i.subsite_miss_1-4 (naturally coded; _i.subsite_miss_1 omitted)

Multiple imputation estimates (gln)
Generalized Linear Model

 d exp(b) std. err. z P>|z| [95% Conf. Int.] mi.off
-----+-----
d end .639156 .008148 -35.11 0.000 .623184 .655364 137.9
 sex .999872 .028130 -0.01 0.991 .945002 1.0566 285.0
 age 1.023571 .003269 7.30 0.000 1.017171 1.030027 338.0
 year8594 .815471 .023626 -6.36 0.000 .769091 .883437 240.3
 i.agegrp_1 .750885 .061206 -3.52 0.000 .639959 .88112 478.1
 i.agegrp_2 .606482 .066892 -4.53 0.000 .488161 .75348 495.9
 i.agegrp_3 .604182 .081514 -3.98 0.000 .472067 .745888 422.1
 i.stage_miss_1 .702454 .042257 -5.87 0.000 .623183 .781833 77.6
 i.stage_miss_2 1.453333 .095146 5.71 0.000 1.271831 1.65554 80.4
 i.stage_miss_3 1.377223 .170219 24.14 0.000 1.058777 1.73106 139.2
 i.subsite_1 1.118924 .048737 2.58 0.012 1.025777 1.20594 19.8
 i.subsite_2 .994577 .031488 -6.16 0.000 .930433 1.08311 118.1
 i.subsite_3 1.064612 .069475 0.96 0.340 .935203 1.21193 92.3
 i.subsite_4 (exposure)
 _j
```

We can now compare the previous MI results to **glm** results on the complete observed dataset

```

log Likelihood = -29031.13648 AIC = 9754546
 BIC = -617436.5

d exp(b) std. err. z Pr>|z| [95% conf. interval]
-----+-----+-----+-----+-----+-----
_cons .6686527 .0078755 -84.37 0.000 .6533938 .6842684
sex .9071460 .0212565 -42.71 0.000 .8648927 .9514394
age 1.0218199 .0035278 28.97 0.000 1.0147542 1.0288857
year3581 .8689074 .0217681 -39.91 0.000 .8252732 .9126369
_1ageyr_1 -.7811657 .0628321 -12.42 0.002 .8846958 -.6776964
_1ageyr_2 .6526808 .0684443 9.54 0.000 .5184302 .8018205
_1ageyr_3 .6721718 .0891566 7.55 0.000 .4986211 .8513211
_1stage_1 .3202007 .0360793 8.88 0.000 .2470002 .3934012
_1stage_2 .8966819 .0464345 19.31 0.000 .8011183 .9924508
_1stage_3 3.2593399 .1165321 28.02 0.000 2.888976 3.649712
_1subsite_1 1.1182079 .0389331 28.71 0.001 1.044318 1.197051
_1subsite_2 .9030066 .0274991 -32.82 0.000 .8066486 .9993849
_1subsite_3 1.0193396 .0544111 18.73 0.000 .9179813 1.126658
y (exposure)
end of do-file

```

## Conclusions & discussion

- ❖ Data are valuable - should not be wasted.
- ❖ Doing nothing, i.e. 'complete case analysis' should not be an option.
- ❖ Missing at random (MAR) crucial to avoid bias, but often un-testable.
- ❖ Great care should be taken in the choice of imputation models.
  1. How many predictors can we include?
  2. Vital status
  3. Survival time

## References

- ❖ Royston P. Multiple imputation of missing values: update of ice. *Stata Journal* 2005; 5: 527—36.
- ❖ Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003;56(1):28-37.
- ❖ Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999;18(6):681-94.
- ❖ Schafer JL. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.
- ❖ Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.
- ❖ Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Second Edition ed. New York: John Wiley & Sons, 1987.