

Stata code for applying the multiple imputation (MI) method to a colorectal cancer data set to handle the incompleteness of tumour information, applied in (Nur, et al, 2009).

This Stata code imputes missing values using switching regression, which is an iterative multivariable regression technique. This is for Stata 9 and later versions, originally publicised in (Royston, 2005a), implementing the MICE (multivariate imputation by chained equations) (Van Buuren et al., 1999) method of multiple multivariate data imputation. Two updates of ice have followed (Royston, 2005b; Royston, 2007). A further update that focus on categorical variables was published two years later. This ice version require Stata 10.1 (Royston, 2009) (Strictly speaking, only the newly implemented negative binomial regression option requires Stata 10.1; all other features work under Stata 9.2 or higher).

The missing observations are assumed to be “missing at random” (MAR) or “missing completely at random” (MCAR) see (Little and Rubin, 1987) for a definition of these terms.

- Step 1. Description of the variables.**
- Step 2. Imputation of incomplete variables.**
- Step 3. Declare data to be survival-time data, and estimate relative survival.**
- Step 4. Estimate the excess hazard of death of the colorectal cancer patients using multivariable regression using a generalised linear model with Poisson error (Dickman et al, 2004), in each of the imputed datasets and combine results using Rubin Rules.**

Step 1: Description of the variables

Variable name	Type	Description	% missing
<i>deprivation</i>	categorical	1: Least deprived, 2, 3, 4, 5: Most deprived	0.0 %
<i>stage</i>	categorical	stage of tumour at diagnosis (1: Stage I, 2:Stage II, 3:Stage III, 4: Stage IV)	39.5 %
<i>treatment</i>	categorical	treatment (1:Surgery only, 2: Surgery and chemo, 3:Surgery and radio, 4:Surgery, chemo, radio, 5: Chemo only, 6: Chemo and radio, 7: Radio only, 8: No treatment)	0.0 %

<i>site</i>	categorical	Site of tumour (1:Colon, 2=Rectosigmoid, 3: Rectum)	0.0 %
<i>charlson</i>	categorical	Charlson index derived for comorbidity conditions occurring in the period 6-18 months before the diagnosis of colorectal cancer (0: none, 1, 2, 3+)	0.0%
<i>hist</i>	categorical	Morphology (1: Adenocarcinoma, 2: Mucinous and serous, 3: Other, 4: Neoplasm Not Otherwise Specified (NOS))	11.6%
<i>grade</i>	categorical	Tumour grade (1:I, 2:II, 3:III/V)	25.0%
<i>sex</i>	binary	Sex (1=male, 2=female)	0.0%
<i>age</i>	continuous	Age at diagnosis	0.0 %
<i>agegrp</i>	categorical	Age group at diagnosis (1:15-44, 2:45-54, 3: 55-64, 4:65-74, 5:75-84, 6:85-99)	0.0%
<i>ageexit</i>	continuous	Age at death/censoring	0.0%
<i>diagyear</i>	continuous	Calendar year of diagnosis	0.0%
<i>time</i>	continuous	Survival time in years	0.0%
<i>timegrp</i>	categorical	survival time (0: <0.5 years, 1:≥0.5 & <1.0, 2: ≥1.0 & <2.0, 3 ≥2.0 & <5.0, 4≥5.0)	0.0%
<i>fup2</i>	binary	survival time (1:first year, 2: second to fifth year). This variable will be used when an interaction term of a predictor variable with follow-up time is to be included in the model.	0.0%
<i>status</i>	binary	status (0:alive, 1:dead)	0.0%

Step 2: Imputation of incomplete variables

We illustrate how to use the Stata command **ice** to impute plausible values for each of the three incomplete variables (stage, morphology and grade). We derive imputation models that include the other two incomplete variables, as well as the complete variables (sex, age, deprivation, comorbidity, treatment and colorectal site), outcome (vital status and follow-up

time in years) and two interactions: one between deprivation and follow-up time, and one between age and follow-up time. For the purpose of imputation, follow-up time is categorized in two intervals of 6 months in the first year, then yearly up to 5 years and then as a single interval for 5 years and later. A total of 10 ‘completed’ data sets are constructed, and saved in the data file ‘impute’.

ice imputes missing values using multiple imputation by chained equations(MICE). The **cmd** (option) specifies the regression command to be used when this incomplete variable becomes the dependent variable in the switching regression procedure. The variables *stage* and *grade* are categorical variables, by default **ice** will treat them as unordered categorical variables, and therefore **mlogit** will be used in the prediction model. We can change this by **cmd** (*stage grade*: ologit) to use ordinal logistic regression instead. One can also use the prefix **o.stage** and **o.grade** telling **ice** to impute missing values for the ordered categorical variables *stage* and *grade* using the **ologit** command (please refer to ice.help for detailed description of the use of these prefix).

Imputed and nonimputed variables are stored to a new file called “impute”. The original data, including missing values, are output by **ice** to the file of imputations, indexed by *_mj* = 0.

Please note that you might need to download some of the programs used in this applied example. To download ice you can type **findit ice** from within Stata. To ensure that you have the current version of **ice** if you already have an older version, you can type

```
which ice
ssc describe ice
ssc install ice, replace
```

Now we are set to generate our imputed data, we chose *m*=10 for ten copies of the imputed datasets. The **seed** option is to ensure that our results are reproducible.

```
ice stage i.treatment i.site i.charlson hist grade sex diagyr
i.deprivation*fup2 i.agegrp*fup2 i.timegrp status, saving
(impute,replace) cmd(stage grade: ologit) seed(123456) m(10) replace
```

#missing values	Freq.	Percent	Cum.
0	16,223	54.88	54.88
1	7,230	24.46	79.33
2	3,051	10.32	89.65
3	3,059	10.35	100.00
Total	29,563	100.00	

Variable	Command	Prediction equation
dep	logit	[No missing data in estimation sample]
stage		dep age treatment site charlson hist grade sex timegrp status diagyr deprivation*fup2 agegrp*fup2
age		[No missing data in estimation sample]
treatment		[No missing data in estimation sample]
site		[No missing data in estimation sample]

charlson		[No missing data in estimation sample]
hist	mlogit	dep stage age treatment site charlson grade sex timegrp status diagyr deprivation*fup2 agegrp*fup2
grade	ologit	dep stage age treatment site charlson hist sex timegrp status diagyr deprivation*fup2 agegrp*fup2
sex		[No missing data in estimation sample]
timegrp		[No missing data in estimation sample]
status		[No missing data in estimation sample]
diagyr		[No missing data in estimation sample]
deprivat ion*fup2		[No missing data in estimation sample]
agegrp*f up2		[No missing data in estimation sample]

```
Imputing 1..2..3..4..5..6..7..8..9..10..
file impute.dta saved
```

Step 3: Declare data to be survival-time data, and estimate relative survival.

We **stset** the data to specify all deaths as events. We then use **strs** to estimate relative survival (the ratio of the observed probability of survival (S) of the cancer patients and the probability of survival that would have been expected (E) if the patients had had the same survival probability as in the general population) using actuarial methods (Dickman, 2006). Results are saved in life tables stratified by *deprivation*, *stage*, *agegrp*, *charlson*, *hist*, *grade*, *sex*, *site* the variables specified in the **by** option. The **strs** command creates new variables e.g. *d* (observed number of deaths in the interval), *d_star* (expected number of deaths in the interval), *start* (start time of interval), *end* (end time of interval). Time Intervals are specified in the **break** (option). **save (replace)** creates two output data sets, *individ.dta* contains one observation for each patient for each life table interval, and *grouped.dta* contains one observation for each life table interval. These are applied to each of the 10 imputed datasets, and therefore ten files of *individ`i`.dta* and *group`i`.dta* are saved where *i*=1,..., 10

'lifetable' is the expected mortality data file.

```
forval i = 1(1)10 {
  use "impute" if _mj=="`i'",clear
  gen id=_mi
  stset ageexit, fail(status) origin(agediag) id(id)
  strs using "lifetable", break(0(.25)1 1.25(.25)2 3(1)8) mergeby(sex age year
  deprivation) diagage(age) diagyear(diagyr) by(deprivation stage age charlson hist
  grade sex site _mi _mj) attage(age) attyear(year) survprob (survprob)
  savind(individ`i`, replace) notable
}
```

We next append the 10 `individ`i`.dta` files, into one file with the name of `individ.dta`

```
use "individ1.dta",clear
local i 2
forval i= 2(1)10 {
append using "individ`i`.dta"
save "individual.dta", replace
}
```

Step 5. Step 4: Estimate the excess hazard of death of the colorectal cancer patients with multivariable regression using a generalised linear model with Poisson error (Dickman et al., 2004), in each of the imputed datasets and combine results using Rubin Rules.

Now that we created our completed `individ.dta` dataset we can fit our generalized linear model to estimate the excess hazard of death adjusting for covariates, in each of the imputed datasets. The command **micombine** pools estimates of the ten models using Rubin Rules. Please note that although we hardly ever need to use Stata's **xi**: dummy variable and interaction creator directly with **ice**, it can still be used in **micombine**.

```
use "individual.dta", clear
xi: micombine glm d i.end i.agegrp i.stage i.hist i.grade i.site sex i.charlson, eform
family(pois) link(rs d_star) lnoffset(y)
```

To test the significance of the interaction of *deprivation* and *fup2*, we save the estimation results of **micombine** using **estimates store**. We then fit the same model with the interaction term and use the likelihood ratio test with the command **lrtest** to test its significance.

```
estimates store A
xi: micombine glm d i.end i.agegrp i.stage i.hist i.grade i.site sex i.charlson i.fup2*deprivation,
eform family(pois) link(rs d_star) lnoffset(y)
lrtest A
```

We carried out a similar likelihood ration test for the significance of the interaction of *agegrp* and *fup2*

```
xi: micombine glm d i.end i.agegrp i.stage i.site i.charlson i.hist i.grade sex
i.deprivation*fup2, eform family(pois) link(rs d_star) lnoffset(y)
```

estimates store B

```
xi:micombine glm d i.end i.stage i.site i.charlson i.hist i.grade sex i.derivation*fup2  
i.agegrp*fup2, eform family(pois) link(rs d_star) lnoffset(y)
```

lrtest B

The final multivariable analysis model is then fitted to estimate the log of excess hazard ratio of death adjusting for socio-demographic and clinical variables, and the interaction between *agegrp* and *fup2* and *derivation fup2*

```
xi:micombine glm d i.end i.stage i.site i.charlson i.hist i.grade sex i.deprivation*fup2  
i.agegrp*fup2, eform family(pois) link(rs d_star) lnoffset(y)
```

Reference list

Dickman PW (2006) Estimating and modelling relative survival in SAS and Stata. www.pauldickman.com/rsmode/index.php, accessed 11 May 2009

Dickman PW, Sloggett A, Hills M, Hakulinen T (2004) Regression models for relative survival. *Stat Med* 23: 51-64

Little RJA, Rubin DB (1987) *Statistical Analysis with Missing Data*. Second Edition John Wiley & Sons: New York

Royston P (2005a) Multiple imputation of missing values: update. *SJ* 5: 188-201

Royston P (2005b) Multiple imputation of missing values: Update of ice. *SJ* 5: 527-536

Royston P (2007) Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *SJ* 7: 445-464

Royston P (2009) Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *SJ* 9: 466-477

Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 18: 681-694