Module 14: Missing Data
Stata Practical

Jonathan Bartlett & James Carpenter
London School of Hygiene & Tropical Medicine
www.missingdata.org.uk
Supported by ESRC grant RES 189-25-0103 and MRC grant G0900724

Contents

Introduction to the Youth Cohort Study dataset

You will be analysing data from the Youth Cohort Study of England and Wales (YCS)[1]. The YCS is a postal survey of young people. We will use data from the 1995 cohort, restricted to those young people who were at comprehensive schools (n=12,884) when the survey took place.

Our analyses will focus on variables recording GCSE attainment, parental socio-economic class, gender, and ethnicity. In particular our interest will focus on models for the young person's GCSE attainment score, with the other variables as covariates or explanatory variables. Such a model is of interest in order to investigate differences in GCSE attainment between ethnic and social economic groups, relative to gender differences (Connolly 2006). Table 14.1 describes the variables included in the dataset.

Table 14.1. Variables contained in Youth Cohort Study dataset

Variable name	Description and coding
t0score2	GCSE score – truncated year 11 exam point score
gender	Gender (1 = boys, 0 = girls)
t0ethnic	Ethnicity 1 = white 4 = black 5 = Indian 6 = Pakistani 7 = Bangladeshi 9 = Other Asian 10 = other response
t0parsc4	Parents' National Statistics Socio-economic classification 1 = managerial & professional 2 = intermediate 3 = working

The GCSE score is formed by assigning numerical scores to the grades obtained by a child at GCSE (A/A*=7 through to grade G=1), truncated at 12 grade A/A*s (giving a maximum score of 84).

The original YCS data also contains a weight variable, based on the sampling scheme used in the survey. Since our aim here is to illustrate the missing data concepts and methods we have introduced, we do not use the weights in this analysis. We emphasise that the analyses shown here are intended to be illustrative of the missing data concepts and methods we have introduced, and should not be interpreted as a substantive analysis of these data.

---

[1] We thank the depositors of the Economic and Social Data Service (ESDS) data collection SN 5765 'Youth Cohort Time Series for England, Wales and Scotland, 1984-2002', and the depositors of the constituent studies, for their permission to make these data available for teaching purposes. We also thank the ESDS (www.esds.ac.uk) for their assistance in obtaining these permissions and through whose website the data were made available to us.

P14.1 The Model of Interest

Throughout the practical we shall assume that our model of interest is the linear regression of GCSE score on gender, ethnicity and parental SEC. Ordinarily we would fit this model in Stata using:

```
xi: regress t0score2 gender i.t0ethnic i.t0parsc4
```

We will keep this model of interest in mind when investigating missingness in the variables and when considering how to handle any missing values.

P14.2 Investigating Missingness

In this section we investigate missingness in the YCS data. Load "10.2.dta" into memory and open the do-file for this lesson:

From within the LEMMA Learning Environment
- Go to **Module 14: Missing Data**, and scroll down to *Stata* **Datasets and Do-files**
- Click "14.2.dta" to open the dataset

P14.2.1 Investigating quantity and patterns of missingness

We begin by investigating how many missing values there are in the variables included in the dataset, using Stata's `misstable` summarize command:

```
. misstable summarize t0score2 t0parsc4 t0ethnic gender
                                                    Obs<.
                                      +-----------------------------
                    |                 | Unique
           Variable |  Obs=.    Obs>.    Obs<. | values      Min        Max
         -----------+------------------------------+-----------------------------
           t0score2 |    129             12,755 |    85        0         84
           t0parsc4 |  1,576             11,308 |     3        1          3
           t0ethnic |    156             12,728 |     7        1         10
         ------------------------------------------------------------------
```

We first note that the gender variable has not been included in the output – this is because the variable has no missing values. Next we see that the parental SEC has the most missing values (1,576), with GCSE score and ethnicity having fewer missing values.

Next we examine the patterns of missingness in these three variables. We use the `misstable` patterns command to tabulate which patterns of missingness occur and how frequently each pattern occurs:

```
. misstable patterns t0score2 t0parsc4 t0ethnic gender , freq
  Missing-value patterns
    (1 means complete)
            |   Pattern
  Frequency | 1  2  3
  ----------+-----------
     11,188 | 1  1  1
            |
      1,422 | 1  1  0
        104 | 1  0  0
         77 | 0  1  1
         41 | 0  1  0
         41 | 1  0  1
          9 | 0  0  0
          2 | 0  0  1
  ----------+-----------
     12,884 |
  Variables are  (1) t0score2  (2) t0ethnic  (3) t0parsc4
```

The output from `misstable` patterns shows, for the specified variables, each pattern of missing data which occurs, ordered according to the frequency with which

they occur. From the first row in the table, we see that there are 11,188 young people for whom all three variables (ethnicity, GCSE score, and parental SEC) are observed. The most common pattern which has some missing values is when GCSE score and ethnicity are observed but parental SEC is missing (n=1,422). The next most commonly occurring pattern is where GCSE score is observed but ethnicity and parental SEC are missing (n=104). We then see that all the other possible missingness patterns occur, but with smaller frequencies.

P14.2.2 Investigating the missingness mechanism

Since missingness occurs in three of the variables in the dataset, we can think of there being an underlying mechanism which determines missingness for each of the variables ethnicity, GCSE score, and parental SEC. Since the majority of missing values occur in the parental SEC variable, however, we shall focus on investigating missingness in this variable, since the analysis is likely most sensitive to assumptions concerning this. The other patterns of missing data we (implicitly) assume are either MCAR or possibly MAR given other observed values.

From the output from `misstable` patterns, we saw that when parental SEC is missing, ethnicity and GCSE score are mostly observed. We can therefore investigate how missingness in parental SEC is related both to these two variables and to the fully observed variable gender.

To investigate which variables are predictive of missingness in the parental SEC variable we first define a binary variable which indicates whether the parental occupation variable is observed (=1) or missing (=0):

```
gen r_t0parsc4=(t0parsc4!=.)
```

Next, we fit a logistic regression model for the variable **r_t0parsc4**, with gender as covariate (we could also have simply performed a chi-squared test):

```
. xi: logistic r_t0parsc4 gender

Logistic regression                              Number of obs   =      12884
                                                 LR chi2(1)      =       1.21
                                                 Prob > chi2     =     0.2716
Log likelihood = -4786.1437                      Pseudo R2       =     0.0001
-------------------------------------------------------------------------------
  r_t0parsc4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
      gender |   .9424493   .0507889    -1.10   0.271     .8479816    1.047441
-------------------------------------------------------------------------------
```

From this we see that there is no evidence (p=0.271) that missingness of parental SEC is related to gender of the young person.
Next we fit the logistic regression model with GCSE score (**t0score2**) as a covariate:

```
. xi: logistic r_t0parsc4 t0score2

Logistic regression                              Number of obs   =      12755
                                                 LR chi2(1)      =     639.66
                                                 Prob > chi2     =     0.0000
```

```
Log likelihood = -4351.1228                          Pseudo R2       =    0.0685

------------------------------------------------------------------------------
   r_t0parsc4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     t0score2 |   1.039299   .0016203    24.72   0.000     1.036128    1.04248
------------------------------------------------------------------------------
```

This gives an estimated odds ratio corresponding to a 1 unit increase in GCSE score of 1.039 (95% CI 1.036 to 1.042, p<0.001). There is thus strong evidence that those with higher GCSE scores are more likely to have parental SEC observed. Or conversely, that those with low GCSE scores are more likely to have parental SEC missing.

At this point we can now say that we have strong evidence against the parental SEC variable being MCAR, because we have found a variable which is predictive of missingness.

Now we fit the logistic regression model with ethnicity as a categorical covariate:

```
. xi: logistic r_t0parsc4 i.t0ethnic
i.t0ethnic        _It0ethnic_1-10     (naturally coded; _It0ethnic_1 omitted)

Logistic regression                              Number of obs   =      12728
                                                 LR chi2(6)      =     547.63
                                                 Prob > chi2     =     0.0000
Log likelihood =  -4266.615                      Pseudo R2       =     0.0603
------------------------------------------------------------------------------
   r_t0parsc4 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_It0ethnic_4 |   .1836708   .0267837   -11.62   0.000     .1380112    .2444364
_It0ethnic_5 |   .5756491   .0929915    -3.42   0.001     .4194238    .7900645
_It0ethnic_6 |   .1492296   .0187741   -15.12   0.000     .1166187    .1909598
_It0ethnic_7 |   .0512417   .0101832   -14.95   0.000     .0347109    .0756451
_It0ethnic_9 |   .3142811   .0668424    -5.44   0.000     .2071486    .4768203
_It0ethni~10 |   .2882334   .0493141    -7.27   0.000     .2061159    .4030668
------------------------------------------------------------------------------
```

Since ethnicity is the only covariate here, the p-value next to the text 'Prob > chi2' in the output corresponds to a joint test that ethnicity is not associated with the probability of parental SEC being missing. There is thus strong evidence that ethnicity is related to this probability. In particular, since Stata takes the lowest level 1=white as the reference category, we see that all of the other ethnic groups are less likely to have parental SEC observed relative to white participants.

So far we have investigated crude associations between the missingness indicator for parental SEC and the other variables. It is also of interest to investigate the independent contribution of the variables in terms of the probability of parental SEC being missing. We therefore conclude the section by fitting the logistic regression model with the three variables gender, GCSE score and ethnicity simultaneously as covariates:

```
. xi: logistic r_t0parsc4 gender i.t0ethnic t0score2
i.t0ethnic        _It0ethnic_1-10     (naturally coded; _It0ethnic_1 omitted)

Logistic regression                              Number of obs   =      12610
                                                 LR chi2(8)      =    1094.72
```

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:

http://www.cmm.bris.ac.uk/lemma

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.