Århus November 19 2013

Applied Statistical Analysis with Missing Data

Exercise 2

Consider the dataset `ess2e03_scand.dta`, which contains a Scandinavian subset of the data collected in a multi-country interview survey known as the European Social Survey. Round 2 was conducted in 2004/05 comprising questions about compliance with last prescribed drug and other relevant predictors (and many more)[1].

Here we will consider the four variables:

cntry	Country of interviewee.
edulvl	Highest level of education
incmean	Annual household income
total_noncompl	Non-compliance at last prescription of a new medication

Descriptives and basic relations

Q1: Investigate the four variables: How are they coded, are they categorical or continuous? How many missing values are there in each variable? What is the pattern of missing values? (Hint: use the command `-misstable summarize-` and `-misstable patterns-`, see `-help misstable-` for more info)

Let us now only consider the complete data set, i.e. where all four variables are observed.

Q2: What is the apparent relation between education level and non-compliance?

Q3: Make a linear regression of income on education? Consider a log-transformation of income, and validate the model.

Predicting missingness

We now consider the dataset in which education is observed (n = 7,707).

Q4: Does the probability of income being missing depend on educational level? On non-compliance?

Q5: Does the probability of non-compliance being missing depend on educational level? Onincome?

[1] For more information on the entire dataset, see http://ess.nsd.uib.no/ess/round2/. For an intro-duction to the objective with respect to non-compliance, see Larsen et al, BMC Public Health (2009, 9:145).