A short guide to inspecting, managing and analysing longitudinal data using SN 8715: "Understanding Society: Longitudinal Teaching Dataset, Waves 1-9, 2009-2018"

Version 1.0, October 2020

Understanding Society

Economic and Social Research Council

Table of Contents

1. Inspecting and managing the datafile (using Stata)

Open the datafile, `longitudinal_td.dta`. **This data fie looks like the data showcased in Table 3.2 in the User Guide (copied below in Table 1.1).**

Table 1.1: Example of a longitudinal data in long format

pidp	wave	age_dv	cob_dv	sex_dv	mstat_dv
1	1	20	Bangladesh	Woman	Single
1	2	21	Bangladesh	Woman	Cohabiting
1	3	22	Bangladesh	Woman	Cohabiting
2	1	53	Jamaica	Man	Married
2	2	54	Jamaica	Man	Married
3	1	59	England	Woman	Divorced
4	2	60	England	Woman	Married

You can inspect the file using the usual Stata commands:

```
describe
summarize
```

Or you can browse the data

```
browse
```

Or list out a few variables for a few cases on the screen

```
sort pidp wave
list pidp wave sex_dv ethn_dv doby_dv jbstat mstat_dv ///
in 1/20, sepby(pidp)
```

Use these commands to examine the data. Are there any odd values for variables? For example, there are negative values for some of the income variables. Consider why this may happen. Take a look at the income section of the user guide to see how these income variables are computed. As the household total and labour income variables include self-employment income which could be negative if there was a loss, these income variables could also be negative. You can check how many individuals there are who live in households with negative gross household monthly income.

```
count if fihhmngrs_dv<0
```

You will see that there are 87 such cases across nine waves. What do you want to do with these cases? Depending on your analysis you can leave these as they are or recode these to 0 or £1.

```
replace fihhmngrs_dv=1 if fihhmngrs_dv<0
```

You can also consider using income quantiles in your analysis in which case only the order of the values will matter, and you will not need to recode the negative values:

```
xtile hhgrinc4=fihhmngrs_dv, nq(4)
```

Some income values are very high and so you could consider excluding these outliers, say the top 1%. You can use this command to produce percentiles and then use Stata's saved commands to identify cases with incomes above the 99th percentile.

```
su fihhmngrs_dv, d
```

Remember Stata considers missing data values (.) as the highest number.

```
generate veryhighinc=1 if fihhmngrs_dv>r(p99) & fihhmngrs_dv<.
replace  veryhighinc=0 if fihhmngrs_dv<r(p99)
```

You can see there are around 2000 individuals with very high incomes.

```
tab veryhighinc
```

This is a longitudinal or panel data across nine waves. But are all variables available in all waves?

```
tabstat _all, by(wave)
```

This will by default produce the mean of every variable in every wave. So, mean of variables that were not asked in a specific wave will be missing for those waves. You can check this list against the Appendix Tables A1-A9 list the variables and the waves in which they appear.

To inspect the panel structure of the data, first tell Stata about the panel structure of the data. In this case individuals are identified by `pidp` **and the time variable is** `wave`**, so:**

```
xtset pidp wave
```

You can see the time/wave pattern of the data by:

```
xtdescribe, patterns(50)
```

As you can see these patterns reflect the response patterns discussed earlier. To check which variables remain constant across waves type:

```
xtsum
```

This command produces summary statistics for all variables taking into account the panel structure of the data. You will see that for some variables the "within standard deviation" is 0. That means these are time invariant variables: there is no variation for that variable across waves for every person. For example: `doby_dv ethn_dv sex_dv cob_dv`

You can produce lagged and lead variables. For example, to produce the value of marital status in the last wave and the next wave, respectively:

```
g l_mstat_dv=L1.mstat_dv
g n_mstat_dv=F1.mstat_dv
```

Table 1.2: Example of a longitudinal data in long format

pidp	wave	age_dv	cob_dv	sex_dv	mstat_dv	l_mstat_dv	f_mstat_dv
1	1	20	Bangladesh	Woman	Single	.	Cohabiting
1	2	21	Bangladesh	Woman	Cohabiting	Single	Cohabiting
1	3	22	Bangladesh	Woman	Cohabiting	Cohabiting	.
2	1	53	Jamaica	Man	Married	.	Married
2	2	54	Jamaica	Man	Married	Married	.
3	1	59	England	Woman	Divorced	.	Married
4	2	60	England	Woman	Married	Divorced	.

To do the same for variable values n waves prior or n waves later use `Ln.` **and** `Fn.`**, respectively.**

Suppose there is a question that is asked once and you assume that variable value does not change over time and you would like to include it in analysing a model using data from many waves. To do this you will need to copy this variable over to other waves. For example, the BMI variable, `bmi_dv`, **is only available in Wave 1. You can copy this over to other waves by using Stata's** `bysort` **and** `egen` **functions to produce a time-invariant value:**

```
bys pidp: egen bmi_dv_fixed=mean(bmi_dv)
```

You can also verify that this variable has valid values across waves and it is the same across waves:

```
tabstat bmi_dv_fixed, by(wave)
xtsum bmi_dv_fixed
```

Table 1.3: Example of a longitudinal data in long format

pidp	wave	age_dv	cob_dv	sex_dv	mstat_dv	bmi_dv	bmi_dv_fixed
1	1	20	Bangladesh	Woman	Single	19	19
1	2	21	Bangladesh	Woman	Cohabiting	.	19
1	3	22	Bangladesh	Woman	Cohabiting	.	19
2	1	53	Jamaica	Man	Married	22	22
2	2	54	Jamaica	Man	Married	.	22
3	1	59	England	Woman	Divorced	25	25
3	2	60	England	Woman	Married	.	25

2. Analysing the data (using Stata) – Part 1

Here are examples of some types of estimations that you can do with this type of longitudinal data.

To estimate wave-on-wave transitions for variables, say, `mstat_dv` **type**:

```
xttrans mstat_dv
```

As you can see, between two consecutive waves, around 1% of single people get married or form same-sex civil partnerships and 4% start living together as a couple (cohabiting).

```
xttrans mstat_dv if sex_dv==1 & age_dv>=30 & age_dv<=39
xttrans mstat_dv if sex_dv==2 & age_dv>=30 & age_dv<=39
```

For 30-39 year old men and women these percentages are higher: 4% and 7% for men, 2% and 6% for women.

You can also check if the mean value of GHQ has changed across waves:

```
mean scghq1_dv, over(wave)
test [scghq1_dv]1 = [scghq1_dv]2 = [scghq1_dv]3 = [scghq1_dv]4 ///
   = [scghq1_dv]5 = [scghq1_dv]6 = [scghq1_dv]7 = [scghq1_dv]8 ///
= [scghq1_dv]9
```

You can run simple pooled cross-section models where you control for the interview year which captures various global factors relevant for that year. For example, in a model of mental health or distress as measured by GHQ:

```
regress scghq1_dv i.sex_dv c.age_dv##c.age_dv i.sf1_dv ///
c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv
```

You can also estimate a lagged dependent variable model, where a lagged value of the dependent variable is included as an explanatory variable:

```
g l_ghq=L1.scghq1_dv
regress scghq1_dv i.sex_dv c.age_dv##c.age_dv i.sf1_dv ///
c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv ///
c.l_ghq
```

You can also estimate this model using various panel data methods, such as a fixed effects model:

```
xtreg scghq1_dv i.sex_dv c.age_dv##c.age_dv i.sf1_dv ///
c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv, fe
```

or a random effects model:

```
xtreg scghq1_dv i.sex_dv c.age_dv##c.age_dv i.sf1_dv ///
c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv, re
```

3. Analysing the data (using Stata): Part 2

3.1 Which weights to use? Why to use weights?

Understanding Society samples are selected with unequal selection probability. Additionally, not everyone selected participates in the survey. This non-response and attrition is not random, that is, particular types of people are more likely to respond. As a result there are more people with some characteristics in the sample than in the population. In other words, unweighted analysis estimates will be biased in favour of the types of individuals over-represented in the sample. Weights are designed to undo this effect by giving higher (or lower) weight to types of individuals who are under (or over) represented in the sample.

An under or over representation of some types of individuals could be by design as in the case of the EMBS in Understanding Society or due to non-random attrition or non-response. The EMBS was designed to include ethnic minority individuals such that there were at least 1000 adult interviews with individuals from these five ethnic groups: Black African, Black Caribbean, Bangladeshi, Pakistani, Indian. These ethnic groups comprise around 8% of the UK population and so their sample size in a representative sample of say, 40,000, would be around 3,200 which would not be enough to study each group separately. Adding the EMBS increased the number of adult interviews of individuals from these five ethnic groups from around 2,400 to 7,400. As discussed earlier not everyone eligible for an interview participates in the survey due to different reasons. For example, those who move are more difficult to trace and locate. As younger populations and recent migrants are more likely to move, the unweighted sample may under represent these groups.

The weights provided adjust for these types of over/under representations. If you are analysing data from the first n waves, then use the weight variable with the suffix `indinus_lw_n`.

This dataset includes a few variables (`sf1_dv, sf12pcs_dv, sf12mcs_dv, scghq1_dv, scghq2_dv, swemwbs_dv, sclfsato`**) which were asked in the self-completion questionnaire. As discussed earlier not everyone who participates in the adult interviews**

completes this additional questionnaire. So, if the model you are estimating includes any of these variables then the analysis sample will be restricted to those who completed the self-completion questionnaire and you will need to account for this additional level of non-response when producing estimates. To do this use the self-completion weights: `indscus_lw_n`.

3.2 Why does sample design matter?

The samples in this datafile were selected using a complex design. Unless you indicate the complex sample design most statistical software assumes the data is from a sample selected with a simple random sample design and the standard errors estimated using these assumptions will be incorrect. Our sample has unequal selection probabilities, clustering and stratification. Clustering means that first a few clusters or groups of population units (say, households) are selected (referred to as Primary Sampling Units) and then all or few of these population units are chosen within each cluster. The variable representing the primary sampling unit (PSU) in this data is psu. Stratification means that the population is first divided into mutually exclusive and exhaustive groups, referred to as strata, based on one or more characteristics of the population units (e.g. region) and then a few or all population units are selected from *every* strata. The variable representing the strata in this data is strata.

Box 1 highlights the different commands and packages available with four key statistical software packages to conduct analysis accounting for weights and sample design.

Box 1 How to account for weights, clustering and stratification in different statistical softwares

To specify the weights, PSU and strata and to conduct analysis taking these into account, use

- SVY suite of commands in Stata
- SVYDESIGN in R
- CSPLAN commands along with COMPLEXSAMPLES in SPSS
- SURVEY procedures in SAS

To know more about weights and sample design in Understanding Society see here. Other resources for learning how to take into account complex sample design:

- Introduction to Understanding Society using Stata, SPSS, SAS & R online courses here
- Section 4.3 of "Guide to Using Weights and Sample Design Indicators with ESS Data" here

3.3 Analysis using weights and accounting for sample design

To account for the sample design and to include weights first tell Stata which variables represent these items. To use all nine waves of data use the longitudinal weight from the last wave: `indinus_lw_9`. But as GHQ was asked in the self-completion questionnaire you will need to use the weight that also accounts for the additional non-response at the self-completion stage: `indscus_lw_9`.

```
svyset psu [pweight = indscus_lw_9], strata(strata)
```

Then produce mean GHQ across waves, and to estimate the GHQ model (described above) using OLS after accounting for weights and sample design, simply type `svy:` at the beginning of the usual Stata estimation commands:

```
svy: mean scghq1_dv, over(wave)
test [scghq1_dv]1 = [scghq1_dv]2 = [scghq1_dv]3 = [scghq1_dv]4 ///
   = [scghq1_dv]5 = [scghq1_dv]6 = [scghq1_dv]7 = [scghq1_dv]8 ///
   = [scghq1_dv]9

svy: regress scghq1_dv i.sex_dv i.ethn_dv c.age_dv##c.age_dv ///
i.sf1_dv c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv
c.l_ghq
```

But as you will see no standard errors have been estimated. Sometimes it may so happen that the sample you are analysing includes strata with only one primary sampling unit. In those cases, it is not possible to compute the strata variance and so Stata will not produce standard errors. In such cases there are a number of methods you can use to work around this problem. We will use one such solution:

```
svyset psu [pweight = indscus_lw_9], strata(strata) ///
singleunit(scaled)

svy: mean scghq1_dv, over(wave)
test [scghq1_dv]1 = [scghq1_dv]2 = [scghq1_dv]3 = [scghq1_dv]4 ///
   = [scghq1_dv]5 = [scghq1_dv]6 = [scghq1_dv]7 = [scghq1_dv]8 ///
   = [scghq1_dv]9

svy: regress scghq1_dv i.sex_dv i.ethn_dv c.age_dv##c.age_dv ///
i.sf1_dv c.fihhmngrs_dv c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv
c.l_ghq
```

You can also estimate a fixed effects model using weights and account for clustering:

```
xtset pidp wave

xtreg scghq1_dv c.age_dv##c.age_dv i.sf1_dv c.fihhmngrs_dv ///
c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv ///
[pw = indscus_lw_9], fe vce(cluster psu)
```

As expected no estimates were produced for time invariant variables `sex_dv` **and** `ethn_dv`.

But you can estimate the model separately by these variable categories or interact these variables for a variable of interest. For example, to estimate separate models for women of some larger ethnic groups in the UK:

```
foreach i in 1 4 9 10 11 14 15 {
      xtreg scghq1_dv c.age_dv##c.age_dv i.sf1_dv c.fihhmngrs_dv ///
      c.hhsize c.ndepchl i.jbhas_dv i.intdaty_dv ///
      if sex_dv==2 & ethn_dv==`i' ///
      [pw = indscus_lw_9], fe vce(cluster psu)
}
```

This pack includes the Stata syntax file, `longitudinalTD_analysis_dofile.pdf` **& output file** `longitudinalTD_analysis_logfile.pdf` **which you can use to compare the results from running the Stata commands discussed in this section. If you are thinking of printing it please note that it is a very long file and so please consider whether you need to print the whole file.**

Appendix A: Variable list and description

Table A1: Identifiers and interview information	
Variable name	**Description**
pidp	Individual identifier (unique within and across waves)
wave	Interview wave
hidp	Household identifier (unique within a wave, not across waves)
buno_dv	Benefit unit within a household
intdaty_dv	Interview year
intdatm_dv	Interview month
intdatd_dv	Interview day
indmode	Interview mode

Table A2: Sampling and weight variables	
Variable name	**Description**
hhorig	Sample origin
psu	Primary sampling unit
strata	Sampling strata
sampst	Sample status
scflag_dv	Whether completed self-completion part of the adult questionnaire
indinus_lw_n	Longitudinal weight to be used when analysing responses to individual adult interview excluding self-completion questions from Waves 1 to n
indscus_lw_n	Longitudinal weight to be used when analysing responses to individual adult interview including self-completion questions (or only responses to self-completion questions) from Waves 1 to n

Table A3: Residential information	
Variable name	**Description**
mvever [a]	Lived at this address whole life *(Waves: 1)*
mvmnth [a]	Month moved to current address, if not living at current address *(Waves: 1)*
mveyr [a]	Year moved to current address, if not living at current address *(Waves: 1)*
distmov_dv	Distance participant moved since last wave in Km *(Waves: 3-9)*
addrmov_dv	Participant changes address postcode since last wave *(Waves: 3-9)*
lkmove	Prefers to move from current address
xpmove	Expects to move from current address
gor_dv	Government office region
urban_dv	Whether participant lives in an urban or a rural area

Table A4: Socio-demographic & household characteristics	
Variable name	**Description**
age_dv	Age in years at time of interview
doby_dv	Year of birth
sex_dv	Sex
ethn_dv	Ethnic group (complete labels are shown in Appendix B)
cob_dv	Country of birth
bornuk_dv	Whether born in UK
yr2uk4	Year arrived into UK
hhsize_dv	Household size
hhtype_dv	Type of household based on household composition
tenure_dv	Whether the accommodation is owned or rented (private, social housing)
hhtype_dv	Composition of household, LFS-version
mstat_dv	De facto marital status
livesp_dv	Lives with spouse in household
cohab_dv	Lives with cohabitee in household
nchild_dv	Number of own children in household
depchl_dv	Whether dependent child - official definition
ndepchl_dv	Number of own dependent children in household
hiqual_dv	Highest educational qualification

Table A5: Health and wellbeing	
Variable name	**Description**
sf1_dv[a]	General health
bmi_dv	Body Mass Index *(Waves: 1)*
sf12pcs_dv[a]	SF-12 Physical Component Summary
sf12mcs_dv[a]	SF-12 Mental Component Summary
scghq1_dv[a]	Subjective wellbeing (GHQ): Likert
scghq2_dv[a]	Subjective wellbeing (GHQ): Caseness
swemwbs_dv[a]	Short Warwick-Edinburgh Mental Well-being Scale *(Waves: 1 4 7)*
sclfsato[a]	Satisfaction with life overall

[a]Asked in self-completion questionnaire. But General Health question was asked by interviewers and not included in the self-completion questionnaire in Wave 1, in all other waves it was included in the self-completion questionnaire.

Table A6: Labour market and job related information	
Variable name	**Description**
jbstat	Current labour force status
jbhas_dv	Whether did any paid work last week, and was this paid or self-employment
jbsoc00_cc	Current job: SOC 2000, condensed
jbsic07_cc	Current job: SIC 2007, condensed
jbnssec8_dv	Current job: Eight Class NS-SEC
jbmngr	Current job: Has managerial duties
jbsize	Current job: No. employed at workplace
jbterm_dv	Current job: Type of job contract
jbsect_dv	Current job: Type of organisation working for
jbhrs	Current job: no. of hours normally worked per week
jbot	Current job: no. of overtime hours in normal week
jbft_dv	Current job: Full or part-time employee
jbotpd	Current job: No. of hours worked as paid overtime
jbpl	Current job: Work location
jbttwt	Current job: Minutes spent travelling to work
workdis	Current job: Distance from work *(Waves: 1 2 4 6 8)*
worktrav	Current job: Mode of transport for journey to work
jbsat	Job satisfaction (1-7)
j2has	Has a second job?
j2semp	Second job: employee or self-employed
j2soc00_cc	Second job: SOC 2010, condensed
j2nssec8_dv	Second job: NSSEC 8 classes
j2hrs	Second job: no. of hours worked per month
jsboss	Self-employed: hires employees
jssize	Self-employed: number of employees
jshrs	Self-employed: hours normally worked per week
jstypeb	Self-employed: nature of employment
jsaccs	Self-employed: draws up profit/loss accounts
jspart	Self-employed: own account or partnership
jspl	Self-employed: work location
jsttwt	Self-employed: commuting time provided
jsttwtb	Self-employed: commuting time
jsworkdis	Self-employed: commuting distance *(Waves: 1)*
jsworktrav	Self-employed: mode of transport to work

Table A6: Labour market and job related information (continued)	
Variable name	**Description**
jbhad	Ever had paid employment *(Waves: 1)*
jlsemp	Last job: Employee or self-employed? *(Waves: 1)*
jlendy	Last job: year left job *(Waves: 1)*
jlendm	Last job: month left job *(Waves: 1)*
jlsoc00_cc	Last job: SOC 2010, condensed *(Waves: 1)*
jlsic07_cc	Last job: SIC 2007, condensed *(Waves: 1)*
jlnssec8_dv	Last job: Eight Class NS-SEC *(Waves: 1)*
jlmngr	Last job: Has managerial duties *(Waves: 1)*
jlboss	Last job: hired employees *(Waves: 1)*
jlsize	Last job: number of people employed at workplace *(Waves: 1)*
paygu_dv	Usual gross monthly pay of current job
paygu_if	Whether paygu_dv that was imputed
paynu_dv	Usual net monthly pay of current job
paynu_if	Whether paynu_dv that was imputed
j2pay_dv	Gross monthly pay of second job
j2paynet_dv	Net monthly pay of second job
j2pay_if	Whether j2pay_dv that was imputed
seearngrs_dv	Gross monthly self-employment earnings
seearnnet_dv	Net monthly self-employment earnings
seearngrs_if	Whether seearngrs_dv that was imputed

Table A7: Income	
Variable name	**Description**
fimngrs_dv	Gross monthly personal income
fimnnet_dv	Net monthly personal income, no deductions
fimngrs_if	Share if fimngrs_dv that was imputed
fihhmngrs_dv	Gross monthly household income (before housing costs)
fihhmnnet1_dv	Net monthly household income (before housing costs), no deductions
fihhmngrs_if	Share of fihhmngrs_dv that was imputed
ieqmoecd_dv	Modified OECD equivalence scale

Table A8: Political behaviour and opinions	
Variable name	**Description**
vote1	supports a particular political party *(Waves: All except 8)*
vote2	closer to one political party than others *(Waves: All except 8)*
vote3	Party would vote for tomorrow *(Waves: All except 8)*
vote4	Which political party closest to *(Waves: All except 8)*
vote5	strength of support for stated party *(Waves: All except 8)*
vote6	level of interest in politics *(Waves: All except 8)*
vote7	voted in last general election *(Waves: 2 7 8 9)*
vote8	Party voted for in last general election *(Waves: 2 7 8 9)*
votenorm	Voting as a social norm (*Waves: 2 3 6 9)*
voteintent	voting intention *(Waves: 2 3 6 9)*
grpbfts	Group benefit from voting *(Waves: 2 3 6 9)*
perbfts	Personal benefit in *voting (Waves: 2 3 6 9)*

Table A9: Environmental attitudes and behaviours *(Waves 1 & 4)*	
Variable name	**Description**
envhabit1	How often leave your TV on standby for the night
envhabit2	How often switch off lights in rooms that aren't being used
envhabit3	How often keep the tap running while you brush your teeth
envhabit4	How often put more clothes on when you feel cold rather than putting the heating on or turning it up
envhabit5	How often decide not to buy something because you feel it has too much packaging
envhabit6	How often buy recycled paper products such as toilet paper or tissue
envhabit7	How often take your own shopping bag when shopping
envhabit8	How often use public transport (e.g. bus, train) rather than travel by car
envhabit9	How often walk or cycle for short journeys less than 2 or 3 miles
envhabit10	How often car share with others who need to make a similar journey
envhabit11	How often take fewer flights when possible

Appendix B: Ethnic group question showcard

White

1. British/English/Scottish/Welsh/Northern Irish
2. Irish
3. Gypsy or Irish Traveller
4. Any other white background

Mixed

5. White and Black Caribbean
6. White and Black African
7. White and Asian
8. Any other mixed background

Asian or Asian British

9. Indian
10. Pakistani
11. Bangladeshi
12. Chinese
13. Any other Asian background

Black / African / Caribbean / Black British

14. Caribbean
15. African
16. Any other Black background

Other ethnic group

17. Arab
97. Any other ethnic group