Exercises for ”Short course: Panel surveys in social and economic research and the treatment of nonresponse”,

24 Feb – 27 Feb 2014
4 ECTS (+ 2 ECTS with completed homework)

Prof. Dr. Ulrich Rendtel

Version 24. Februar 2014

In order to do these exercises you need:

- Access to the internet.
- Access to SAS and the LEM EXE-file.
- Access to the SAS-programs Result_soepinfo, Generate_flat, Generate_long_file, Generate_panel, divorce.
- Access to the SAS-data sets New, Flat, Long, Analyse, Human_cap, Hmohiv, Universe.

Meta information

Enter the English version of the SOEPINFO from the SOEP homepage.

Generate from the SOEP a panel data set containing some variables that are useful in the analysis of earnings. The variables of interest are: Monthly gross-income, tenure (i.e. period at the same firm), satisfaction with earned income, sex and year of birth.

1 Use the SOEPINFO URL. Use icon "Main actions" and select the questionnaire search. You have to browse the 2005 questionnaire. Mind the naming of the variables which depends on the position in the questionnaire. Select the variables for satisfaction and gross-income into the basket.

2 Use the "Basket action" "Item and Frequencies" select the corresponding variables names for the years 2006 to 2008 into the basket. Note, that you did not find a variable for tenure which is a so-called generated variable.

3 Use the "Varname" facility in the "Main actions" menu. You should expect generated variables in a file ending with "PGEN" (which stands for generated variables at person level). Browse the list and select "length of work with firm". Select again the variable names for the other waves into the basket.

4 Save the content of your basket into a txt-file for later use.

5 Use the SAS program generator facility under the basket actions for all variables in the basket. In the dialog program select:

- Balanced panel
- Subsample F (2000)
- Private Households
- Both sexes

6 Copy the SAS program from the SOEPINFO to the SAS-Editor window.

- You have to adapt the SAS-library settings to your environment (One library where the SOEP data are and the library (named "library") where the format file is located.

- Check the generated SAS file. There are more variables than you asked for!
- What is the meaning of the negative values -1 and -2?

Data editing before analysis

The data set New generated from SOEPinfo is a raw file not suited for direct panel analysis.

7 Rename the variables to time-indexed variables, for example satisf_1 to satis_4 for the satisfaction scores. You will find the file with the basket information on the original variable names useful.

8 Use Proc Panel to transform the flat-file into the long format.

9 Delete for each point in time the observations with missing values. Note that the SOEP uses its own missing values.

10 Proc Panel expects for each unit at least two observations. Otherwise the procedure stops with an error message. It is up to the user to clear the data base from units with only one observation. One way to manage this task is to use the FIRST and LAST dummy variables that are generated by the ”SET BY” statement. If the first record of a unit is also the last record the unit should be skipped.

Now you are in a position to run a panel analysis!

Do it yourself

You are interested in the commuting behaviour and the satisfaction with leisure time at the individual level. You find the SOEP data for 2010 (File BAP) and 2011 (File BBP) in the SAS-Cloud data set. First you have to retrieve the information about the variable names from SOEPinfo. For this exercise you need only these names, not the SAS-code generated from this platform.

- Start with the questionnaire for file BAP, to see the the approximate position of the questions in the questionnaire. Then open the variable search window and browse the variable names. (Hint: Look for question 1 and 44).
- Create a SAS-Data set that is ready for the use of Proc Panel.
- Produce a Spaghetti plot of Km commuted over time.

Linear Panel Analysis

Use data set Analyse. You want to regress the tenure on satisfaction and income.

11 Use Proc Panel to estimate the FE and RE Effects model. Use also the Pooled statement.

12 What is the result of the Within-estimator?

13 What is the result of the Hausman test?

14 Use time dummies in the RE model by a minor modification in Proc Panel.

15 Use the data set human_cap to estimate a standard human capital model with logearnings as dependent variable regressed on education years, marital status, experience experience squared. The data set is in the long format and has approximately 320 000 observations. Increase the number of observations that enter the analysis from 1 000 to 10 000 to 100 000. Compare the parameter estimates for the within -, the FGLM - and the pooled estimator.

Mixed Models

Use the simulation program generate_panel which is similar to the program for the generation of the spaghetti plots in the course materials.

16 Vary the program so that covariate X_i is correlated with α_i. Use the RE-Model and compare the FGLM- and the Within estimator. What about the result from the pooled estimation. What happens if we increase the number of units or the number of points in time.

17 Use Proc Mixed to specify a model where the intercept and the coefficient of X is regarded as random with unspecified covariance.

Graphical Modeling

18 **Conditional independence:** Suppose we have three variables A, B and C such that $A \otimes B|C$. Derive from the definition of conditional independence the Loglinear representation of the joint contingency table.

19 **Sex before marriage and out of marriage (extramarital sex, infidelity, side step)** and the consequences (divorce or not divorce) is the focus of this exercise. Also displayed is the variable sex (gender) of the corresponding persons. Below you find the SAS command lines to read in this 4 dim contingency table. (Note, these data are somewhat outdated, as the great majority supposes to have no sexual experience before marriage.)

```
  DATA Divorce;
  DO mstatus='divorced', 'married';
    DO sex='male', 'female';
      DO Sex_b='Yes  ', 'No';
        DO Sex_o='Yes  ', 'No';
        INPUT number @@;   OUTPUT;
  END; END; END; END;
  LABEL mstatus='Marital*Status'  sex='Gender'
        Sex_b='Sex before'
        Sex_o='Sex extramarital'; * infidelity, Seitensprung;
  CARDS;
 28  60  17  68  17  54  36 214
 11  42   4 130   4  25   4 322
;
```

20 Use Proc CATMOD to check whether marital status is independent from sex

21 Is sex before marriage independent from extramarital sex?

22 Is extramarital sex independent from gender (sex) ?

23 Find a graphical model for the 3 variables sex, sex_b and sex_o. The model must fit the data! Write the graph of the model and give an interpretation in terms of conditional independence.

24 Find a graphical model for all 4 variables mstatus and the three sex-variables. The model must fit the data. Write the graph of the model and give an interpretation in terms of conditional independence. Give a substantial interpretation of the estimated highest interaction term.

25 **Markov chains**: Suppose you have 4 panel waves and let S_t $(t = 1, 2, 3, 4)$ indicate the state at wave t. A second order Markov chain is characterized by the fact that the distribution of S_t depends on the last **two** preceding states. Write the corresponding graphical model. What is the Loglinear representation?

26 On page 52 of the course materials you find a latent Markov model. Suppose one could observe the latent states what would be the Loglinear representation of this model?

Survival Analysis

The data set "hmohiv" contains the survival times (in month) of 100 HIV-patients in the early 90-s. (Survival chances have improved substantially since that time!) There are two additional variables "age" at the beginning of the record and whether the patient is a drug user. There is also an censoring indicator "Censor" with 0 indicating a right censored observation.

27 Create a Kaplan-Meier plot that compares the survival functions of the drug users with other patients. Did all patients die? What can be said about the longest lifetimes of the deceased patients?

28 What is the effect of age and drug use on the baseline hazard? Is the proportional hazard assumption justified?

Use of LEM

29 The following LEM program commands refer to the estimation of a NMAR model on page 148 of the course materials. R refers to residential mobility which makes the labour force status at time 2 a missing value.

```
res 1                * No. response variables
man 2                * No. of manifest variables
dim 2 3 3            * No. of values of resp. + manifest vars
lab R A B            * Labels of resp. manifest vars.
sub AB A             * Observed tables
mod A B|A  {AB} R|AB {RB} *  Models for tables. Here: R depends only on B
dat [4221 308 358 233 181 208 313 55 1113 * Table AB
    2278 294 558]                         * Table A
```

Start the LEM program. What are the estimated values of $P(B|A)$ for the transitions between labour force states? Is the impact of labour force state B significant? Which part of the corresponding Loglinear model is tested?

30 You want to put other restrictions on $P(R = 1|A = a, B = b)$:

$$(P(R = 1|A = a, B = b)) = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

This amounts to collapsing the groups with medium and low mobility. How does the LEM Program have to be modified? How change the estimates of the $P(B|A)$? Was it a good idea to collapse the two groups?

31 How would the `mod` statement look like if we would assume MCAR ("Missing completely at random")?

Design based estimation

Use the data set mysas.universe as a statistical universe. It contains the 2000–data of the SOEP cross-national file data with 13119 persons.

32 Select a Simple Random Sample (SRS) without replacement from this universe. The sample size shall be 1000. Use Proc Surveyselect.

33 Use Proc Means to estimate the total of the annual earnings (variable earnings) from the sample by the HT-estimator.

34 You want to use some standard calibration setting: Calibration with respect to age*sex groups and marital status. For this purpose compute the year of birth and then birth cohorts brackets of length 10 years starting from 1930. Use Proc Means to calculate the totals. With respect to family status you have to consider that the states 5 (separated) and 6 (exceptional group) are too seldom to form an own group. They are collapsed with state 4 (divored).

35 Use Proc Surveyreg to compute the regression of earnings on the age*sex groups and the marital status groups. Use the "estimate" statement to compute $t'_x\hat{\mathrm{B}}$ and use the output statement for the to collect the residuals of the sample.

36 Estimate the HT-estimator of the residuals and its variance. Now you are in the position to compute the GREG and its variance.

37 Compare the HT- and the GREG-estimate with the population value. Compare also the standard deviation of the two estimates. What is the reason for the small gain from the use of the GREG in this case?