Examination Leaflet on "Short course: Panel surveys in social and economic research and the treatment of nonresponse", 24 Feb – 27 Feb 2014 4 ECTS

Prof. Dr. Ulrich Rendtel

Date of examination: Wednesday 12 March 2014

Questions + Answers!

Fill in your name:
Fill in your enrollment number:

Use the space between the questions for your answers. If necessary use the backside of the sheet.
This test consists of 5 blocks with different items. Select 4 out of the 5 blocks. In each block 6 points can be gained. There is a maximum of 24 points. The test is passed if 12 points are earned. The exercises with a star count twice.

1 Block on general questions about panel design

1 * One alternative to the panel approach to gain longitudinal information is the use of a retrospective questionnaire. An apparent difficulty is the recall burden. Ignore this difficulty for a moment. Suppose you want to gain information about health risks. For that reason you collect a sample of persons at the age of 50 and ask them about smoking habit, drinking behaviour, etc.. Why are the conclusions from such a cross-sectional sample with respect to health risks misleading?

Answer: Because the persons that already died are not included.

2 List 3 arguments for running a rotation panel instead of a permanent panel!

Answer: increasing respondent burden, cumulated panel attrition, only interest in short histories, for example poverty over 4 years.

3 Tracing of residential movers in a panel may turn out a demanding task. Describe several strategies to cope with this difficulty.

Answer: Follow-up by telephone; Contact letters to panel households; information by neighbours; Postal service on new address;

4 Incentives and letters to motivate panel members for further participation is a difficult task. Why? Discuss some pro's and con's:

Answer: Motivation letters may not reach the household in the case of a residential move, the letter may mention some facts the respondent likes/dislikes. Financial incentives may have a selective effect. Incentives like payments cannot be stopped in later panel waves.

5 In a household panel certain households are over-sampled. Which are these households?

Answer: Households with fusions: persons from the existing population move into other households after the start of the panel. Newborns and immigrants do NOT count here.

2 Block on Linear Panel Models and SAS

1 Describe two different analysis settings different from the examples given in the course: In the first case a fixed effects model is appropriate while in a second example the random effects model should not be applied. Justify your examples.

Answer: The fixed effects model applies for a country panel, for example, the units are the EU member states. Justification: Results are conditional on the member states.

A random effects model specifies the covariance of the residual terms in a special way. If this specification is does not apply the random effect model should not be used. For example, there may be a declining correlation within the units over time in contrast to the assumption of the constant correlation of the random effects model.

2 The variable `marital_status` can be used as a covariate in human capital models. The effect of this covariate has somewhat different interpretations when it is estimated by an Within-estimator or by an FGLS-estimator. Explain the difference:

Answer: With the within-estimator we measure the effect of unmarried persons becoming married persons. The FGLS estimates the averaged income differential between married and un-married persons in the population, irrespective whether they had their marriage in the time period of interest or not.

3 Consider a panel on the school performance of pupils over time. A random effects model might be regarded here as to simple. How could such a view be supported? Describe two possible extensions of the random effects model that reflect your arguments.

Answer: The random effects model assumes that the residuals of different pupils are independent over time. This may be regarded as invalid. The residuals of pupils with the same teacher may be correlated, or the scores within the same class correlate. If different schools are involved, the residuals of pupils from the same school may be correlated.

4 Suppose you have an insurance panel, i.e. you count the number of car accidents per contract over subsequent years. In this case the Normal distribution for the error terms will not be a reasonable assumption.

However, there is a straightforward extension of the linear model. Formulate the model:

Answer: Let $Y_{i,t}$ be the number of accidents of contract i in year t. The distribution of the $Y_{i,t}$ is Poisson with expectation $\mu_{i,t}$. A standard parametrisation of the poisson distribution uses the *log*-function as a link between the $\mu_{i,t}$ and a systematic term $X'_{i,t}\beta$. A straightforward extension is given by using a normal distribution for the intercept of the model, resulting in : $\log(\mu_{i,t}) = \beta_{0,i} + X'_{i,t}\beta$

5 * Transformation of a long file into a flat file by a SAS data step.

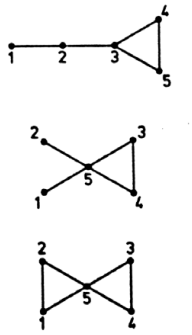
Suppose you have a SAS file `long` with the variables `id` as unit identifier, `wave` as wave identifier, `Y` as the value of the dependent variable and `X` as the covariate. Suppose you have an unbalanced panel with 3 waves. You want to create a file `flat` with one row per person and variables Y_1, Y_2, Y_3 and X_1, X_2, X_3. This can be achieved in one SAS data step. Write the SAS code that manages this task. (**Hint:** Use the `Output` statement and the `First` and the `Last` variables.

Answer:

```
  data flat;
  set long; by id;
  if first.id =1 then do; Y_1=.; Y_2=.; Y_3=.; X_1=.;X_2=.;X_3=.; end;
    if wave=1 then do;
      Y_1=y; x_1=x;
end;
    if wave=2 then do;
      Y_2=y; X_2=x;
end;
    if wave=3 then do;
      Y_3=y; x_3=x;
end;
   If last.id =1 then output;
   Retain Y_1 Y_2 Y_3 X_1 X_2 X_3;
```

3 Block on Discrete Panel Models

The following figure shows three graphs. Each graph connects 5 variables, which are named $1, 2, 3, 4, 5$. Each of the graphs represents a graphical model.



1 * What are the cliques of the graphs in the figure above?

Answer: Figure above:

$$\{1, 2\}, \{2, 3\}, \{3, 4, 5\}$$

Figure in the middle:

$$\{1, 5\}, \{2, 5\}, \{3, 4, 5\}$$

Figure below:

$$\{1, 2, 5\}, \{3, 4, 5\}$$

2 Give an interpretation of the graphs in terms of conditional independence (Use the symbol $\otimes$ to indicate independence.)

Answer: Figure above

$$1 \otimes (3, 4, 5)|2$$

$$(1, 2) \otimes (4, 5)|3$$

Figure in the middle

$$(1 \otimes 2 \otimes (3, 4)|5)$$

Figure below

$$((1, 2) \otimes (3, 4)|5)$$

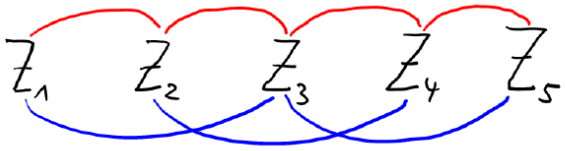
3 What is the algebraic representation of the model at the bottom of the figure in terms of interactions?

Answer:

$$\begin{aligned} & 1*2*5+1*5+1*2+2*5+1+2+5 \\ + \quad & 3*4*5+3*5+3*4+4*5+3+4 \end{aligned}$$

4 * Let $Z_t \quad (t = 1, \ldots T)$ represent the states of a unit over time. The sequence is said to follow a Markov chain of **second** order if $P(Z_t|Z_{t-1}, Z_{t-2}, \ldots, Z_1) = P(Z_t|Z_{t-1}, Z_{t-2})$. This model has a simple graphical representation. Write the corresponding graph and the algebraic representation of this model for a panel with 5 waves.

Answer:



4 Block on Survial Analysis

The treatment of right-censoring in the likelihood approach uses the following likelihood contributions L_i of spell i:

$$L_i = \begin{cases} f_T(t_i), & \text{if spell i is not censored, i.e.: } C_i = 0 \\ P(T_i \geq t_i), & \text{if spell i is censored, i.e.: } C_i = 1 \end{cases}$$

where $f_T(t)$ is the density function of T.

1 * The above likelihood uses an implicit assumption about censoring C_i and the duration of the spell. Can you describe this assumption in terms of a simple formula?

Answer:

$$P(T_i \geq t_i | C_i = 1) = P(T_i \geq t_i | C_i = 0) = P(T_i \geq t_i)$$

2 What does this condition mean substantially?

Answer: The condition means that the survival time is independent from censoring.

3 Give a real life example, where this condition is violated.

Answer: This assumption is violated, for example, in the case of therapy studies where the candidates don't want to show that their therapy is no longer successful. In this case the remaining time after censoring at time t_i is different from $P(T_i \geq t_i)$.

4 Suppose you observe the survival times $5, 6, 7, 8, 8, 9^+, 10, 15$, where $+$ stands for a right censored observation. Compute the Kaplan Meier estimate of the survival function.

Answer:

t	5	6	7	8	9	10	15
$\hat{S}(t)$	0.875	0.75	0.625	0.375	0.375	0.1875	0.0

5 What is the lowest value which is computed by the Kaplan Meier estimator? What is its value in the case of a sample with no censoring?

Answer: The lowest value is computed for the largest observed point in time with no censoring. In the example, it is $\hat{S}(15) = 0.0$

In the case of no censoring the lowest value of $\hat{S}(t)$ is equal to $\hat{S}(t) = 0$

5 Block on Nonresponse

1 Describe the difference between **Missing at Random** and **Missing on observables** by an example, which is different from the one in the course.

Answer: Suppose we want to regress y on a covariate vector X. There are some observed variables z which are excluded from the regression model. Suppose that x and z are always observed and the observability of y is described by $P(R|y, x, z)$. Then

$$P(R|y, x, z) = \begin{cases} P(R|x), & \text{Missing at Random;} \\ P(R|z), & \text{Missing on Observables.} \end{cases}$$

Example: y is income, x is education, z is level of neighborhood.

2 Consider a regression model $y_{i,t} = X_i'\beta_t + \epsilon_{i,t}$. How can one characterize MAR between wave one and wave 2 in terms of a sampling procedure?

Answer: If R_i is the attrition indicator between wave one and two, then MAR is equivalent to

$$P(R_i = 1|y_{i,2}, X_i) = P(R_i = 1|X_i)$$

The sampling procedure that generates the respondent sample is equivalent to stratified sampling within groups of persons with equal covariate values.

3 In the above setting: Describe a selection process, which is **not** MAR. Give a practical example where such a process is reasonable.

Answer: If attrition depends intrinsically on $Y_{i,2}$:

$$P(R_i = 1|y_{i,2}, X_i) = P(R_i = 1|y_{i,2})$$

If $Y_{i,t}$ is income at wave t, it may be reasonable to assume that attrition is lower for higher incomes if the survey is mainly concerned with income (because the questionnaire is more interesting for high income people).

4 * Recall the Markov model for the Fade-Away effect of the initial nonresponse bias in a panel survey. Under what condition on the transition matrix $\mathbf{P}$ the nonresponse bias would disappear after **one** wave?

Answer: If all rows of $\mathbf{P}$ are equal, the distribution on the state space does not depend on the state at time one. The distribution at wave 2 is the steady state distribution of $\mathbf{P}$.

5 Recall the use of the pattern mixture model which is implicitly used in comparisons of an estimate of the first panel wave with the full sample and an estimate with a sample, which is restricted to units with permanent panel participation. There are some situations where the power of this test is zero! Describe such a scenario.

Answer: Consider the case of a regression model $y_{i,t} = X_i'\beta_I + \epsilon_{i,t}$ for population I and $y_{i,t} = X_i'\beta_{II} + \epsilon_{i,t}$ for population II. If attrition depends on temporal changes of $y_{i,1} - y_{i,2}$ then the levels of $Y_{i,1}$ at wave one remain unchanged, also the relationship between $y_{i,1}$ and $X_{i,1}$ remains unchanged in the restricted sample. Hence the comparison of the estimation with full sample and the restricted sample will indicate no difference.

You have reached the end of this test!