Exercise 2

Applied Longitudinal Data Analysis

Deadline: Please upload your assignment by Monday (February, 19), 5 p.m. Upload **one** file only (.pdf). Include the R-code into the Appendix. Cite data, readings and preferably also R-packages. Label and number all figures and tables.

Exercise 2.1 (Preparation)

a) Provide a short text that explains how you arrived at the sample size for your analytical sample. For the construction of your sample, proceed as follows:
   - Select a country and outcome of your choice (home leaving, first birth, first job etc.).
   - Reduce the sample to either male or female respondents.
   - Reduce the sample to the cohorts 1940-1999.
   - Apply listwise deletion.

b) Construct the dependent variable (EVENT, TIME). How many events will enter your analysis?

c) Construct the independent variables (COHORT and EDU). As to EDU: Classify this variable into meaningful categories (either two or three categories). Take into account the sample sizes but also aspects of content. Briefly describe how you constructed the variable and justify your categorization.

d) Provide the sample statistics for education (EDU) by birth cohort and briefly describe patterns.

Exercise 2.2 (Log Rank)

a) Do you expect any differences in the transition to your event of interest by level of education? Do you expect any differences across birth cohorts? Do you assume that educational differences have narrowed across cohorts? Formulate testable hypotheses. Buttress your hypotheses. Cite at least one reading. This may be from the readings on moodle (see lecture 1 and 2). You may also draw on other sources. If you draw on other readings, make sure that the paper is from a credible source, such as a peer-reviewed journal.

b) Test your hypotheses. Estimate the survival functions and apply a Log-Rank-test.

Exercise 2.3 (Discussion)

How do you evaluate the patterns that you found in the data. Do you see any need for policy intervention? (max 250 words)

Exercise 2.4 (Limitation)

Omitted variable bias is an omnipresent problem. Name one variable that may have biased your results. Explain why this variable would have been particularly important to include for your country and outcome of choice. How are the results affected because you failed to account for this variable? (max 300 words)

Appendix

Name	Realization	Class	Description
gndr	1 Male 2 Female NA Refusal	haven	Gender
yrbrn	1928 … 1999 NA Refusal	haven	Year of birth
lnwyys	2018 2019	haven	Year of interview
eisced	0 Not possible to harmonise 1 ISCED I , less than lower secondary 2 ISCED II, lower secondary 3 ISCED IIIb, lower tier upper sec. 4 ISCED IIIa, upper tier upper sec. 5 ISCED IV, advanced vocationa 6 ISCED V1, lower tertiary/ BA level 7 ISCED V2, higher tertiary., >= MA 55 Other	haven	Highest level of education, ISCED
cntry	AT Austria BE Belgium BG Bulgaria CH Switzerland CY Cyprus CZ Czech Republic DE Germany EE Estonia FI Finland FR France GB Great Britain HU Hungary IE Ireland IT Italy NL Netherlands NO Norway PL Poland RS Serbia SE Sweden SI Slovenia	haven	Country
lvpntyr	0 still in parental home 1111 Never lived with a parent 1943 … 2019 NA Refusal	haven	Year left parental home
evmar	1 yes 2 no	haven	Ever married?
maryr	1938 … 2019	haven	Year first marriage
evlvpt	1 yes 2 no	haven	Ever lived with partner?
lvptnyr	1938 … 2019	haven	Year started stared living with a partner
evpdemp	1 yes 2 no	haven	Ever had paid job
pdempyr	1938 … 2019	haven	Year first job
bthcld	1 yes 2 no NA Refusal	haven	Ever had a child
fcldbrn		haven	Year first child was born

New variables

Name	Realization	Class	Description
TIME		numeric	Age at event for cases with event Age at censoring for censored cases
EVENT	0 1	numeric	Variable that equals 1 for cases with event and 0 otherwise
COHORT	1940-1969 1970-1999	factor	Year of birth (grouped)
EDU	TBD	factor	Level of education