University of Helsinki
Department of Social Research

Topics in Survey Methodology and Survey Analysis, fall 2012

PERSONAL HOMEWORK ASSIGNMENT

Credits: 2 ECTS credits

Final product: Written report (10-15 pages plus selected annexes)
Tentative structure:
Title page (title, author, course, department, date, student id number)
Text part (divided into suitable sections and subsections)
References (literature)
Annexes (extracts from output, selected pieces of program code)

Delivery of final product by **5 November 2012** (in PDF format) as email attachment:
risto.lehtonen@helsinki.fi
or as paper copy:
Risto Lehtonen
University of Helsinki, Department of Social Research
P.O. Box 68 (Gustaf Hällströmin katu 2b), 00014 Helsingin yliopisto

The OHC data set provides the empirical data set (download SAS or SPSS version from course webpage).
SAS (Version 9.2 or 9.3) or SPSS (PASW Statistics 18 or 20) can be used in the analysis.

Exercise 1. Properties of the sampling design of the OHC data set

Describe the technical properties of the sampling design underlying the OCH data set. What properties of the sampling design should be taken into account for proper statistical inference when analyzing the OHC data set? Why? What happens if these properties are ignored?

Exercise 2. Exploratory data analysis and comparison of approaches

a) Let us consider the subject matter variables in the OHC data set. Describe the types of variables (qualitative, binary, continuous…). Produce descriptive statistics (point estimates and standard error estimates, selected frequency tables) for the variables, by taking into account the complexities (stratification, clustering) of the sampling design.
b) Carry out a similar analysis as in point a by assuming that the data arises from a SRS design.
c) Compare the results of points a and b. Give explanation of possible differences and draw conclusions.

Exercise 3. Logistic ANCOVA

a) Select a binary study variable and a set of explanatory variables from the list of subject matter variables in the OHC data set. Select a proper approach (design-based, model-based) for model fitting. Give motivation for your choice. Fit a logistic ANCOVA model by taking into account the complexities of the sampling design or the hierarchical (multilevel) structure of the data, under the chosen approach. Please also consider interaction terms in model fitting. Report results on estimated logistic regression coefficients (point estimates, standard error estimates, t test statistics and p-values). Calculate odds ratio estimates and their standard error estimates and report the results. Give interpretation of results. Draw conclusions.
b) Carry out a similar analysis as in point a by assuming that the data arises from a SRS design.
c) Compare the results of points a and b. Give explanation of possible differences and draw conclusions.

You are encouraged to consult the VLISS application at http://mathstat.helsinki.fi/VLISS/ for help.

Contents of the SAS data set OHC

The CONTENTS Procedure

Data Set Name	A.OHC	**Observations**	7841
Member Type	DATA	**Variables**	12
Engine	V9	**Indexes**	0
Created	20. syyskuuta 2011 tiistai 11:30:19	**Observation Length**	96
Last Modified	20. syyskuuta 2011 tiistai 11:30:19	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Variables in Creation Order

#	Variable	Type	Len	Label
1	ID	Num	8	Element identifier
2	STRATUM	Num	8	Stratum identifier
3	SEX	Num	8	Gender
4	AGE	Num	8	Age in years
5	AGE2	Num	8	Age under/over 45
6	PHYS	Num	8	Physical health hazards of work
7	CHRON	Num	8	Chronic morbidity
8	PSYCH	Num	8	Psychic strain - 1st princomp
9	PSYCH2	Num	8	Psychic strain - dichotomy
10	PSU	Num	8	Primary sampling unit (Cluster)
11	wd	Num	8	Design weight
12	wa	Num	8	Rescaled weight (Analysis weight)