

Topics in Social Statistics, fall 2010

PERSONAL HOMEWORK ASSIGNMENT

Credits: 2 cu

Final product: Written report (some 10 pages plus annexes)

Tentative structure:

Title page (title, author, course, department, date)

Text part (divided into suitable sections and subsections)

References (literature)

Annexes (extracts from output, selected pieces of program code)

Delivery of final product by **30 October 2010** (in PDF format) as email attachment:

risto.lehtonen@helsinki.fi

or as paper copy:

Risto Lehtonen

Helsingin yliopisto, Matematiikan ja tilastotieteen laitos

PL 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 9.1.3 or 9.2) or SPSS 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? 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. Fit a logistic ANCOVA model by taking into account the complexities of the sampling design. Please also consider the interaction terms. 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 can consult the VLISS application for help.