UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Exam in: STK4600 Statistical methods for social sciences.
Day of exam: Thursday June 4, 2015
Exam hours: 14.30 – 18.30
This examination paper consists of 3 pages.
Appendices: None
Permitted materials: Lecture notes, student's own notes and approved calculator

Make sure that your copy of this examination paper is complete before answering.

Exercise 1

In a city of 72500 people, a simple random sample of four households is selected from the 25000 households in the population to estimate the average cost on food per household for a week, denoted by μ. The first household in the sample had 4 people and spent a total of kr. 1800 in food that week. The second household had 2 people and spent kr. 1200. The third, with 4 people, spent kr. 2400. The fourth, with 3 people, spent kr. 1680.

a) The variable of interest, y, is the cost of food for a week for a household. Identify the sampling units and any auxiliary information associated with the units.

b) Consider the sample mean based estimator of μ. Derive the estimate of μ and find the standard error of the estimator.

c) Describe an estimator that uses the auxiliary information and derive the estimate. Find the standard error of the estimator.

d) Based on the results in parts b) and c), which estimator is to be preferred? Give a reason for your choice.

Exercise 2

Assume we take an election survey after a national election, to estimate the voting participation. Of a random sample of 1000, only 500 responded in the survey. The response rates for the 3 age groups 18-29, 30- 60, and 61+, were 30, 61 and 52.5. The data were:

Age	Total sample	Response sample	The number of voters in the response sample	Total in population
18-29	300	90	60	2 500 000
30-60	500	305	280	4 000 000
61 and over	200	105	80	1 500 000

a) Assume the nonresponse is MCAR and consider the sample mean based estimator of the proportion of voters in the population. Compute the estimate and a 95% confidence interval for the proportion of voters in the population. The true voting proportion is 78 percent. What can you say about the MCAR assumption?

b) To correct for the bias in part a) we shall poststratify according to the age groups. Find the poststratified estimate for the voting proportion in the population.

Continued on page 2

c) Derive the standard error of the poststratified estimator and compute a 95% confidence interval.

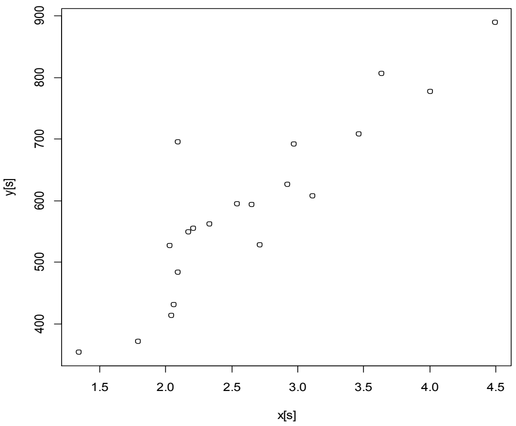
d) Under what condition is the poststratified estimator unbiased? Is that the case here? If not, how would you attack this problem?

Exercise 3

Consider the population of 6194 California schools with Academic Performance Index (API) computed for 1999 and 2000. We shall estimate the average API for 1999 for the population using as auxiliary variable x = average parental education level (avg.ed). A simple random sample s of 20 schools were selected with the following results.

y= api99	x = avg.ed
595	2.54
354	1.34
431	2.06
695	2.09
608	3.11
555	2.21
484	2.09
414	2.04
528	2.71
549	2.17
527	2.03
806	3.63
594	2.65
890	4.49
627	2.92
692	2.97
372	1.79
562	2.33
708	3.46
777	4.00

A scatter plot of the data:



Continued on page 3

Final exam in STK4600

We shall consider the following two possible population models:

Model 1: $E(Y_i) = \beta x_i$ and $Var(Y_i) = \sigma^2 x_i$. The Y_i's are uncorrelated.
Model 2: $E(Y_i) = \beta_1 + \beta_2 x_i$ and $Var(Y_i) = \sigma^2$. The Y_i's are uncorrelated.

We have the following values of various statistics that you can use in answering the questions in this exercise:

- Mean of x in the sample is 2.6315, mean of x in the population is 2.7935
- The estimated value of regression coefficient $\beta_2$is equal to 162.5
- The estimated σ^2 in model 1 is equal to 2571.6
- The estimated σ^2 in model 2 is equal to 4209.3
- The sample variance of x equals 0.628
- The sample variance of y equals 20572.5

a) Looking at the scatter plot, which of the two models do you think describes the data best?
b) Estimate the mean api99 based on model 1. Derive the model-based standard error and an approximate 95% model-based confidence interval.
c) Estimate the mean api99 based on model 2. Derive the model-based standard error and an approximate 95% model-based confidence interval.
d) Compare the estimates and standard errors in a) and b) with the sample-mean based estimate and its model-based standard error. What can you conclude regarding the value of including x as an auxiliary variable in the estimation?
e) The true value of mean api99 is 632. Try to give an explanation on why the three estimated values differ in distance from the true value.