Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 75/28 22 August 1975

DESIGN AND ESTIMATION PROBLEMS WHEN ESTIMATING A REGRESSION COEFFICIENT FROM SURVEY DATA

By

Ib Thomsen

CONTENTS

		Page
	Abstract	2
1.	Introduction	3
2.	Non-existence of a minimum variance unbiased, linear estimator of a	4
3.	Optimal strategy in a subclass of unbiased linear estimators	9
4.	Estimation of b	11
5.	Estimation of σ^2, var($\hat{a}_S^*$), and var($\hat{b}_S^*$)	12
6.	Acknowledgements	13
7.	References	13

ABSTRACT

The values of a variable x are assumed known for all elements in a finite population. Between this variable and another variable Y, whose values are registered in a sample survey, there is the usual linear regression relationship, viz.,

$$E(Y_i|x_i) = ax_i + b$$

with $\text{var}(Y_i|x_i) = \sigma^2$ and $\text{cov}(Y_i, Y_j|\underset{\sim}{x}) = 0 \quad (i,j = 1,2,-,N)$.

This paper considers problems of design and of estimation of a and b. The following Godambe type theorem is proved: There exists no minimum variance unbiased linear estimator of a and b. We also prove that the usual estimators of a and b have minimum variance if attention is restricted to the class of linear estimators unbiased in any given sample.

1. INTRODUCTION

Assume that we have a finite population of N distinguishable elements labelled by the integers $1,2,\ldots,N$, with the associated values $(x_1,Y_1),\ldots(x_N,Y_N)$. Furthermore, we shall assume that

(i) $x_1,\ldots,x_N$ are given numbers,

(ii) $U_1,\ldots,U_N$ are uncorrelated random variables with $E(U_i) = 0$, $\text{var}(U_i) = \sigma^2$,

(iii) $Y_i = ax_i + b + U_i$, where a and b are unknown constants.

Our aims are to make a sample design and to estimate a, b, and σ^2. Conditioned on a given set of values of x_i $(i = 1,2,\ldots,N)$ we shall look for linear, unbiased estimates of a, b which over repeated samples from the finite population give the least variance for all possible values of a, b, and $\sigma > 0$.

Under certain assumptions about the design we shall prove that there exist no minimum variance unbiased, linear estimate of the regression coefficient. In Section 3 we shall restrict the class of estimates and show that within this restricted class the usual estimate of the regression coefficient has minimum variance.

Recently the problems involved in estimating the finite population regression coefficient from a sample have been discussed in Kish (1970) and Frankel (1971). In the present paper we are concerned with estimating parameters in an assumed structure, rather than estimating parameters in a specific population.

Our model we think is useful when estimating a behavioral econometric model in which the endogeneous variable, Y, is considered to result from a generating process comparable to a chance mechanism. Haavelmo (1944). In textbooks in econometrics the discussion are concentrated on finding "good" estimates of a, b, and σ^2 for a given set of values of x_i in the sample. Klein (1963), Malinvaud (1968), Theil (1971). In these textbooks it is proven that conditioned on a given set of values of x_i in the sample the usual estimates are Markov-estimates. In this paper, however, we want to take into account the variation due to sampling. We have several reasons for doing this, among others:

1. In addition to "good" estimates of a, b, and σ^2 we want to find "good" sampling designs for estimating a and b.

2. In Section 2 below is given an estimate of a, which may have a smaller variance than the usual regression estimate, when the sampling variation is taken into account.

3. On the given assumptions the sample is an ancillary statistic and this brings us up against the question of whether the inference should be made conditional given the sample, or unconditionally, i.e., over repeated samples from the finite population. Although many implicitly seem to prefer conditional inference on grounds of common sense, there exists to our knowledge no theory justifying this position.

2. NON-EXISTENCE OF A MINIMUM VARIANCE UNBIASED, LINEAR ESTIMATOR OF a

A sample of size n is defined quite generally to be an ordered sequence of n of the population labels $i_1,i_2,\ldots,i_n$ (repetitions allowed) together with the sequence of their associated observed characteristic values

$$(x,y) = (x_{i_1},y_{i_1}),(x_{i_2},y_{i_2}),\ldots,(x_{i_n},y_{i_n}).$$

Let

y_i denote the realization of Y_i in the population, and

let
$$\underset{\sim}{x} = (x_1,\ldots,x_N),$$
$$\underset{\sim}{Y} = (Y_1,\ldots,Y_N),$$
$$\underset{\sim}{y} = (y_1,y_2,\ldots,y_N).$$

A sample design is then defined by some finite set $\mathcal{S}$ of ordered sequences, s, together with a probability measure assigned by choosing a function $p(s) > 0$, $\Sigma p(s) = 1$, where p(s) is the probability of choosing the sample s. The probability that label i is included in the sample at least once we shall denote π_i. Define a stochastic variable, S, with $p(S=s) = p(s)$, $s \in \mathcal{S}$ such that S and $\underset{\sim}{Y}$ are independent.

As estimates of a and b we shall consider the general class of linear estimates defined in Horvitz and Thompson (1952) and Godambe (1955)

$$\hat{a}_s(\underset{\sim}{y}) = \sum_{\lambda \in s} \beta_{\lambda s} y_\lambda ,$$

where $\beta_{\lambda s}$ can be assigned to each label, λ, and each sequence $s \in \mathcal{S}$ before selection and observation of the sample; thus $\beta_{\lambda s}$ may well depend upon $\underset{\sim}{x}$. ($\sum_{\lambda \in s}$ is the sum over all different labels in s.)

For a given design and a given $\underset{\sim}{x}$ we want $\hat{a}_S(\underset{\sim}{Y})$ to be unbiased, i.e.

$$E(\hat{a}_S(\underset{\sim}{Y})) = \sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s}(ax_\lambda + b)$$

$$= a \sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s} x_\lambda + b \sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s}$$

$$= a.$$

For $\hat{a}_S(\underset{\sim}{Y})$ to be unbiased we must have that

(1) $$\sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s} x_\lambda = 1, \text{ and}$$

(2) $$\sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s} = 0.$$

We also find that

$$E(\hat{a}_S(\underset{\sim}{Y}))^2 = \sigma^2 \sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s}^2 + \sum_{s \in \mathcal{S}} p(s) \{ \sum_{\lambda \in s} \beta_{\lambda s}(ax_\lambda + b)\}^2 ,$$

from which follows that

(3) $$\operatorname{var}(\hat{a}_S(\underset{\sim}{Y})) = \sigma^2 \sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \beta_{\lambda s}^2 + \sum_{s \in \mathcal{S}} p(s) \{ \sum_{\lambda \in s} \beta_{\lambda s}(ax_\lambda + b)\}^2 - a^2.$$

We shall state an assumption and prove that under this assumption there exists no $\{\beta_{\lambda s}\}$ which minimizes (3) under conditions (1) and (2).

Assumption 1

We assume that there exist two samples s_1 and s_2 with a common label and such that

$$\sum_{i \in s_1} (x_i - \tilde{x}) \neq \sum_{i \in s_2} (x_i - \tilde{x}), \quad \text{where}$$

$$\tilde{x} = \sum_{i=1}^{N} \pi_i x_i \Big/ \sum_{i=1}^{N} \pi_i .$$

Theorem 1

Under assumption 1 there exists no $\{\beta_{\lambda s}\}$ such that for any $\sigma > 0$ (3) is minimized for all values of a and b under conditions (1) and (2).

Proof:

Assume that $\{\beta^*_{\lambda s}\}$ minimizes (3) under (1) and (2) for all values of a and b. Then for all λ and all s including λ we have that

$$(4) \qquad \sigma^2 \beta^*_{\lambda s} + \{ \sum_{i \in s} \beta^*_{is} (a x_i + b) \} (a x_\lambda + b) - \mu x_\lambda - \gamma = 0,$$

where μ and γ are lagrangian multipliers. Equations (4) must be satisfied for all admissible values of a and b for any given value of $\sigma > 0$. We shall insert two sets of values for a and b, and show that the two corresponding sets of conditions on $\{\beta^*_{\lambda s}\}$ cannot be fulfilled simultaneously.

(i) a = b = 0, and $\sigma = 1$

Inserting these values into (4) we see that $\beta^*_{\lambda s} = \beta^*_{\lambda}$ for all λ and all s including λ, and it follows that

$$(5) \qquad \beta^*_{\lambda} = \mu x_\lambda + \gamma .$$

As $\beta^*_{\lambda s} = \beta^*_\lambda$ we can rewrite (1) and (2) in the following way:

$$(1^*) \qquad \sum_{i=1}^{N} \pi_i x_i \beta_i^* = 1, \quad \text{and}$$

$$(2^*) \qquad \sum_{i=1}^{N} \pi_i \beta_i^* = 0.$$

From (5), (1^*), and (2^*) we find that

$$\mu = 1 / \{ \sum_{i=1}^{N} \pi_i x_i^2 - (\sum_{i=1}^{N} \pi_i x_i)^2 / \sum_{i=1}^{N} \pi_i \} = 1 / S_x^2 ,$$

and

$$\gamma = - \mu \sum_{i=1}^{N} \pi_i x_i / \sum_{i=1}^{N} \pi_i = - \tilde{x} / S_x^2 .$$

From this, it follows that

$$(6) \qquad \beta_i^* = (x_i - \tilde{x}) / S_x^2 .$$

We now insert a second set of values for a and b.

(ii) $a = 0$, $b = 1$, and $\sigma = 1$.

Inserting these values into (4), and using $\beta^*_{\lambda s} = \beta^*_\lambda$ we have that

$$(7) \qquad \beta_\lambda^* + \{ \sum_{i \in s} \beta_i^* \} - \mu x_\lambda - \gamma = 0$$

for all λ and all s including λ. Now by assumption 1 we can choose two samples s_1 and s_2 such that they have at least one label in common. Then from (7) it follows that

$$(8) \qquad \sum_{i \in s_1} \beta_i^* = \sum_{i \in s_2} \beta_i^* .$$

Inserting (6) in (8) we find that

$$(9) \qquad \sum_{i \in s_1} (x_i - \tilde{x}) = \sum_{i \in s_2} (x_i - \tilde{x})$$

for any two samples in the design with a common label. Applying assumption 1 we see that (9) is not satisfied for at least two samples in the design, and we have proved theorem 1. □
To illustrate the result we shall give an example:

Example 1

For a given design the following estimate of a is unbiased.

$$\hat{a}_S^{**}(\underset{\sim}{Y}) = (p(S)^{-1}/|\mathcal{S}|) \sum_{i \in S} Y_i(x_i - \bar{x}(S))/ \sum_{i \in S} (x_i - \bar{x}(S))^2 ,$$

where $|\mathcal{S}|$ denotes the number of samples in the design, and $\bar{x}(s)$ denotes the sample mean of x. Furthermore, we have that

$$\mathrm{var}(\hat{a}_S^{**}(\underset{\sim}{Y})) = \sigma^2 \sum_{s \in \mathcal{S}} \{p(s) |\mathcal{S}|^2 \sum_{i \in s} (x_i - \bar{x}(s))^2\}^{-1} +$$

$$+ a^2 \{ \sum_{s \in \mathcal{S}} (p(s)|\mathcal{S}|^2)^{-1} - 1\}.$$

Let $\hat{a}_S(\underset{\sim}{Y})$ denote the usual estimate of a. Then we find that

$$\mathrm{var}(\hat{a}_S(\underset{\sim}{Y})) = \sigma^2 \sum_{s \in \mathcal{S}} p(s) / \sum_{i \in s} (x_i - \bar{x}(s))^2.$$

We shall find a population and a design such that $\mathrm{var}(\hat{a}_S^{**}(\underset{\sim}{Y}))$ is not uniformly larger than $\mathrm{var}(\hat{a}_S(\underset{\sim}{Y}))$.

The population consists of three elements with the associated x-values $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. The design consists of two samples, $s_1 = \{1, 2\}$ and $s_2 = \{1, 3\}$, with $p(s_1) = 2/3$ and $p(s_2) = 1/3$. In this case we find that $\mathrm{var}(\hat{a}_S(\underset{\sim}{Y})) = 1.5\,\sigma^2$ and $\mathrm{var}(\hat{a}_S^{**}(\underset{\sim}{Y})) = (9\,\sigma^2 + a^2)/8$. It follows that $\mathrm{var}(\hat{a}_S(\underset{\sim}{Y}))$ is not uniformly smaller than $\mathrm{var}(\hat{a}_S^{**}(\underset{\sim}{Y}))$.

3. OPTIMAL STRATEGY IN A SUBCLASS OF UNBIASED LINEAR ESTIMATORS

A strategy involves two things, the sampling design and the estimation procedure.

One "natural constraint" to put on an estimate of a is that it should be unbiased over the hypothetical population for any given sample, i.e.

$$E(\hat{a}_S(\underset{\sim}{Y}) \mid S = s) = a, \text{ or}$$

$$\sum_{i\in s} \beta_{is} x_i = 1 \text{ and } \sum_{i\in s} \beta_{is} = 0 \tag{10}$$

for all samples in the design. Evidently this class of estimates is a subclass of all unbiased linear estimates. We now minimize (3) under conditions (10).

From (10) it follows that $\Sigma p(s)\{\Sigma\beta_{\lambda s}(ax_\lambda + b)\}^2 = a^2$, and we can therefore minimize $\Sigma p(s)\Sigma\beta_{\lambda s}^2$ under conditions (6), which gives

$$\hat{\beta}_{\lambda S} = \{x_\lambda - \bar{x}(S)\} / \{\sum_{i\in S} x_i^2 - n(S)\,\bar{x}^2(S)\}, \tag{11}$$

where $\bar{x}(S)$ is the mean of all x values in the sample after the removal of duplicates, and $n(S)$ os the number of different labels in the sample.

Remark 1

It should be noted that $\hat{\beta}_{\lambda S}$ is independent of the design (except that we must have $\{\sum_{i\in S} x_i^2 - n(s)\,\bar{x}^2(S)\} > 0$ for all samples in the design). One consequence of great practical interest is that the researcher does not need to know much about the design to calculate $\hat{\beta}_{\lambda S}$.

Define

$$\hat{a}_S^{*} = \sum_{\lambda\in S} \beta_{\lambda S}^{*} Y_\lambda = \frac{\sum_{\lambda\in S} [x_\lambda - \bar{x}(S)][Y_\lambda - \bar{Y}(S)]}{\sum_{i\in S} [x_i - \bar{x}(S)]^2},$$

which is the usual estimate of the regression coefficient.

Then $$\text{var}\,(\hat{a}_S^{*}) = \sigma^2 \sum_{s \in \mathcal{S}} p(s) \Big/ \sum_{i \in s} (x_i - \bar{x}(s))^2 ,$$

which is not independent of the design.

The design which minimizes $\text{var}\,(\hat{a}_S^{*})$ consists of choosing the units which minimize $\sigma^2 \sum_{s \in \mathcal{S}} p(s) \Big/ \sum_{i \in s} (x_i - \bar{x}(s))^2$.

Remark 2

The optimal design is of little practical interest for several reasons. We shall give two:

1. One would almost never design a survey to estimate a single regression coefficient.
2. In economics and social research the linear relationship is an approximation, and the researcher would like to study the residuals over the whole range of X.

Corollary 1

If the sample is a simple random sample of size n, we have:

$$\text{var}(\hat{a}_S^{*}) \geq \frac{1}{n-1} \frac{\sigma^2}{S_x^2} ,$$

where

$$S_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \quad \text{and} \quad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i .$$

Proof:

Applying Jensen's inequality one finds that

$$\sum_{s \in \mathcal{S}} p(s) \left[\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}(s))^2} \right] \geq \frac{\sigma^2}{\sum_{s \in \mathcal{S}} p(s) \left[\sum_{i=1}^{n} (x_i - \bar{x}(s))^2 \right]} = \frac{\sigma^2}{(n-1) S_x^2} . \square$$

4. ESTIMATION OF b

The class of linear estimators is again

$$\hat{b}_S = \sum_{\lambda \in S} \gamma_{\lambda S} Y_\lambda ,$$

where $\gamma_{\lambda S}$ can be assigned to each y_λ and $s \in \mathcal{S}$ before selection and observation of the sample.

For a given design we want $\hat{b}_S$ to be unbiased, i.e.,

$$E(\hat{b}_S) = b,$$

which implies that we must have

$$(12) \qquad \sum_{s \in \mathcal{S}} p(s)\Big[\sum_{\lambda \in s} \gamma_{\lambda s} x_\lambda\Big] = 0, \text{ and}$$

$$(13) \qquad \sum_{s \in \mathcal{S}} p(s)\Big[\sum_{\lambda \in s} \gamma_{\lambda s}\Big] = 1.$$

We also find that

$$(14) \qquad \operatorname{var}(\hat{b}_S) = \sigma^2\Big\{\sum_{s \in \mathcal{S}} p(s) \sum_{\lambda \in s} \gamma_{\lambda s}^2 + \sum_{s \in \mathcal{S}} p(s)\Big\{\sum_{\lambda \in s} \gamma_{\lambda s}(ax_\lambda + b)\Big\}^2 - b^2 .$$

Under the same assumptions on $\mathcal{S}$ and $\underset{\sim}{x}$ as in section 2 we have that there exists no minimum variance linear, unbiased estimate of b.

Again, if we restrict the class of estimates, as in section 3, to the class that satisfies

$$(15) \qquad \sum_{\lambda \in s} \gamma_{\lambda s} x_\lambda = 0, \text{ and}$$

$$(16) \qquad \sum_{\lambda \in s} \gamma_{\lambda s} = 1$$

for all samples in the design, we find that (14) is minimized by

$$\gamma_{\lambda s}^{*} = \Big[\frac{1}{n(s)} - \frac{[x - \bar{x}(s)]}{\sum_{i \in s} [x_i - \bar{x}(s)]^2} \bar{x}(s)\Big] = \frac{1}{n(s)} - \beta_{\lambda s}^{*} \bar{x}(s).$$

It follows from (14) that

$$\operatorname{var}(\hat{b}_S^{*}) = \sigma^2 \sum_{s \in \mathcal{S}} p(s)\Big\{\sum_{\lambda \in s}\Big[\frac{1}{n(s)} - \frac{[x_\lambda - \bar{x}(s)]}{\sum_{i \in s} [x_i - \bar{x}(s)]^2} \bar{x}(s)\Big]^2\Big\} ,$$

where

$$\hat{b}_S^{*} = \sum_{\lambda \in S} \gamma_{\lambda S}^{*} y_\lambda = \bar{y}(S) - \hat{a}_S^{*} \bar{x}(S).$$

Corollary 2

For any design where $\{ \sum_{i \in s} [x_i - \bar{x}(s)]^2 \} > 0$ for all samples in the design, we have that

$$\mathrm{cov}(\hat{a}_S^{*}, \hat{b}_S^{*}) = \sigma^2 \sum_{s \in \mathcal{S}} p(s) \bar{x}(s).$$

5. ESTIMATION OF σ^2, $\mathrm{var}(\hat{a}_S^{*})$ AND $\mathrm{var}(\hat{b}_S^{*})$

Denote

$$Q_{0S} = \sum_{i \in S} (y_i - \hat{a}_S^{*} x_i - \hat{b}_S^{*})^2.$$

Corollary 3

For any design in which $\{ \sum_{i \in s} (x_i - \bar{x}(s))^2 \} > 0$ for all samples in the design, we have that

$$E(Q_{0S}) = (\bar{n} - 2) \sigma^2,$$

where $\bar{n} = \sum_{s \in \mathcal{S}} n(s) p(s)$.

Proof:

Applying $E(Q_{0S}) = \sum_{s \in \mathcal{S}} p(s) E(Q_{0s} \mid S=s)$ we find that

$$E(Q_{0S}) = \sigma^2 \{ \sum_{s \in \mathcal{S}} p(s) n(s) - 2 \} = \sigma^2 (\bar{n} - 2).$$

Corollary 4

For any design in which $\{ \sum_{i \in s} (x_i - \bar{x}(s))^2 \} > 0$ for all samples in the design, we have that

$$E \left[Q_{0s} / \{ (n(s)-2) \sum_{i \in s} (x_i - \bar{x}(s))^2 \} \right] = \mathrm{var}(\hat{a}_S^{*}).$$

Proof:

The proof is similar to the proof of corollary 3.

Moreover, the estimates of $\mathrm{var}(\hat{b}_S^{*})$ are found to be the usual estimates. From corollary 3 and 4 it follows that the researcher does not need to know much about the design to estimate variances of $\hat{a}_S^{*}$ and $\hat{b}_S^{*}$.

6. ACKNOWLEDGEMENTS

Part of this work was done while the author was on leave of absence at Institute for Social Research, Ann Arbor, Michigan. I am grateful to Professor L. Kish and Professor W. Ericson, University of Michigan, for helpful comments and suggestions. The draft was finished after my return to the Central Bureau of Statistics, where I received helpful suggestions from Dr. J.M. Hoem and Mr. Arne Amundsen.

7. REFERENCES

[1] Frankel, M.R. (1971): Inference from Survey Samples. Institute for Social Research, Ann Arbor, Michigan.

[2] Godambe, V.P. (1955): A Unified Theory of Sampling from Finite Populations. J.R.S.S. Ser. B, 269-278.

[3] Haavelmo, T. (1944): The Probability Approach in Econometrics. Econometrica Vol. 12. Supplement.

[4] Horvitz, D.G. and Thompson, D. (1952): A Generalization of Sampling without Replacement from a Finite Population. J. Am. Statist. Ass. 47, 668-670.

[5] Kish, L. and Frankel, M.R. (1970): Balanced Repeated Replication from Standard Errors. J. Am. Statist. Ass. 65, 1071-94.

[6] Klein, L.R. (1963): An Introduction to Econometrics. Prentice-Hall, Inc., N.Y.

[7] Malinvaud, E. (1968): Statistical Methods of Econometrics. North-Holland Publishing Company, Amsterdam.

[8] Theil, H. (1971): Principles of Econometrics. New York: John Wiley and Sons.