MODELS FOR ESTIMATION OF SAMPLING ERRORS[1]

P.D. Ghangurde[2]

This paper presents results of an empirical study on fitting log-linear models to data on estimates of characteristics and their coefficients of variation (CV) from the Canadian Labour Force Survey. The characteristics were classified into groups on the basis of design effects and models were fitted to data on estimates of characteristic totals and their CVs over twelve month period. The models can be used in situations where estimates of CV are needed for new characteristics, and for providing more precise estimates of reliability of estimates based on past data. The problem of evaluation of fit of the models is considered.

1. INTRODUCTION

This paper presents results of an evaluation study on models for estimation of coefficient of variation (CV) of estimates of characteristics based on the Canadian Labour Force Survey (LFS). The LFS is a monthly household survey with a stratified multi-stage area sample design with a sample size of approximately 55,000 households.

Each month estimates of CV are calculated for a set of characteristics using Keyfitz method of variance estimation based on Taylor series approximation [4], [5]. However, computation of appropriate variance estimates for all estimates tabulated from a large scale survey such as the LFS is not possible due to operational constraints of time and

---

[1] Presented at the American Statistical Association Annual Meeting in Detroit, August 1981.

[2] P. D. Ghangurde, Census and Household Survey Methods Division, Statistics Canada.

costs. The model-based estimates of CV can be used to obtain preliminary estimates of reliability for new characteristics based on the past data, and when estimates of CV for an extended period (e.g. one year) are needed. The models can also be used for obtaining concise estimates of reliability, e.g. alphabetic indicators for ranges of CV.

In section 2 the linear and non-linear models used for estimation of totals and proportions are explained. Sections 3 and 4 review considerations made in forming groups, fitting models and evaluation of goodness of fits.

2. THE MODELS

The LFS is a monthly household survey in which dwelling is the final stage sampling unit. Each of the ten provinces in Canada are divided into economic regions which consist of groups of counties with similar economic structure. The economic regions are divided into geographic strata and multi-stage area samples are drawn without replacement with two stages in self-representing strata in the large urban centres and three or four stages in the non-self-representing strata in rural areas. The sample selection in the initial stages is with probability proportional to population size and that in the last stage, in which dwellings are selected from clusters, being systematic.

The design-based estimates within strata are obtained by weighting the data by inverse of probabilities of selection. An adjustment of the basic weight for non-response and ratio estimation within age-sex groups, which are post-strata, is used to obtain final estimates. The census-based population projections for age-sex groups within each province are used as auxiliary variable totals for ratio estimation. More details on the sample design and estimation are given in [5].

The variance estimates of various characteristics at the province level are obtained by Taylor series approximation assuming that the primary sampling units (psus) within non-self-representing strata are selected independently. In self-representing strata the sampled clusters are divided into two groups, which are treated as pseudo-psus and are assumed to have been selected independently. The variance estimate for an estimated characteristic total at Canada level is the sum of corresponding provincial variance estimates [5]. The variance of an estimate $\hat{X}$ of a characteristic total X in a province can also be expressed as

$$V(\hat{X}) = F\ (W-1)\ X\ (1 - \frac{X}{P}), \qquad (1)$$

where P = population for the province,

W = inverse sampling ratio,

F = design effect for the characteristic, and

n = sample size (persons).

The expression (1) for $V(\hat{X})$ relates the variance obtained for the complex ratio estimate based on a stratified multi-stage sample design to the variance of the estimate based on a simple random sample of the same size drawn from the finite population of size P. The sampling variance of an estimate of total based on a simple random sample of size n $(= \frac{P}{W})$ is the usual binomial variance with finite population correction. The term, F, the design effect, represents a factor by which variance is increased due to the effect of such factors as sampling procedure at each stage, the extent of stratification and post-stratification, size of units at various stages and clustering of counts of the characteristic in the province. It may be noted that stratification and post-stratification usually reduce the variance and clustering increases variance of an estimate.

In general, design effects tend to be greater than one due to clustered sample design of the LFS. The labour force status categories such as "employed", "unemployed" by age-sex groups tend to have lower design effects due to post-stratification by age-sex which decreases their variance. Those for labour force status by particular industry tend to

be large due to their location in specific areas. Design effects are known to be related to measures of homogeneity and average size of clusters. Models expressing their relationships have been developed for many surveys. In a study on components of variance in the LFS the design effects and measures of homoegeneity have been analyzed for a number of characteristics [2].

A measure of precision of estimates which is independent of the level of the estimate and the scale is coefficient of variation. The $CV(\hat{X})$ is given by

$$CV(\hat{X}) = \sqrt{F(W-1)\ (\frac{1}{X} - \frac{1}{P})}\ . \tag{2}$$

By taking logarithms to base e on both sides of (2) we have an equation relating CV, X and P given by

$$\log CV(\hat{X}) = \frac{1}{2}\ \log F(W-1) - \frac{1}{2} \log X + \frac{1}{2} \log\ (1 - \frac{X}{P}). \tag{3}$$

Because of the third term on the right, the equation (3) is not linear in log CV and log X, even if F(W-1) is assumed constant. However, for small values of X the contribution of the third term is negligible. A model based on (3) is given by

$$\log CV(\hat{X}) = A + B \log X + \varepsilon, \tag{4}$$

where A and B are parameters of the model and ε is the error term. The estimate of parameter B will differ from $-\frac{1}{2}$ depending on the extent to which B log X approximates $\frac{1}{2}$ log $[X/(1 - \frac{X}{P})]$ over the range of X. In an evaluation of fits of (4) and of an alternative model (5) given by

$$\log CV(\hat{X}) = A + B \log \frac{X}{(1 - \frac{X}{P})} + \varepsilon, \tag{5}$$

the goodness of fit for the two models as shown by R^2, the ratio of regression sum of squares to total sum of squares, was found to be

quite close. The model (4) is linear in log X and log CV and is simpler than model (5).

A non-linear model corresponding to (4) is given by:

$$CV(\hat{X}) = A' X^{B'} + \varepsilon, \qquad (6)$$

where A' and B' are parameters of the model and ε is the error term. The two models (4) and (6) were fitted to data on monthly estimates and their CVs for 90 characteristics in each of 10 provinces and Canada.

3. GROUPING OF CHARACTERISTICS

The monthly design effects of LFS estimates for January-December 1980 for each of 90 characteristics excluding total population for each province and Canada were averaged and plotted to decide the ranges for the two groups. In each province, the first group consists of characteristics with design effects greater than D.

Table 1 shows the boundary values D for group I and II in each province and at Canada level, and the number of characteristics in group II. The grouping of characteristics was done by arranging characteristics in increasing order of average design effects. The boundary value D was selected so that the assumption of equal design effects was satisfied as far as possible in group I. The second group consists of all remaining characteristics where the assumption of equal design effects is more crude. Most characteristics pertaining to labour force status by age-sex groups fall in group I. "Employed by industry" and "duration of unemployment" mostly fall in group II. The average design effects differ substantially between provinces and for Canada. More refined grouping of characteristics on the basis of models for design effects is being investigated.

It may be noted that about 80% of the characteristics in each province and for Canada, have been classified in group I. For obtaining a

conservative estimate of CV for a new characteristic models based on group II can be used. For a characteristic for which monthly estimates of CV are routinely produced the models for the group in which the characteristic falls, can be used to obtain approximate estimate of CV with a greater precision than that based on monthly data.

In the following section the assumptions made in fitting the models (4) and (6) are explained and model fits are evaluated.

4. EVALUATION OF MODELS

The basis of fitting the log-linear model (4) is to treat the model as a simple linear regression model in $y = \log CV(\hat{X})$ and $x = \log X$ and to obtain estimates of parameters A and B in the linear regression framework. The usual assumptions of independence of errors and constant variance have been made. Under these assumptions, R^2 provides a measure of fit of the model. The values of the estimated parameters and coefficients of determination, R^2, for group I and II in 10 provinces and Canada are given in Table 2. The actual fitting of these models was done by using SAS utility.

All R^2 values are significant and quite high indicating that the fits are very good. The error plots do not show any patterns to conclude that the assumption of constant variance is not satisified. Under these assumptions and normality of errors $CV(\hat{X})$ has a log-normal distribution with constant CV for any value of X.

The non-linear model (6) was fitted by Gauss-Newton method using SAS utility. The initial values of parameters A' and B' were assumed to be 1.00 and -0.50 respectively. The number of iterations required to reach convergence was at most 8 for each province and Canada, the convergence criterion being that the relative difference between successive error sum of squares is less than 10^{-8}. Table 3 shows values of estimated parameters and errors sum of squares for Canada Group II. The errors are approximately normally distributed as shown by normal probability plots.

Since it is of interest to compare the fits of the non-linear model for provinces, Canada and the two groups it is necessary to have a criterion of goodness of fit. In the non-linear model, the total sum of squares is not equal to the total of regression and error sums of squares. A criterion R'^2 can be defined as

$$R'^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2},$$

where $\hat{Y}_i$'s are estimated CVs based on the model, Y_i's are observed CVs and $\bar{Y}$ their mean. The summation extends over N, the number of characteristics in the group multiplied by 12, the number of months. In the linear case $R^2 = R'^2$. However, in the non-linear case $R^2 \neq R'^2$ since the total sum of squares is not equal to regression sum of squares plus error sum of squares due to product term not being zero.

The errors $(Y_i - \hat{Y}_i)$ will be small when the fit is good giving a value of R'^2 close to 1, the errors $(Y_i - \hat{Y}_i)$ will be large when the fit is poor giving a small value of R'^2. When all the points lie on the fitted curve i.e. $Y_i = \hat{Y}_i$ for all i, $R'^2 = 1$. However, in general no lower bound to R'^2 seems to exist. The values of R'^2 shown in Table 4 tend to be greater for group I as compared to group II, which has 13 to 21 characteristics out of the total of 90.

Although the log-linear model (4) was fitted to data on logarithms of estimates and their CVs and its fit seems to be good, the fitted models for provinces and Canada are used for estimation of CV of estimates. In order to compare the fit of the transformed model to original data of estimates and their CVs, these data and the transformed model corresponding to (4) were plotted for the two groups in 10 provinces and Canada. From these charts it can be concluded that the transformed model corresponding to (4) fits the data of estimates and their CVs better than the non-linear

model (6), especially for small values of estimates. The plots of these models for Canada group II are shown on Chart 1 and 2.

5. CONCLUDING REMARKS

The characteristics considered are total persons with labour force status by age-sex, industry, marital status and total persons with various ranges of duration of unemployment. However, the models can also be used for proportions instead of totals. The models are not applicable to estimates for subprovincial areas such as urban centres or groups of economic regions, since design effects for these areas are more unstable and can be much higher due to the effect of ratio-adjustment based on projected population at province level [1].

An assumption made in the use of models for a new characteristic is that its design effect is close to the average for the group. This requires finer grouping of characteristics of various types possibly on the basis of models relating design effects with measures of homogeneity for these characteristics. In fitting the models, it was assumed that errors are uncorrelated and that independent variable is fixed. Since twelve monthly estimates for each characteristic were used, there could be correlation in errors for estimates for a given characteristic. Extension of the study to models with errors in independent variable and correlated errors is being considered.

A problem in evaluation of fit of non-linear models, whether actually fitted to data or transformed from linear models, is the lack of a criterion for comparison of fits of different models. The criterion suggested in section 4 may be appropriate for comparison of fits of a model to different data sets, but may not work for different models.

TABLE 1: DESIGN EFFECT BOUNDARY VALUES AND NUMBERS OF CHARACTERISTICS IN GROUPS I AND II*

Province	Boundary Value (D)	Number of Characteristics	
		Group I	Group II
Newfoundland	2.3	75	15
P.E.I.	1.9	73	17
Nova Scotia	1.9	74	16
New Brusnwick	2.2	77	13
Quebec	1.9	73	17
Ontario	1.7	69	21
Manitoba	2.0	76	14
Saskatchewan	2.8	76	14
Alberta	2.1	71	19
British Columbia	2.3	73	17
Canada	1.9	77	13

* A characteristic belongs to Group I if its design effect (averaged over the 12-month period from January to December 1980) is less than or equal to the boundary value D. If the average design effect is greater than D, then the characteristics is in Group II.

TABLE 2: REGRESSION COEFFICIENTS AND R^2 FOR LOG-LINEAR MODEL

Province	Group	Regression Coefficient A	B	R^2
Newfoundland	I	3.3119	-0.5723	0.9534
	II	3.7757	-0.6101	0.9377
P.E.I.	I	2.7962	-0.5617	0.9485
	II	3.1796	-0.5885	0.8887
Nova Scotia	I	3.4612	-0.5837	0.9702
	II	3.6412	0.5257	0.8717
New Brunswick	I	3.2782	-0.5545	0.9606
	II	3.7544	-0.6017	0.9357
Quebec	I	4.3298	-0.5942	0.9686
	II	4.3093	-0.5216	0.9127
Ontario	I	4.3825	-0.6053	0.9736
	II	4.1796	-0.5009	0.9633
Manitoba	I	3.5155	-0.5926	0.9619
	II	3.8769	-0.5640	0.9166
Saskatchewan	I	3.3796	-0.5700	0.9544
	II	3.5478	-0.4423	0.8994
Alberta	I	3.6960	-0.5968	0.9678
	II	3.7526	-0.5090	0.9513
B.C.	I	3.9847	-0.5750	0.9621
	II	3.9814	-0.4708	0.8410
Canada	I	4.3458	-0.5936	0.9703
	II	4.2357	-0.5191	0.9699

TABLE 3: NON-LINEAR LEAST SQUARES: GAUSS-NEWTON METHOD

CANADA (GROUP II)

Iteration	A'	B'	Residual S.S.
0	1.00000000	-0.50000000	3401.93232121
1	15.22076853	-0.23647629	461.76322678
2	26.47981387	-0.36743343	322.67707190
3	51.94184546	-0.51147529	248.68405130
4	57.29455529	-0.47434886	99.32440727
5	58.32558100	-0.48419609	96.57832290
6	58.28627964	-0.48409502	96.57810754
7	58.28746710	-0.48409960	96.57810746

TABLE 4: R'^2 FOR GROUP I AND II

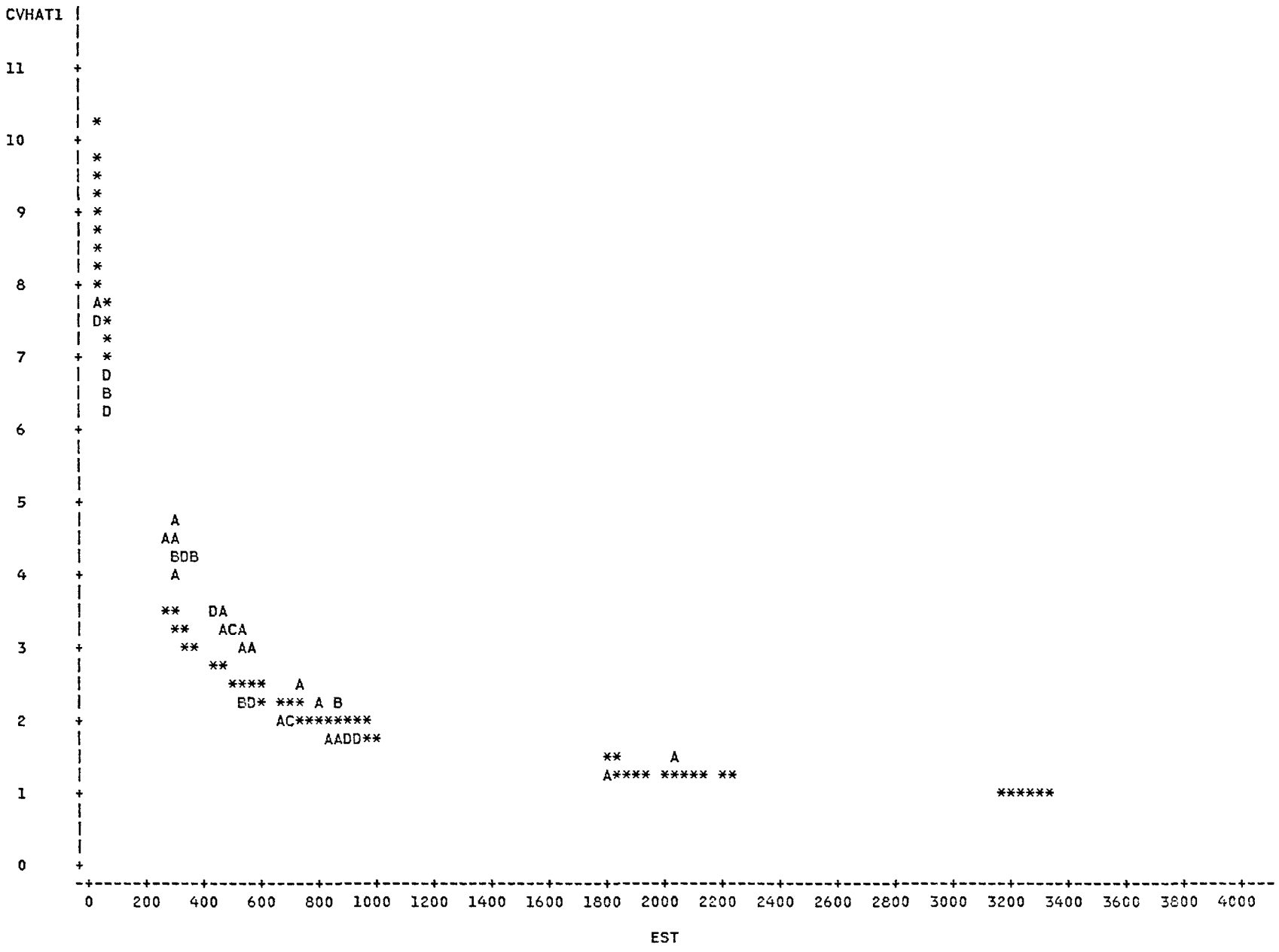
Province	Group	N*	$R'^2 = 1 - \frac{\text{Error S.S.}}{\text{Total S.S.}}$
Newfoundland	I	866	0.9362
	II	190	0.8835
P.E.I	I	827	0.8925
	II	294	0.7285
Nova Scotia	I	872	0.9790
	II	192	0.7813
New Brunswick	I	908	0.9990
	II	156	0.8639
Quebec	I	859	0.9800
	II	204	0.7804
Ontario	I	823	0.9632
	II	252	0.9208
Manitoba	I	895	0.9691
	II	168	0.8137
Saskatchewan	I	896	0.9436
	II	168	0.8196
Alberta	I	845	0.9701
	II	228	0.8852
B.C.	I	868	0.9319
	II	204	0.7786
Canada	I	923	0.9665
	II	156	0.9286

* N for group I can be less than 12 (no. of characteristics) due to exclusion of characteristics with zero estimates.

CHART 1

#1 LOG-LINEAR MODEL (transformed)
PROV=CANADA GROUP=2

PLOT OF CVHAT1*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.

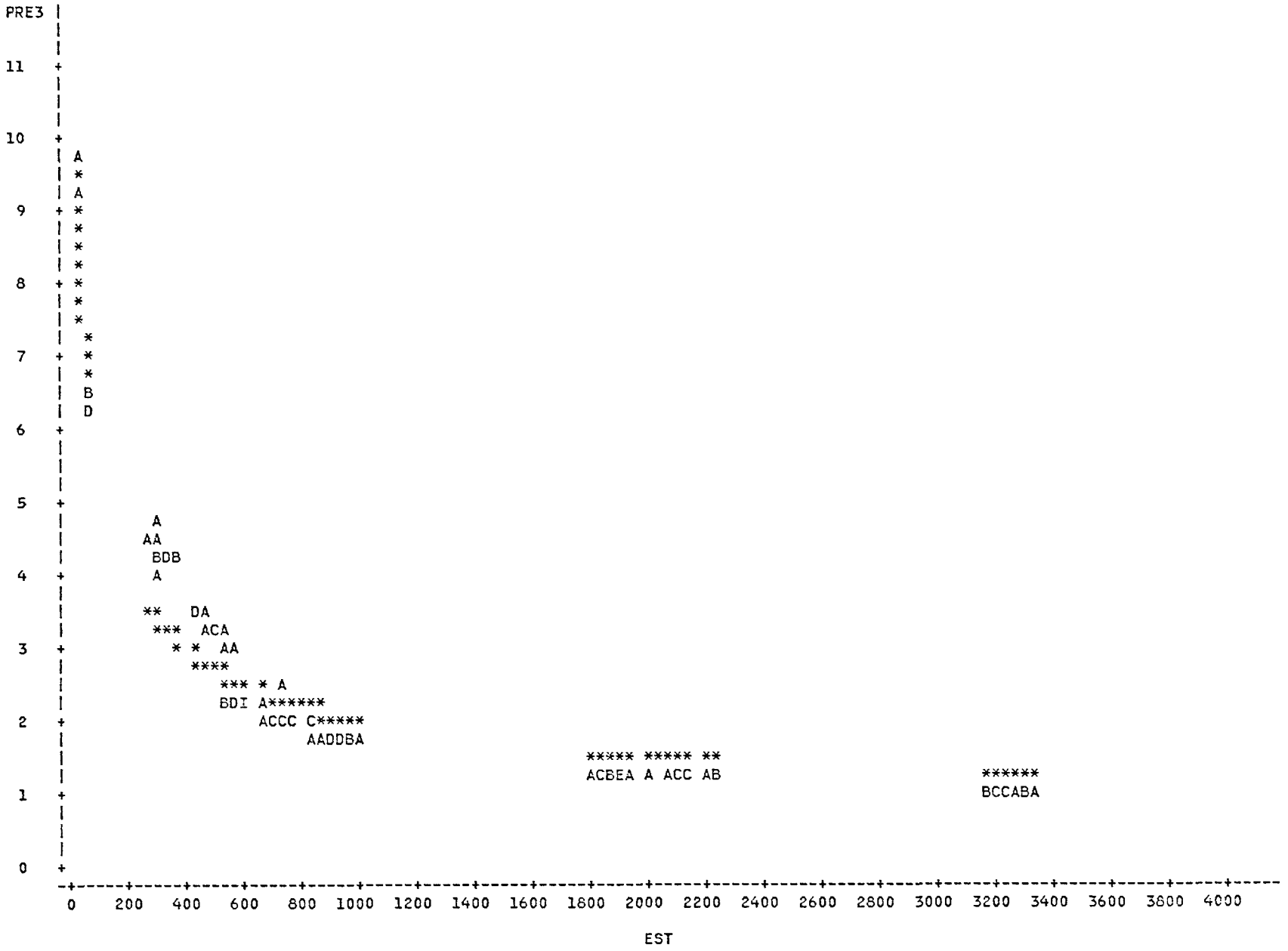


NOTE: 98 OBS HIDDEN

CHART 2

#3 SIMPLE POWER MODEL
PROV=CANADA GROUP=2

PLOT OF PRE3*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



ACKNOWLEDGEMENTS

The author would like to thank the referee for helpful comments.

REFERENCES

[1] Ghangurde, P.D. and Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics, Theory and Methods, A 10(22), 2327-38.

[2] Gray, G.B. and Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Survey Methodology, Vol. 2, No. 1., 1-30.

[3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", Presented at Meeting of the International Association of Survey Statisticians.

[4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are selected from Each Stratum". Journal of the American Statistical Association, 52, 503-510.

[5] Platek, R. and Singh, M.P. (1976), Methodology of the Canadian Labour Force Survey. Catalogue 71-526 occasional.

[6] Sprent, P. (1969) "Models in Regression and Related Topics", Methuen and Co.

The Editorial Board wish to thank Mr. R.E. Drover, Chairman, Publication Board and the staff of Administrative Services Division for their continuing support of the production of this Journal.

Acknowledgement is also due to N. Brien, M. Fluet, P. Foy and G. Kriger for their proofreading and preparation of final content.

Thanks are also due to Mrs. D. Edirisinghe, for her patient typing of the Journal, as well as for the execution of numerous other duties associated with its production. Finally, the Editorial Board wish to thank the following persons who have served as referees during the past year.

D.A. Binder
R.G. Carter
G.H. Choudhry
D.P. Dixon
J.D. Drew
S. Earwaker
M. Fluet
P. Foy
G.B. Gray
M.A. Hidiroglou
G. Kriger
S. Kumar
M.L. Lawes
I. Macredie
M.J. March
A. Satin
K.P. Srinath
L. Swain
P.F. Timmons