

OVERVIEW OF ISSUES IN REGRESSION ANALYSIS USING COMPLEX SURVEY DATA

Barbara Chun¹

ABSTRACT

Many surveys have sampling designs which make use of the population structure through stratification, clustering, and unequal probabilities of selection. Analyses of these data often ignore aspects of the design and proceed under simple random sampling assumptions which can lead to invalid inference. In regression analysis, in order to study the relationship between variables, one attempts to fit a model to the data. The question which arises is how sample data derived from a complex design can be used for inference about model parameters. A summary of the issues involved in model-based versus design-based approaches to inference are presented. Conditions for ignorability of the sampling design and methods used to protect against bias and model misspecification are discussed. An example of a regression analysis which incorporates sampling design is presented using data from the Survey of Work Arrangements which is a Canadian Labour Force Survey supplement.

KEY WORDS: Model-based; design-based; ignorable sample design; regression analysis; complex survey design.

RÉSUMÉ

Plusieurs sondages ont des plans d'échantillonnage qui font usage de la structure de la population à travers la stratification, la création des grappes et des probabilités de sélection inégales. L'analyse de ces données ignore souvent certains aspects du plan et procède sous hypothèse d'un tirage aléatoire simple qui peut cependant amener des inférences invalides. En analyse de régression, en vue d'étudier la relation entre les variables, les gens tentent d'ajuster un modèle aux données. La question naturelle est comment les données d'échantillonnage obtenues par plans complexes peuvent être utilisées afin de faire de l'inférence concernant les paramètres du modèle. On présente un résumé des possibilités découlant d'une approche basée sur le modèle comparativement à une approche basée sur le plan d'échantillonnage pour faire de l'inférence. Les conditions d'ignorabilité du plan d'échantillonnage et les méthodes utilisées pour se protéger du biais et de la mauvaise spécification du modèle sont discutées. Un exemple d'une analyse de régression qui incorpore le plan d'échantillonnage en utilisant le logiciel SUDAAN est présenté à l'aide de données provenant de l'Enquête de l'organisation du travail qui est un supplément de l'Enquête sur la population active.

MOTS CLÉS: Modèle; plan de sondage; ignorabilité du plan d'échantillonnage; analyse de régression; enquête complexe.

1. INTRODUCTION

1.1 Description of the Problem

In many practical applications, data being analysed come from complex survey designs which are characterised typically by multiple stages of sampling, unequal probabilities of selection, stratification, and clustering. Ordinary least squares methods (OLS) used in most standard statistical packages for regression analysis are based on the assumption that the data are independently and identically distributed. The issues which arise in regression analysis using complex survey data is the subject of this paper.

1.2 Organisation of the Paper

In section 2 assumptions of the classical linear regression model, a description of model-based and design-based inference, and the idea of ignorability of sampling design are presented. Section 3 presents an example of regression analysis using two different software packages to illustrate the effect of taking complex design into account in the analysis. In section 4, a brief summary is given.

¹ Barbara Chun, Statistics Canada, Ottawa, K1A 0T6, chunbar@statcan.ca

2. LITERATURE REVIEW

2.1 Linear Regression Model

The following is the classical linear regression model described in most texts and used in many standard statistical packages relating a variable Y and a vector of variables \mathbf{x} for each unit i (Neter et al., 1983):

$$Y_i = \alpha + \mathbf{x}_i' \beta + \epsilon_i$$

Assumptions of the model are:

- linearity: $E(\epsilon_i | \mathbf{x}_i) = 0 \forall i$
- homoscedasticity: $\text{Var}(\epsilon_i | \mathbf{x}_i) = \sigma^2 \forall i$
- independence between observations:
 $\text{Cov}(\epsilon_i, \epsilon_j | \mathbf{x}_i, \mathbf{x}_j) = 0 \forall i \neq j$
- normality of the errors: $\epsilon_i \sim N(0, \sigma^2)$

The ordinary least squares estimator for β is

with variance $\text{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. The

OLS estimator $\hat{\beta}$ has the desirable property of being the minimum variance, unbiased estimator of β under this model.

In many practical applications, as in the case of samples from complex surveys, these assumptions are violated (Landis et al., 1982). The standard estimation procedure of OLS does not incorporate any probability weighting or population structure such as clustering, stratification or measures of size into the analysis (Pfeffermann and Holmes, 1985). For such sampling situations, it is now well known that, in general, ordinary least squares regression estimates will be biased even asymptotically (Nathan and Holt, 1980).

2.2 Complex Survey Design

Survey populations have complex structures which may be incorporated into survey designs through stratification, clustering, and unequal probabilities of selection. Gains in efficiency of estimators can be realised through stratifying the population into homogenous groups; administrative convenience and cost savings can result from clustered sample selection; and special interest in sub-groups may result in over-sampling certain segments of the population. This complexity must be taken into account in the analysis if valid inferences are to result (Kish and Frankel, 1974). For example, clustered selection tends to introduce positive correlations between the errors of the model which can often have serious consequences if ignored (Skinner et al., 1989; Pfeffermann, 1996).

There are several reasons why complex design is not taken into account in analysis (Skinner et al., 1989).

Standard statistical software such as SAS, SPSS, and BMDP are familiar to many analysts and have procedures already in place. Many analyses are secondary analyses of data which are not provided at the micro-level. For confidentiality reasons, stratum and cluster information may not be available on the data file.

Methods proposed to account for the complex sample design involve the use of weights in a design-based approach or modification of the model in the model-based approach. Excellent discussions of inference based on the design-based randomisation principle compared with the model-based approach is given in Royall and Cumberland, 1981, and Brewer and Mellor, 1973.

2.3 Model-Based vs. Design-Based Inference

2.3.1 Model-Based Inference

In model-based inference, each population element is considered a random variable and the finite population is one realisation of an assumed super-population. The target parameter is a characteristic of the model. For the super-population approach, expectation is defined over the possible realisations under the model of the population, conditional on the sample units observed. Thus inference is based on the sampling distribution of statistics over repeated realisations y_1, \dots, y_N generated by the model ξ , with the selected sample held fixed (Skinner et al.; Pfeffermann and Holmes, 1985). The analysis is done as if the only source of variation were random sampling from the hypothetical super-population.

The main argument for model-based inference rests on the assumption that, if a model is appropriate, then a sampling plan and the best estimator for that model can be determined, which may result in an estimator with sampling errors that are smaller than for alternative possible designs or estimators (Hansen et al., 1983). This may have several advantages if the model is appropriate. For example, if the assumed model is appropriate, then inferences can be based on relatively small samples. If a super-population model applies to all population units then model-based inference is appropriate because of certain optimality properties under the assumed model (Binder, 1983; DuMouchel and Duncan, 1983).

In regression analysis one may possibly model different regression equations for different clusters, or model a single regression line which allows for intracluster correlations between residual terms belonging to the same cluster. The unknown model parameters can be estimated using maximum likelihood,

Bayesian, or some other optimal strategies (Holt et al., 1980).

2.3.2 Ignorable Sampling Schemes

Conditions for ignorability of a sampling design are presented and formalised in Pfeffermann (1996) and Pfeffermann (1991). The basic idea is that a design is ignorable if it is of known form and depends only on the known values, \mathbf{Z} , of the design variables, such as stratification factors, clusters, measures of size etc., i.e., $P(\mathbf{I} | \mathbf{Y}, \mathbf{Z}) = P(\mathbf{I} | \mathbf{Z})$, where $\mathbf{I} = (I_1 \dots I_N)$ is the sample indicator variable such that $I_i=1$ if i is in the sample and $I_i=0$ otherwise, and $\mathbf{Y} = (Y_1 \dots Y_N)$ is the vector of response variables associated with unit i .

Then design information can be incorporated into the model so that the actual selection probabilities carry no extra information. The ignorability of the sampling design depends on the available design information and also on the model and parameters of interest. If the regressor variables in a regression model include all the design variables, the sampling design is ignorable for estimating the regression model and standard methods apply.

The ignorability of the design may be tested by testing the significance of the difference between the best (optimal) estimator of the vector of regression coefficients under a particular model which assumes that the design is ignorable, $\hat{\beta}$, and the weighted least squares estimator $\hat{\beta}_w$ (Pfeffermann, 1996; Pfeffermann, 1991).

The hypothesis test is $H_0: \lim_{n \rightarrow \infty} p \lim (\hat{\beta} - \hat{\beta}_w) = 0$ as $n \rightarrow \infty$. The test statistic is $\hat{\lambda} = \hat{\mathbf{D}}' [\hat{\mathbf{V}}(\hat{\mathbf{D}})]^{-1} \hat{\mathbf{D}}$ where $\hat{\mathbf{D}} = \hat{\beta} - \hat{\beta}_w$ and $\hat{\mathbf{V}}(\hat{\mathbf{D}})$ is an estimator of the variance-covariance matrix of $\hat{\mathbf{D}}$. If $\hat{\mathbf{V}}(\hat{\mathbf{D}})$ is the randomisation variance-covariance matrix then $\hat{\lambda} \sim F_{p, n-2p-L}$ where L is the number of strata. The choice between $\hat{\beta}$ and $\hat{\beta}_w$ is influenced by the choice of a model-based approach to inference vs. an approach based on randomisation within a finite population in which no particular model is assumed (Korn and Graubard, 1995; Pfeffermann, 1991). A comparison of both weighted and unweighted parameter estimates can also be a useful diagnostic check of model adequacy (DuMouchel and Duncan, 1983).

2.3.3 Design-Based Inference

In a design-based approach, inference is based on the sampling distribution of statistics over repeated samples generated by the sampling design, with finite population values y_1, \dots, y_N fixed. The target parameter is a characteristic of the finite population. Expectation is

defined over all possible samples that would be obtained from the finite population under a specified sample design (Skinner et al.; Pfeffermann and Holmes, 1985).

Fitting models that closely approximate the behaviour of a heterogeneous population and incorporate all elements of a complex design is not always practical (Pfeffermann, 1996). Model parameters should be replaced by descriptive population quantities (DPQ) which are robust and interpretable (Pfeffermann, 1991). Assuming that the finite population is a simple random sample from the super-population, then the DPQ, B , which is the estimator of the model parameter, J , in the case of a census, is a model-consistent estimator for the model parameter. A design-based estimator for the DPQ, \hat{B} , is then a consistent estimator of the model parameter (Pfeffermann, 1996), i.e.,

$$\hat{B} - \beta = \underbrace{\hat{B} - B}_{\text{design-based inference}} + \underbrace{B - \beta}_{\text{tends to be small}}$$

where the error in

In regression analysis, define the target parameter of interest as the least squares solution in the case of a census, i.e., the DPQ, $B = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y} =$

$$\left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i$$

and estimate B using, for example, the weighted least squares estimator

$$\hat{B} = (\mathbf{X}_s' \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{W}_s \mathbf{Y}_s = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i \right).$$

B has a clear and meaningful interpretation even if the model fails to hold, i.e., it is the slope of the best-fitting prediction equation, in the least-squares sense, for all the units in the population. It is consistent and in many situations the loss of efficiency is relatively small in the case of surveys where the sample size is large (Hansen et al., 1983). Weights can be used to test and protect against nonignorable sampling designs which could cause bias and can be used to protect against misspecification of the model holding in the population (Pfeffermann, 1996; Kott, 1991; Pfeffermann, 1991). Methods for incorporating weights in the inference process are described in Pfeffermann, 1991 and Pfeffermann and Holmes, 1985.

3. EXAMPLE

The Survey of Work Arrangements (SWA) was administered as a supplement to the 1995 November Canadian Labour Force Survey (LFS). It was designed to collect detailed information on conditions of work and pay. The LFS sample design is based upon a stratified multi-stage cluster design employing probability sampling at all stages of the design.

SUDAAN (Survey Data Analysis Software, Release 7.0, Research Triangle Institute, Research Triangle Park, NC 27709) is a sample survey software package that is designed to take into account complex design elements such as stratification, clustering, and unequal probabilities of selection. Regression coefficients are computed by weighted least squares and variance estimates are calculated using the Taylor linearization technique. An analysis of variance model was fit to the SWA data in order to evaluate the relationship between hourly salary and several independent variables using SUDAAN. Design variables such as stratum and psu information were obtained from the LFS data file. The same model was fit using SAS with the WEIGHT option to produce a weighted least squares analysis. The weights were used in the SAS analysis to provide a comparison of a model-based approach which "naïvely" attempts to incorporate selection probabilities into the analysis to produce variance estimates with design-based properties and purely design-based variance estimation which does not rely on a model. The results are given in Table 1 in the Appendix.

Both analyses calculate weighted least squares parameter estimates $\hat{\beta}$. However, the SAS estimates of precision do not take into account the complex design features of the data. The standard errors for the SUDAAN regression are consistently higher (and hence the t-values lower) than for the SAS analysis except for the eastern provinces and Saskatchewan. Yet, for most t-tests, we would have come to the same conclusions with both analyses. However, we would have rejected the null hypothesis that $\beta=0$ using SAS but not using SUDAAN for the age 55-64 group, i.e., we would have concluded that there is a statistically significant difference in hourly salary between the age group 55-64 and the reference group if design had not been taken into account in the variance estimation.

Failure to recognise the intracluster correlation in estimation of variance in our example may be leading to understatement of confidence intervals and overstatements of precision. Thus test statistics based on downwardly biased estimates of variances would result

in associations appearing to be more significant than they really are.

4. SUMMARY

Failure to account for the sample selection process may bias the inference (Korn and Graubard, 1995). One example of a model-based approach to deal with this is to augment the regression equation with design variables so that the design becomes ignorable and standard methods apply. This should lessen the potential bias of the sampling design while retaining the efficiency of OLS. In a design-based approach one can define the target parameter as the OLS estimator in the case of a census and use probability-weighted estimators to protect against non-ignorable designs and model misspecification (Pfeffermann, 1996). Standard statistical packages may produce unbiased point estimators for the regression coefficients but inferences about the precision of will be incorrect (Skinner et al., 1989). Design-based methods for allowing for complex designs in the standard error of include adjusting SRS-based standard errors using an estimated design effect or using a robust variance estimation technique such as Taylor linearization, balanced repeated replication, or jackknife (Skinner et al., 1989; Kott, 1991).

REFERENCES

- Binder, D.A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys". *International Statistical Review*, 51, 279-292.
- Brewer, K.R.W., Mellor, R.W. (1973). "The Effect of Sample Structure on Analytical Surveys". *Austral. J. Statist.*, 15:3, 145-152.
- DuMouchel, W.H., Duncan, G.J. (1983). "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples". *JASA*, 78:383, 535-543.
- Hansen, M.H., Madow, W.G., Tepping, B.J. (1983). "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys". *Journal of the American Statistical Association*, 78:384, 776-793.
- Holt, D., Smith, T.M.F., Winter, P.D. (1980). "Regression Analysis of Data from Complex Surveys". *J. R. Statist. Soc. B*, 143:4, 474-487.
- Kish, L., Frankel, M.R. (1974). "Inference from Complex Samples". *J. R. Statist. Soc. B*, 36, 1-37.

- Korn, E.L., Graubard, B.I. (1995). "Analysis of Large Health Surveys: Accounting for the Sampling Design". *J. R. Statist. Soc. A*, 158:2, 263-295.
- Kott, P.S. (1991). "A Model-Based Look at Linear Regression With Survey Data". *The American Statistician*, 45:2, 108-112.
- Landis, J.R., Lepkowski, J.M., Eklund, S.A., Stehouwer, S.A. (1982). "A Statistical Methodology for Analyzing Data from a Complex Survey: The First National Health and Nutrition Examination Survey". *Vital and Health Statistics*, Series 2, No. 92.
- Nathan, G., Holt, D. (1980). "The Effect of Survey Design on Regression Analysis". *J. R. Statist. Soc. B*, 42:3, 377-386.
- Neter, J., Wasserman, W., Kutner, M.H. (1983). *Applied Linear Regression Models*. Homewood, Illinois: Richard D. Irwin, Inc.
- Pfeffermann, D., Holmes, D.J. (1985). "Robustness Considerations in the Choice of a Method of Inference for Regression Analysis of Survey Data". *J. R. Statist. Soc. A*, 148:3, 268-278.
- Pfeffermann, D. (1993). "The Role of Sampling Weights when Modeling Survey Data". *International Statistical Review*, 61:2, 317-337.
- Pfeffermann, D. (1996). "The Use of Sampling Weights for Survey Data Analysis". *Statistical Methods in Medical Research*, 5, 239-261.
- Royall, R.M., Cumberland, W.G. (1981). "An Empirical Study of the Ratio Estimator and Estimators of Its Variance". *JASA*, 76:373, 66-88.
- Skinner, C.J., Holt, D., Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons Ltd.

ACKNOWLEDGEMENTS

The author would like to thank Milorad Kovacevic, Georgia Roberts, and S. Kumar for their valuable comments. Thanks also to Lee Grenon for his contribution to the analysis of the Survey of Work Arrangements data.

APPENDIX

Table 1. Comparison of weighted least squares regression in SUDAAN and SAS (SAS results appear in parentheses)

Independent Variables and Effects	Beta Coeff.	SE Beta B=0	T-Test T-Test B=0	P-value
Intercept	9.53	0.32 (0.24)	29.75 (39.23)	0.0000 (0.0001)
LFSEX				
male	2.64	0.12 (0.09)	21.99 (28.71)	0.0000 (0.0001)
MARSTAT ¹				
married	1.54	0.16 (0.11)	9.54 (14.13)	0.0000 (0.0001)
other	1.10	0.27 (0.18)	4.05 (6.18)	0.0001 (0.0001)
EDUCCAT ²				
0-8 years	-2.26	0.26 (0.23)	-8.69 (-10.01)	0.0000 (0.0001)
some secondary	-0.88	0.15 (0.14)	-5.80 (-6.38)	0.0000 (0.0001)
some post-secondary	0.62	0.18 (0.15)	3.40 (4.02)	0.0007 (0.0001)
post-secondary cert.	1.34	0.14 (0.11)	9.61 (11.78)	0.0000 (0.0001)
university degree	4.61	0.24 (0.14)	19.24 (31.94)	0.0000 (0.0001)
TENCAT ³				
1-6 months	-0.57	0.16 (0.13)	-3.54 (-4.35)	0.0004 (0.0001)
7-12 months	-0.60	0.19 (0.16)	-3.21 (-3.69)	0.0013 (0.0002)
6-10 years	1.45	0.17 (0.12)	8.70 (12.12)	0.0000 (0.0001)
11-20 years		0.19 (0.13)	14.27 (20.52)	0.0000 (0.0001)
over 20 years	3.72	0.24 (0.17)	15.66 (22.34)	0.0000 (0.0001)
SIZE ⁴				
< 20	-2.09	0.17 (0.12)	-12.02 (-17.55)	0.0000 (0.0001)
20-99	-1.25	0.16 (0.11)	-7.66 (-11.15)	0.0000 (0.0001)
over 500	1.38	0.21 (0.14)	6.51 (9.53)	0.0000 (0.0001)
CLASS OF WORKER				
public sector	1.41	0.20 (0.14)	6.97 (9.85)	0.0000 (0.0001)
PROV ⁵				
Nfld	-2.36	0.24 (0.33)	-9.97 (-7.10)	0.0000 (0.0001)
PEI	-3.28	0.24 (0.63)	-13.45 (-5.17)	0.0000 (0.0001)
NS	-2.93	0.22 (0.24)	-13.49 (-12.11)	0.0000 (0.0001)
NB	-2.56	0.19 (0.27)	-13.20 (-9.36)	0.0000 (0.0001)
QUE	-0.97	0.16 (0.10)	-6.03 (-9.43)	0.0000 (0.0001)
Man	-2.11	0.22 (0.21)	-9.65 (-10.02)	0.0000 (0.0001)
Sask	-1.79	0.18 (0.24)	-9.84 (-7.61)	0.0000 (0.0001)
Alta	-1.07	0.17 (0.14)	-6.20 (-7.58)	0.0000 (0.0001)
BC	0.94	0.21 (0.13)	4.59 (7.24)	0.0000 (0.0001)
AGECAT ⁶				
15-19	-0.62	0.24 (0.23)	-2.59 (-2.75)	0.0097 (0.0060)
20-24	-1.22	0.19 (0.15)	-6.40 (-7.91)	0.0000 (0.0001)
35-44	1.06	0.15 (0.11)	6.86 (9.69)	0.0000 (0.0001)
45-54	1.57	0.20 (0.13)	8.02 (11.88)	0.0000 (0.0001)
55-64	0.42	0.29 (0.19)	1.44 (2.16)	0.1497 (0.0306)
65-69	-1.09	1.16 (0.73)	-0.94 (-1.50)	0.3474 (0.1349)
INDGRP ⁷				
agriculture & other primary	3.16	0.40 (0.31)	7.98 (10.06)	0.0000 (0.0001)
manufacturing	0.68	0.21 (0.15)	3.22 (4.50)	0.0013 (0.0001)
construction	2.94	0.36 (0.28)	8.24 (10.69)	0.0000 (0.0001)
transportation	1.38	0.34 (0.26)	4.04 (5.30)	0.0001 (0.0001)
communications	1.99	0.28 (0.22)	7.24 (8.97)	0.0000 (0.0001)
trade	-0.93	0.19 (0.15)	-4.81 (-6.33)	0.0000 (0.0001)
finance	1.09	0.31 (0.19)	3.52 (5.61)	0.0004 (0.0001)
public administration	1.01	0.28 (0.20)	3.60 (5.05)	0.0003 (0.0001)
OCCGRP ⁸				
managerial, admin.	3.68	0.21 (0.15)	17.79 (24.77)	0.0000 (0.0001)
prof. & technical	3.00	0.21 (0.15)	14.39 (20.00)	0.0000 (0.0001)
sales	0.14	0.21 (0.19)	0.65 (0.74)	0.5176 (0.4569)
service	-1.27	0.18 (0.16)	-6.95 (-7.84)	0.0000 (0.0001)
primary	-1.37	0.48 (0.41)	-2.84 (-3.32)	0.0046 (0.0009)
construction	0.28	0.19 (0.16)	1.48 (1.75)	0.1384 (0.0800)
transportation	1.31	0.33 (0.28)	3.96 (4.74)	0.0001 (0.0001)
fabricating, materials handling	-0.42	0.33 (0.28)	-1.26 (-1.52)	0.2080 (0.1292)
JOB STATUS				
permanent	0.16	0.19 (0.14)	0.88 (1.18)	0.3782 (0.2379)
COVERED				
yes	0.65	0.15 (0.10)	4.40 (6.27)	0.0000 (0.0001)
ENROLLED				
yes	-0.39	0.18 (0.15)	-2.15 (-2.56)	0.0314 (0.0105)

¹ reference group single

² reference group high school graduate

³ reference group 1-5 years

⁴ reference group 100-500

⁵ reference group Ontario

⁶ reference group 25-34

⁷ reference group business, community and personal services

⁸ reference group clerical