Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 551-559
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Simple Linear Regression Model for Hidden/Hard-to-Reach/Elusive Populations

Maahi Tuahiru, Mubarika Alhassan, Haadi Abdul-Rahaman

Department of Statistics, Tamale Technical University, Tamale, Ghana
Email: tmaahi@tatu.edu.gh, amubarika@tatu.edu.gh, ahaadi@tatu.edu.gh

How to cite this paper: Tuahiru, M., Alhassan, M. and Abdul-Rahaman, H. (2017) Simple Linear Regression Model for Hidden/Hard-to-Reach/Elusive Populations. *Open Journal of Statistics*, **7**, 551-559.
https://doi.org/10.4236/ojs.2017.74037

Received: June 16, 2017
Accepted: July 21, 2017
Published: July 24, 2017

Abstract

Estimation methods have over the years been a problem for Statistician especially in sectors that have to do with Hidden/Hard-to-Reach population. In this paper, a regression model was derived for Elusive/Hard-to-Reach/Hidden populations. This was achieved by modelling the Multiplicity Estimator given by Birnbaum and Sirken (1965) into a regression model. The paper also gave the least-squares estimation of the unknown parameters β_0 and β_1, and σ^2.

Keywords

Multiplicity Estimator, Simple Linear Regression, Least Squares Estimation

1. Introduction

Perhaps equally important to conducting a research or survey is the estimation(s) of parameters. If it is a survey or research using conventional methods, then estimations of parameters become quite easy, as well developed probability estimators have been developed for the estimation of parameters.

It is however not the case in the studies of elusive or hard-to-reach populations. This is largely due to the fact that most of the sampling methods like network sampling used in the studies of these populations are non-probability sampling methods. Concerns are therefore always on how to estimate parameters and proper application of the method to achieve unbiased results. Three unbiased estimators for such designs were derived by Birnbaum and Sirken [1]. These estimators basically addressed the effect of multiplicity on selection probabilities of reported patients. Out of these estimators, the *multiplicity estimator* was the simplest and most robust and is now generally used whenever network sampling is used. In the *multiplicity estimator* each observation is divided by its

multiplicity and since multiplicity is proportional to the draw-by-draw selection probability, this estimator is akin to the Hansen Hurwitz estimator.

Brief History of Network or Multiplicity Sampling

Network sampling is a technique that assures unbiased estimation when the same observation units are eligible to be counted at (linked to) multiple selection units in the survey [2]. Network sampling is a technique which may be used to increase the efficiencies of sample surveys aimed at producing estimates about rare populations.

It differs from classical survey sampling with respect to the counting rule paradigm for linking population elements to the selection units at which they are countable in the survey [3]. Whiles classical survey sampling uses unitary counting rules, such as de jure and de facto residence rules in household surveys, that seek to uniquely link each person to one and only household, network sampling, on the other hand, seeks to capitalize on duplicate counting of population elements by using multiplicity counting rules, such as friendship and kinship rules in household surveys, that link the same person to multiple households of their friends or relatives. Strictly speaking however, network sampling is not a sampling technique because it does not specify the rules for selecting a sample. It can be applied in all sample surveys when a multiplicity counting rule is used for linking individual observation units to multiple selection units.

Network sampling has also been used as a synonym for multiplicity sampling, but Kish [4], thinks it's a needless and confusing redundancy. However, the name has come to stay and the two are used now interchangeably.

Network sampling emerged as an unexpected finding in the early 1960s in response to estimation problems involving a sample survey of medical providers designed to estimate prevalence of cystic fibrosis—a relatively rare genetic disease of children.

A pilot of a national stratified survey of physicians and hospitals was conducted in 1959, in three New England states to estimate the prevalence of medically diagnosed cases of cystic fibrosis [5]. Cystic fibrosis had been identified as a distinct entity in the mid-1930s and in the late 1950s. Diagnostic tests were still relatively crude and test results were often ambiguous when the survey was conducted. The procedures by which the pilot survey sought to evaluate diagnostic validity yielded information that disclosed an unanticipated survey design problem.

Medical sources involved in the survey reported all patients they had treated for cystic fibrosis since 1952, identified each patient, and reported the patient's date of birth, sex, and the medical findings supporting the cystic fibrosis diagnosis. They also identified referral medical sources, if any, that treated each of their patients, and the referral sources were subsequently queried for supplementary diagnostic information about the patients. After the survey was completed, the diagnostic information reported by the original and referral medical sources was

combined to assess the certitude of the cystic fibrosis diagnoses.

In the assessment process, it was determined that the original sample of 1600 medical sources had reported about 650 distinct cystic fibrosis patients and these patients had been treated by over 1000 different medical sources [6]. Unexpectedly, more than two-thirds of the patients had been treated by multiple medical sources. Unbiased estimation of cystic fibrosis prevalence was not a problem in the pilot survey because virtually all the cystic fibrosis patients were reported by medical providers in the certainty strata. Otherwise, this would have been a problem because matching to eliminate duplicate reports would be insufficient to ensure unbiased estimation.

Subsequent works on network sampling have reviewed these estimators. The Horvitz-Thompson estimator for network sampling, in which each person's inclusion probability is determined by the multiplicities, was also given by Birnbaum and Sirken.

Nathan [7] and Sirken [8] [9] have concentrated on the multiplicity estimator. Levy [10] and Sirken and Levy [11] examined ratios of multiplicity estimators, which could be used, for example, to estimate the proportion of an ethnic group with a rare disease. The effects of reporting errors through the linkages in network sampling—cases in which, for example, the patient's household may be more reliable at reporting the disease than a relative's household, were evaluated by Czaja *et al.* [12].

The stratified multiplicity estimator given also by Birnbaun and Sirken [1] deals with complications that arise in stratified selection units. In this case a given observational unit may be linked to selection units in more than one stratum making observations in different strata not independent as in conventional stratified sampling.

A simplified and unified review using the multiplicity approach for estimation in multiple frame surveys was given by Mecatti and Singh [13]. In their paper they considered the connection between Multiple Frame sampling, indirect sampling and Network sampling and showed how all estimators can be expressed as a multiplicity-adjusted estimator. Multiple Frame estimators was classified into two class, separate frame approach (SEP) and combined frame approach (COMB). It was shown that the unbiased Multiple Frame estimators such as the Kalton-Anderson COMB estimator and the Hartley SEP estimator with known α^H could be expressed in a Generalized Multiplicity-adjusted Horvitz-Thompson GMHT form introduced earlier by Singh and Mecatti [14] [15]. The GMHT class can be extended by relaxing the assumption of unbiasedness to approximate unbiasedness.

Mecatti [16] proposed a single frame multiplicity estimator for multiple frame survey. The estimator was proposed using the multiplicity approach because multiplicity estimators required less information about unit domain membership.

Laska *et al.* [17], proposed a model-based multiplicity estimation for popula-

tion size. Their estimator utilized two items determined for each survey participant: the number, u, among the w lists in S and the number, j, among all K lists on which each survey participant appears. In its traditional form, selection units were chosen using probability sampling and the statistical properties of the estimator derived from the sampling mechanism. Here, selection units were purposively chosen to maximize the chance that they were "typical" and a model-based analysis was used for inference. If the sample were typical, the ML estimators of N and $E(J)$ were unbiased. If a condition on the second moment of U/J were satisfied, the model-based variance of the estimator of N based on a purposively chosen typical sample was smaller than one based on a randomly chosen sample. Methods to test whether the typical assumption was valid using data from the survey were not yet available.

Multiplicity estimation continue to be one of the main estimation method for the U.S department of Health and Human Service, Centers for Disease Control and Prevention and the National Center for Health Statistics when target population seem to be rare/hidden. It has also been applied by various individuals in surveys that has to do with the estimation of rare/hidden/hard-to-reach populations.

Hing and Burt [18], described average annual estimates of nonfederal, office-based physicians who saw patients in the United States during 2005-2006. The report used a multiplicity estimator from the physician sample to estimate the number and characteristics of medical practices with which physicians were associated. Selected physician estimates of characteristics obtained only in 2006 were also presented, as well as selected trends in physician practice characteristics between 2001-2002 and 2005-2006.

Multiplicity estimation has been used in estimation for Service Based Enumeration (SBE) in census 2000 in the U.S. [19]. The SBE was designed to provide people with no usual residence an opportunity to be enumerated. Even though the multiplicity estimation procedures were used, there was a decision not to use the multiplicity estimator.

The rational was that the ratio of the multiplicity estimate to the number of persons actually enumerated in shelters (4.25 nationally) is probably too high due to the high percentage of persons responding "1" to the shelter usage question.

They felt this percentage was too high based on results from National Survey of Homeless Assistance Providers and Clients (NSHAPC) and other national findings.

Although the total national level multiplicity estimate of nearly 1 million persons was reasonably close to what was expected, using the multiplicity estimation results to distribute these persons to local areas and service facilities was not statistically defensible due to response bias to the usage questions, particularly in shelters.

Johnston *et al.* [20] conducted a survey to collect baseline measurements of HIV and syphilis prevalence and sexual risk behaviors among men who have sex with men (MSM) in Agadir and Marrakech, Morocco, and provide strategic information to improve outreach programmes. Respondent-driven sampling was

used to recruit men who reported having anal sex with another man in the last 6 months, aged 18 years and older and living in either Agadir or Marrakech for the past 6 months, regardless of nationality. Data were analyzed with the multiplicity estimator using respondent-driven sampling analysis tool. 323 MSM in Agadir and 346 in Marrakech were recruited into the survey. Most MSM in both cities reported being < 25 years, being unemployed, bisexual and in a couple with both a man and a woman. Most reported selling sex and having sex with women. HIV prevalence was 5.6% in Agadir and 2.8% in Marrakesh; syphilis was 7.0% in Agadir and 10.8% in Marrakesh. Among MSM who tested positive for HIV, 31.6% in Agadir and 56.4% in Marrakesh were co-infected with syphilis. HIV and syphilis findings coupled with high risk activities indicate the need for expanding programmes targeting MSM throughout Morocco. Selling sex and sex with women may be a strategy to cope with extreme stigma towards MSM. Criminalization and discrimination of MSM in Morocco underscores the urgent need for long-term and sustainable risk reduction through legal reforms and promotion and protection of human rights.

Seamless phase II/III clinical trials offer an efficient way to select an experimental treatment and perform confirmatory analysis within a single trial. However, combining the data from both stages in the final analysis can induce bias into the estimates of treatment effects. Methods for bias adjustment developed thus far have made restrictive assumptions about the design and selection rules followed. In order to address these shortcomings, Robertson *et al.* [21] derived a uniformly minimum variance conditionally unbiased estimator for two-stage seamless phase II/III trials. Their framework allowed for the precision of the treatment arm estimates to take arbitrary values, could be utilized for all treatments that were taken forward to phase III and was applicable when the decision to select or drop treatment arms was driven by a multiplicity-adjusted hypothesis testing procedure.

In many research problems, it is of interest to study the effects that some variables exert on others. One sensible way to describe this relationship is to relate the variables by some sort of mathematical equation. This is necessary for elusive populations since non-probability sampling methods are usually employed to estimate the population parameters of these populations. As such these populations usually do not meet the assumptions of the classical regression model. It is therefore imperative that a different regression model be derived for elusive populations.

This paper will contribute in this direction by modelling the Multiplicity Estimator given by Birnbaun and Sirken [1] into a regression equation. It will continue with the estimation of β_0, β_1 and σ^2.

2. Modelling Multiplicity Estimator into a Regression Equation

When the response variable, denoted by y, is continuous and believed to depend linearly on k variables $x_1, x_2, \cdots, x_k$ through unknown parameters $\beta_0, \beta_1, \cdots, \beta_k$

then this linear (where "linear" is used to indicate linearity in the unknown parameters) relationship is given as

$$y_i = \sum_{j=0}^{k} \beta_j x_{ji} + \varepsilon_i \tag{1}$$

where $x_{0i} = 1$ for all $i = 1, 2, \cdots, n$.

The term ε_i is unobservable random error representing the residual variation and is assumed to be independent of the systematic component $\sum_{j=0}^{k} \beta_j x_{ji}$.

It is also assumed that $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$; hence, $\mathrm{E}(y_i) = \sum_{j=0}^{k} \beta_j x_{ji}$ and $\mathrm{Var}(y_i) = \sigma^2$.

This paper will apply the above in modelling the Multiplicity Estimator given below

$$\hat{\tau}_m = \frac{M}{n} \sum_{i \in s} \frac{y_i}{m_i} \tag{2}$$

where τ_m = Population Total

M = Number of selection units in the population

m_i = Multiplicity of the i^{th} observational unit

y_i = Indicator variable, equals to 1 if the unit has characteristics of interest and 0 otherwise.

n = Sample population

In rendering the statistical properties of the multiplicity estimator transparent, it is sometimes simplified as

$$\hat{\tau}_m = \frac{M}{n} \sum_{j=1}^{n} w_j \tag{3}$$

where

$w_j = \sum_{i \in A_j} \frac{y_i}{m_i}$ is the sum of the y_i / m_i for all observational units linked with selection unit j.

The Multiplicity Estimator will be modelled as a function of k covariates, where

$$w_j = \sum_{i=1}^{n_i} w_{ij} \tag{4}$$

and

$$w_{ij} = \sum_{i=0}^{k} \beta_i x_{ij} \tag{5}$$

That is, we assume the relationship between $\hat{\tau}$ and the covariates to be

$$\hat{\tau}_m = \frac{M}{n} \sum_{i=0}^{k} \beta_i x_{ij} \tag{6}$$

This paper will concentrate on modelling the Multiplicity estimator into a simple linear regression model.

That is

$$w_{ij} = \beta_0 + \beta_1 x_i \tag{7}$$

and

$$\hat{\tau}_m = \frac{M}{n}\left(\beta_0 + \beta_1 x_i\right) \tag{8}$$

3. Estimation of β_0, β_1 and σ^2

Using a random sample of n observations $\hat{\tau}_{m1}, \hat{\tau}_{m2}, \cdots, \hat{\tau}_{mn}$ and the accompanying fixed values $x_1, x_2, \cdots, x_n$, we can estimate the parameters β_0, β_1 and σ^2. To obtain the estimates β_0 and β_1, we use the method of least squares, which does not require any distributional assumptions (for maximum likelihood estimators based on normality). That is, we estimate β_0 and β_1 so that the sum of the squares of the differences between the observations $\hat{\tau}_{mi}$ and the straight line is a minimum.

Thus the least-squares criterion is

$$S\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n}\left(\hat{\tau}_m - \frac{M}{n}\left(\beta_0 + \beta_1 x_i\right)\right)^2 \tag{9}$$

To find the values of β_0 and β_1 that minimize Equation (9), we differentiate with respect to β_0 and β_1, and set the results equal to 0:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(\hat{\tau}_m - \frac{M}{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = 0$$

and

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(\hat{\tau}_m - \frac{M}{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)x_i = 0$$

Simplifying Equations ((10) and (11)) yields,

$$\sum_{i=1}^{n}\hat{\tau}_m = M\hat{\beta}_0 + \hat{\beta}_1\frac{M}{n}\sum_{i=1}^{n}x_i \tag{10}$$

$$\sum_{i=1}^{n}\hat{\tau}_m x_i = \hat{\beta}_0\frac{M}{n}\sum_{i=1}^{n}x_i - \hat{\beta}_1\frac{M}{n}\sum_{i=1}^{n}x_i^2 \tag{11}$$

The solution to (10) and (11) is given by

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n}\hat{\tau}_m}{M} - \hat{\beta}_1\sum_{i=1}^{n}\frac{x_i}{n} \tag{12}$$

$$\hat{\beta}_1 = \frac{n}{M}\frac{\sum_{i=1}^{n}\hat{\tau}_m x_i}{\sum_{i=1}^{n}x_i^2} - \hat{\beta}_0\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2} \tag{13}$$

The method of least squares does not yield an estimator of $\mathrm{Var}\left(\hat{\tau}_m\right) = \sigma^2$; minimization of $S\left(\beta_0, \beta_1\right)$ yields only $\hat{\beta}_0$ and $\hat{\beta}_1$. To estimate σ^2, we use the definition

$$s^2 = \frac{SSE}{n-2}, \quad s^2 = \frac{\sum_i\left(\hat{\tau}_m - \frac{M}{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}{n-2}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by Equations ((12) and (13)) respectively.

4. Conclusion

The paper modelled the multiplicity estimator into a regression equation and proceeded to give estimates for β_0, β_1 and σ^2. To verify that $\hat{\beta}_0$ and $\hat{\beta}_1$ in (12) and (13) minimize $S(\beta_0, \beta_1)$ in (9), we can examine the second derivatives or simply observe that $S(\beta_0, \beta_1)$ has no maximum and therefore the first derivatives yield a minimum. Going forward, hypothesis testing and confidence interval for $\hat{\beta}_1$ can be derived.

Acknowledgements

We thank the editor and the referee for their comments. Appreciation also goes to Prof. S. K. Nokoe for looking through the initial work.

References

[1] Bimbaum, Z.W. and Sirken, M.G. (1965) Design of Sample Surveys to Estimate the Prevalence of Rare Diseases. Vital and Health Statistics. National Centre for Health Statistics, Washington DC.

[2] Sirken, M.G. (1977) Network Sampling. In: *Encyclopedia of Biostatistics*, John Wiley and Sons, New York, 2977-2985.

[3] Sirken, M.G. (1998) Network Sampling. In: *Encyclopedia of Biostatistics*, Volume 4, John Wiley and Sons, New York, 2977-2986.

[4] Kish, L. (1988) Multipurpose Sample Designs. *Survey Methodology*, **14**, 19-32.

[5] Kramm, E.R., Crane, M.M., Sirken, M.G. and Brown, M.L. (1962) A Cystic Fibrosis Pilot Study in Three New England States. *American Journal of Public Health*, **52**, 2041-2057. https://doi.org/10.2105/AJPH.52.12.2041

[6] Sirken, M. (2005) Network Sampling. Encyclopedia of Biostatistics, 5. https://doi.org/10.1002/0470011815.b2a16043

[7] Nathan, G. (1976) An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules. *Journal of the American Statistical Association*, **71**, 808-815. https://doi.org/10.1080/01621459.1976.10480951

[8] Sirken, M.G. (1972) Stratified Sample Surveys with Multiplicity. *Journal of the American Statistical Association*, **67**, 224-227. https://doi.org/10.1080/01621459.1972.10481236

[9] Sirken, M.G. (1972) Variance Components of Multiplicity Estimators. *Biometrics*, **28**, 869-873. https://doi.org/10.2307/2528769

[10] Levy, P.S. (1977) Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations. *Journal of the American Statistical Association*, **72**, 758-763. https://doi.org/10.1080/01621459.1977.10479952

[11] Sirken, M.G. and Levy, P.S. (1974) Multiplicity Estimation of Proportions Based on Ratios of Random Variables. *Journal of the American Statistical Association*, **69**, 68-73. https://doi.org/10.1080/01621459.1974.10480129

[12] Czaja, R.F., Snowdon, C.B. and Casady, R.J. (1986) Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules. *Journal of the American Statistical Association*, **81**, 411-419.

https://doi.org/10.1080/01621459.1986.10478285

[13] Mecatti, F. and Singh, A.C. (2014) Estimation in Multiple Frame Surveys: A Simplified and Unified Review Using the Multiplicity Approach. *Journal de la Société Française de Statistique*, **155**, 51-69.

[14] Singh, A. and Mecatti, F. (2009) A Generalized Multiplicity-Adjusted Horvitz Thompson Class of Multiple Frame Estimators. Book of Abstract. Siena, 75-77.

[15] Singh, A. and Mecatti, F. (2011) Generalized Multiplicity-Adjusted Horvitz-Thompson Type Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, **27**, 633-650.

[16] Mecatti, F. (2007) A Single Frame Multiplicity Estimator for Multiple Frame Surveys. *Survey Methodology*, **33**, 151-158.

[17] Laska, E.M., Meisner, M. and Wanderling, J. (2009) Model-Based Multiplicity Estimation of Population Size. *Statistics in Medicine*, **28**, 2230-2252. https://doi.org/10.1002/sim.3614

[18] Hing, E. and Burt, C.W. (2008) Characteristics of Office-Based Physicians and Their Medical Practices: United States, 2005-2006. *Vital & Health Statistics*, **13**, 1-34.

[19] Kohn, F., ZuWallack, R., and Griffin, R. (2001) Multiplicity Estimation for Service Based Enumeration in Census 2000. *Proceedings of the Annual Meeting of the American Statistical Association*, Washington DC, 5-9 August 2001.

[20] Johnston, L.G., *et al.* (2015) Estimating the Size of Hidden Populations Using Respondent-Driven Sampling Data: Case Examples from Morocco. *Epidemiology*, **26**, 846-852. https://doi.org/10.1097/EDE.0000000000000362

[21] Robertson, D.S., *et al.* (2016) Unbiased Estimation in Seamless Phase II/III Trials with Unequal Treatment Effect Variances and Hypothesis-Driven Selection Rules. *Statistics in Medicine*, **35**, 3907-3922. https://doi.org/10.1002/sim.6974