THE ANALYSIS OF SURVEY DATA

ALASTAIR SCOTT

Abstract. Gad Nathan was one of the first people to appreciate the problems that arise when conventional techniques for statistical analysis are applied to data from a complex survey. He did pioneering work in a number of areas, including regression and the analyisis of categorical data. In particular, it was this latter work that led to my own work with J.N.K. Rao on adapting chi-squared and likelihood-ratio tests for loglinear-models to survey data. In this paper, I discuss the origins of this work and its application to likelihood-ratio tests for more general models, as well as to analogues of model selection criteria such as AIC.

Keywords: Likelihood ratio tests; AIC; multistage sampling; survey weights; Rao-Scott tests

1. Introduction

The analysis of survey data has become very big business in recent years, driven in particular by public access to the results of large medical and social surveys such as the National Health and Nutrition Examination Surveys (NHANES) in the US or the British Household Panel Survey in the UK. To give just one indication of the extent of the scale of this, Google Scholar lists more than 34,000 papers containing both the words "NHANES" and "regression" in the abstract. Hundreds of similar (if mostly smaller) studies are being analyzed around the world every year.

Gad Nathan was one of the very first people to appreciate how important analysis was going to become, and to draw attention to the problems that arise when conventional techniques for statistical analysis are applied to data from a complex survey. He did pioneering work in a number of areas, including regression (Nathan & Holt, 1980) and the analyisis of categorical data Nathan, 1972). In particular, it was this latter work, and discussions with Gad on problems arising from the paper, that led Jon Rao and me to our work on adapting chi-squared and likelihood-ratio tests for loglinear-models to survey data.

In general, researchers who are analysing data sets like NHANES know what analysis they want to do — fit a (linear, logistic, Cox, ...) regression model etc — and would be able to implement it using a standard statistical package if the data had been collected through a simple random sample,. There are problems with the technical details of the analysis when the data is collected via a complex survey with varying selection probabilities and multi-stage sampling. However, the underlying population, and what researchers want to know about it, are not changed by the method of data collection. Most researchers still want to use the same techniques to answer

these questions as they would with a random sample and, in our experience, they want to implement them using programs that mimic familiar software as closely as possible.

Since those early discussions with Gad 35 years, much of this has become possible and all the main packages have survey versions for implementing standard techniques such as linear or logistic regression. There are still some useful quantities missing from these packages, however. The most notable of these are quantities related to likelihood — likelihood-ratio test statistics, deviances, AIC, BIC, etc. In this paper, we build on that original work with Jon Rao to fill these gaps. More specifically we show that the theory underlying the Rao—Scott tests for log-linear models with categorical data applies almost unchanged to likelihood ratio tests in general In addition, following the approach of Takeuchi[1976] for possibly-misspecified models, we show that AIC can be extended to enable comparison of models fitted with survey data by replacing p in the usual penalty term by the trace of the Rao–Scott design effect matrix.

2. Basic set-up

Suppose that we have observations $\{(y_i, \boldsymbol{x}_i); i \in s\}$ on a response variable, y, and a vector of possible explanatory variables, $\boldsymbol{x}$, from a sample, s, of n units drawn from a finite population or cohort of N units using some probability sampling design. Let π_i be the probability of selecting the ith unit with this design, with $w_i = 1/\pi_i$ the associated weight (perhaps adjusted to compensate for non-response and frame errors by, for example, calibration to known population totals).

We assume that the finite population values are generated independently from some distribution with density $g(y, \boldsymbol{x})$. This is much less restrictive than it might appear at first sight: we can generate populations with very complex spatial correlation structures by, for example, measuring extra variables such as latitude and longitude and sorting on them (see Lumley & Scott, 2013, for a more detailed discussion).

Suppose that, after plotting the data and carrying out other preliminary investigations, we decide that we want to fit a parametric model, $f(y \mid \boldsymbol{x}; \boldsymbol{\theta})$, for the marginal conditional density of y given $\boldsymbol{x}$. We do not assume that the true model is a member of this parametric family. It follows from standard work on AIC (see Chapter 2 in Claeskens & Hjort, 2008, for example) that the best fitting model in our class, in the sense of minimizing the Kullback-Leibler distance between it and the super-population model $g(\cdot)$, is obtained by setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ satisfies

the superpopulation score equation

$$\boldsymbol{U}(\boldsymbol{\theta}) = E_g\left\{\frac{\partial \log f(y \mid \boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} = \mathbf{0}. \tag{1}$$

Since $\boldsymbol{U}(\boldsymbol{\theta})$ is just a vector of population means for any fixed value of $\boldsymbol{\theta}$, we can estimate it from our sample. Let

$$\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i\in s} w_i \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \text{ say,} \tag{2}$$

where $\ell_i(\boldsymbol{\theta}) = \log f(y_i \mid \boldsymbol{x}_i;\boldsymbol{\theta})$, be the Horvitz-Thompson estimator of $\boldsymbol{U}(\boldsymbol{\theta})$ and let $\widehat{\boldsymbol{\theta}}_n$ be the value we obtain by setting $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}_n)$ equal to $\mathbf{0}$. Then, under mild conditions, $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}^*$ as $n, N \to \infty$. This is the basis of the approach developed by Fuller (1975) for linear regression and by Binder (1983) for more general regression models. It is the approach underlying all the major statistical packages for survey analysis and the one that we shall adopt here.

We shall assume the asymptotic setting and regularity conditions of Th 1.3.9 in Fuller (2009). We have a sequence of finite populations assumed to be random samples from a fixed super population. As we noted above, this is much less restrictive than it might sound. The regularity conditions impose restrictions on the superpopulation (finite fourth moments), on the sequence of sampling designs (a central limit theorem for Horvitz-Thompson estimators) and on the estimating functions $\mathbf{U}_i$ (continuous second derivatives). Let $\widehat{\boldsymbol{\mathcal{J}}}(\boldsymbol{\theta})$ be the analogue of the observed information matrix defined by

$$\widehat{\boldsymbol{\mathcal{J}}}(\boldsymbol{\theta}) = -\frac{\partial \widehat{\mathbf{U}}}{\partial \boldsymbol{\theta}^T} = -\frac{1}{N}\sum_{i\in s} w_i \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$$

and set $\boldsymbol{\mathcal{I}} = E\left\{\widehat{\boldsymbol{\mathcal{J}}}\right\}$. Then it follows from the theorem above that

$$\mathbf{V}(\boldsymbol{\theta}^*)^{-\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}),$$

where $\mathbf{V}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}^{-1} Cov\{\widehat{\boldsymbol{U}}\}\boldsymbol{\mathcal{I}}^{-1}$, as $n, N \to \infty$. We can estimate $\mathbf{V}(\boldsymbol{\theta})$ by $\widehat{\mathbf{V}} = \widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}}_n)^{-1}\widehat{\mathbf{V}}_U(\widehat{\boldsymbol{\theta}}_n)\widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}}_n)^{-1}$ where $\widehat{\mathbf{V}}_U(\boldsymbol{\theta})$ is an estimate of $Cov\left\{\widehat{\boldsymbol{U}}\right\}$.

Note that $\boldsymbol{\theta}^*$ could be replaced by $\boldsymbol{\theta}_N$, the solution of the finite population score equation

$$\boldsymbol{U}_{\text{pop}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

in the previous paragraph. The only change needed is in the interpretation of $Cov\{\widehat{\boldsymbol{U}}\}$. Now it would mean the expected value of the covariance under the distribution generated by repeated sampling from a fixed finite population rather than the covariance over the combined operation of drawing finite population values from the superpopulation and then drawing a sample from the resulting finite population.

In Lumley & Scott(2012) we used this set-up to construct an analogue of the likelihood ratio test, based on $\widetilde{\ell}(\boldsymbol{\theta}) = n\widehat{\ell}(\boldsymbol{\theta}) = \frac{n}{N}\sum_{i\in s} w_i\ell_i(\boldsymbol{\theta}))$, that has many of the properties of an ordinary likelihood ratio test. (We multiply by n to ensure that we get the same value as we would with a standard regression program when we have a simple random sample with weights $w_i = N/n$.) Write $\boldsymbol{\theta}$ in the form $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$, where $\boldsymbol{\theta}_2$ is $q \times 1$, and suppose that we are interested in testing the hypothesis $H_0 : \boldsymbol{\theta}_2^* = \boldsymbol{\theta}_{20}$. Let $\widehat{\boldsymbol{\theta}}_0$ be the solution of $\widehat{\boldsymbol{U}}_1(\boldsymbol{\theta}_0) = \mathbf{0}$, where $\boldsymbol{\theta}_0 = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_{20} \end{pmatrix}$ and $\widehat{\boldsymbol{U}}_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i\in s} w_i\ \partial\ell_i/\partial\boldsymbol{\theta}_1$. Then our pseudo likelihood-ratio test statistic is given by

$$\Lambda = 2\left\{\widetilde{\ell}(\widehat{\boldsymbol{\theta}}_n) - \widetilde{\ell}(\widehat{\boldsymbol{\theta}}_0)\right\}. \tag{3}$$

If the regularity conditions of Th 1.3.9 in Fuller (2009) are satisfied and $H_0 : \boldsymbol{\theta}_2^* = \boldsymbol{\theta}_{20}$ is true, then we show that

$$\Lambda = 2\{\widetilde{\ell}(\widehat{\boldsymbol{\theta}}_n) - \widetilde{\ell}(\widehat{\boldsymbol{\theta}}_0)\} \sim \sum_1^q \delta_i Z_i^2,$$

where $Z_1, \ldots, Z_q$ are independent $N(0,1)$ random variables and $\delta_1, \ldots, \delta_q$ are the eigenvalues of $\boldsymbol{\Delta} = n\left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}\right)\mathbf{V}_{11}$ where $\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \boldsymbol{\mathcal{I}}_{21} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix}$ and $\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$. If our sample had been a random sample from the superpopulation, then $n\mathbf{V}$ would be equal to $\boldsymbol{\mathcal{I}}^{-1}$. Using the standard form for the inverse of a partitioned matrix, it follows that $n\mathbf{V}_{11}$ would be equal to $\left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}\right)^{-1} = \mathbf{V}_{11}^{(0)}$, say. Thus we can write the matrix $\boldsymbol{\Delta}$ in the form $\boldsymbol{\Delta} = \mathbf{V}_{11}^{(0)-1}\mathbf{V}_{\mathbf{11}}$. By analogy with the simple scalar case, we call $\boldsymbol{\Delta}$ the "design-effect matrix" and the eigenvalues, $\delta_1, \ldots, \delta_q$, "generalized design effects", as in Rao & Scott (1981, 1984).

In the next section, we build on all this to construct an analogue of AIC for use with survey data.

3. AIC

Our development follows that for random sampling in Claeskens & Hjort (2008). The appropriately-weighted Kullback-Leibler distance of $f(\widehat{\boldsymbol{\theta}}_n)$ from the true model is

$$\begin{aligned} KL(g(\cdot \mid \boldsymbol{x}), f(\cdot \mid \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_n) &= \int\int \log \frac{g(y \mid \boldsymbol{x})}{f(y \mid \boldsymbol{x}; \widehat{\boldsymbol{\theta}}_n)}\ g(y, \boldsymbol{x}) dy d\boldsymbol{x} \\ &= \int\int \log g(y \mid \boldsymbol{x})\ g(y, \boldsymbol{x}) dy d\boldsymbol{x} - \ell(\widehat{\boldsymbol{\theta}}_n), \end{aligned}$$

with $\ell(\boldsymbol{\theta}) = E_g\{\log f(y \mid \boldsymbol{x}; \boldsymbol{\theta})\}$. The first term is the same across all models so we are interested in $\ell(\widehat{\boldsymbol{\theta}}_n)$, which is a random variable through its dependence on $\widehat{\boldsymbol{\theta}}_n$. The AIC strategy is to estimate $Q_n = E_g\{\ell(\widehat{\boldsymbol{\theta}}_n)\}$ for each candidate model and then select the model with the largest value of Q_n. This is equivalent to searching for the model with the smallest estimated Kullback-Leibler distance from the true model.

A naive first estimator of Q_n would be $\widehat{\ell}(\widehat{\boldsymbol{\theta}}_n)$ where

$$\widehat{\ell}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i \in s} w_i \log f(y_i \mid \boldsymbol{x}_i : \boldsymbol{\theta}) = \frac{1}{n}\widehat{\ell}(\boldsymbol{\theta}),$$

as in the previous section. This turns out to be an overestimate. More precisely,

$$E_g\{\widehat{\ell}(\widehat{\boldsymbol{\theta}}_n) - \ell(\widehat{\boldsymbol{\theta}}_n)\} = \frac{1}{n} tr\{\boldsymbol{\Delta}_M\} + o_p(n^{-1}), \tag{4}$$

where $\boldsymbol{\Delta}_M = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)\mathbf{V}(\boldsymbol{\theta}^*) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}\mathbf{V}_U(\boldsymbol{\theta}^*)$. A sketch of the proof is given in the appendix. This result leads to $\widehat{\ell}(\widehat{\boldsymbol{\theta}}_n) - tr\{\boldsymbol{\Delta}_M\}/n = n^{-1}\left[\widehat{\ell}(\widetilde{\boldsymbol{\theta}}) - tr\{\boldsymbol{\Delta}_M\right]$ as a bias-corrected estimate. Following the usual practice for independent sampling, we multiply by $2n$ to obtain

$$AIC_W = 2\widetilde{\ell}(\widehat{\boldsymbol{\theta}}_n) - 2tr\{\boldsymbol{\Delta}\}$$

as our modified version of AIC for survey data. (Note that the design effect matrix $\boldsymbol{\Delta}_M$ depends on the model being fitted.) Under simple random sampling, where the weights are constant, AIC_W reduces to Takeuchi's robust version of AIC. If, in addition, our class $f(y \mid \boldsymbol{x} : \boldsymbol{\theta})$ contains the true model $g(y \mid \boldsymbol{x})$, then $\boldsymbol{\Delta}$ is the $p \times p$ identity matrix and we get the conventional expression for AIC.

References

[1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 29–34.

[2] Chambers, R.L. and Skinner, C.J. (eds) (2003) *Analysis of Survey Data.* New York: Wiley.

[3] Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhya C*, **37**, 117-132.

[4] Fuller, W. (2009). *Sampling Statistics*. New York: John Wiley and Sons.

[5] Lumley, T. (2013). "survey: analysis of complex survey samples". R package version 3.29-3. http://cran.r-project.org/package=survey

[6] Lumley, T. and Scott, A.J. (2012). Partial likelihood ratio tests for the Cox model under complex sampling. *Statistics in Medicine*, **31**, 409-427.

[7] Lumley, T. and Scott, A.J. (2013). Fitting GLMs with survey data. 5174-5181 *Proceedings of the Survey Research Methods Section, Amer. Statist. Assoc,*, 5174–5181.

[8] Nathan, G. (1972). On the asymptotic power of tests for independence in contingency tables from stratified samples. *Journal of the American Statistical Association*, **67**, 917–920.

[9] Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, B*, **42**, 377–386.

[10] Rao, J.N.K., and Scott, A.J. (1981), The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables *Journal of the American Statistical Association*, **76**, 221–230.

[11] Rao, J.N.K., and Scott, A.J. (1984), On chi-squared tests for multi-way tables with cell proportions estimated from survey data *Annals of Statistics*, **12**, 46–60.

[12] Skinner, C.J., Holt, D., and Smith, T.M.F. (eds) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

[13] Takeuchi,K. (1976). Distribution of informational statistics and a criterion of model fitting. *SuriKagaky* **153**, 12–18. In Japanese.

Department of Statistics, University of Auckland, Auckland, NZ

E-mail address: a.scott@auckland.ac.nz