Extensions of Rao-Scott tests: fitting GLMs with Survey Data

Thomas Lumley and Alastair Scott
Department of Statistics
The University of Auckland,
Private Bag 92019, Auckland, New Zealand
September 29, 2012

Abstract

Data from complex surveys are being used increasingly to build the same sort of explanatory and predictive models used in the rest of statistics. Unfortunately the assumptions underlying standard statistical methods are not even approximately valid for most survey data. The problem of parameter estimation has been largely solved through the use of weighted estimating equations, and software for most standard statistical procedures is now available in the major statistical packages. With one notable exception, a big gap in the output from these packages is an analogue of the likelihood ratio test and related quantities like AIC. The exception is the Rao-Scott test for loglinear models in contingency tables. It turns out to be straightforward to extend this test to many other situations, in particular to Generalized Linear Models. We show that the asymptotic null distribution of a natural analogue of the likelihood-ratio statistic is a linear combination of chi-squared random variables whose coefficients are eigenvalues of a matrix product that does not involve the inverse of the estimated covariance matrix.

Key words: Likelihood-ratio tests; complex sampling; statistical computing.

1. Introduction

Traditional sample survey methods were developed primarily for the estimation of descriptive statistics like means, proportions and totals, rather than for analysis. However, the analysis of survey data has become big business in recent years, driven in particular by public access to the results of large medical and social surveys such as NHANES in the US, the British Household Panel Survey in the UK, or HILDA in Australia. To give just one indication of the extent of the literature, GoogleScholar lists more than 32,000 papers with "NHANES" and "regression" in the abstract, and almost 2,000 with "NHANES" and "generalized linear models". Hundreds of similar (albeit mostly smaller) studies are being analyzed around the world every year. Fortunately, most of the traditional work on descriptive statistics can be used directly to develop methods for analysis.

What do the researchers who are analysing data sets like NHANES want from their analysis? If the data had been collected through a simple random sample, there would be no problem − they would simply use a standard statistical package to carry out whatever analysis they thought appropriate to answer the question of interest − fit a (linear, logistic, Cox) regression model, and so on. There 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. Thus, most researchers still want to use the same techniques to answer these questions. Moreover, not only do researchers want to use the same techniques, they want to implement them using programs that mimic familiar software as closely as possible.

After a lot of work by many people over the last 25 years or so, much of this is now possible. In particular, we would like to pay tribute to the contributions made by David

Binder and Gad Nathan, both of whom died very recently. We focus on fitting Generalized Linear Models (GLMs) in this paper. Our general aim is to summarize what has been done, to point to gaps where more needs to be done, and to make a start on filling some of those gaps. In particular, we extend the tests developed by Rao & Scott (1981, 1984) for loglinear models to arbitrary Generalized Linear Models.

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, $\boldsymbol{x}$, where s is a sample 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 calibrated to known population totals to compensate for non-response and frame errors). 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 $\boldsymbol{x}$. (Note that plotting survey data is not at all straightforward but good routines for this are now available - see Chapter 4 in Lumley, 2010, for a detailed description.)

If the sample was a simple random one, we would probably fit the model by maximum likelihood and obtain our estimate, $\widehat{\boldsymbol{\theta}}$ say, by solving the likelihood equations,

$$\mathbf{U}_{\rm srs}(\boldsymbol{\theta}) = \sum_{i \in s} \mathbf{U}_i(\boldsymbol{\theta}) = \sum_{i \in s} \frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

where $\ell_i = \ell_i(\boldsymbol{\theta}) = \log\{f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\}$. Under mild regularity conditions, this would be a consistent estimator of the solution, $\boldsymbol{\theta}_{\rm pop}$ say, of the population (or census) estimating equation

$$\mathbf{U}_{\rm pop}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1}$$

If $\boldsymbol{\theta}_{\rm pop}$ is a sensible population quantity to be estimating when we have a simple random sample, then it must be equally sensible when we have a more complex sample from the same population. The important question is thus "Is $\boldsymbol{\theta}_{\rm pop}$ a sensible thing to be estimating in the first place?"

Clearly $\boldsymbol{\theta}_{\rm pop}$ would be a natural target if we believed that $\{y_1, y_2, \ldots, \boldsymbol{y}_N\}$ could be modelled sensibly as a sequence of independent observations with densities $f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ for some arbitrary sequence $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. Most real finite populations will have a much more complex covariance structure than this. However, as suggested by Binder (1983), the solution of equation (1) will give a consistent estimator of $\boldsymbol{\theta}$ under much more realistic assumptions (basically provided the finite population is selected from some superpopulation with the right marginal structure in such a way that $E\{\mathbf{U}_{\rm pop}(\boldsymbol{\theta})\} = \mathbf{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\rm pop}$, cf using estimating equations with a "working independence model" for longitudinal data as in Liang & Zeger, 1986). If we think of $\boldsymbol{\theta}_{\rm pop}$ as estimating the regression parameter in some superpopulation, then we could get a more efficient estimate by using a more realistic working covariance matrix. However, as shown by Pepe and Anderson (1994) and Pan, Louis and Connett (2000), there are real dangers when we move away from a diagonal working covariance structure. For example, we need to model the expected value of $\boldsymbol{y}_i$ given the covariates of units whose response is correlated with $\boldsymbol{y}_i$ and to model it correctly.

To avoid such complications, it seems sensible to stick with the simple, robust working independence structure for general purpose software. This is particularly apposite for surveys like NHANES where samples are large and any squared bias term will dominate the mean squared error.

We still need to think about the role of the finite population parameter $\boldsymbol{\theta}_{\rm pop}$. Some traditional survey statisticians (see Kish & Frankel, 1974, for example) would regard estimating such finite population quantities as the ultimate objective of any inference, but it seems to us that researchers fitting a model almost always intend the results to be used well beyond the particular finite population from which the sample was drawn.

What do the researchers themselves want? In reality, we suspect that few of them give much thought to the question. Richard Peto is one of the few who writes explicitly about it. In Yusuf, Collins, & Peto (1984), he and his co-authors say:
"A key principle underlying the argument that (such studies) can provide medically relevant conclusions involves careful distinction between 'quantitative' and 'qualitative' interactions. A qualitative interaction is one where true treatment effects in different subgroups do not even point in the same direction, whereas a quantitative interaction is one whereby the direction of the treatment effect is similar..... Our expectation is not that *all* qualitative interactions are unlikely, but merely that *unanticipated* qualitative interactions are unlikely ..."

This is closer to a superpopulation approach than the traditional finite population framework, but perhaps a random effects superpopulation model in which the parameters vary among populations. The difference between estimating a finite population parameter and a fixed superpopulation parameter vanishes in large populations. This is not true if our parameter varies over populations. Unfortunately we only have observations from one possible population so there is no chance of estimating the between-population variance component in such cases. Hence we are always going to have an under-estimate of the true mean squared error, no matter how large our population or sample.

3. Some Formalities

Return to the problem of estimating $\boldsymbol{\theta}_{\rm pop}$, defined as the solution of (1):

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

For any fixed value of $\boldsymbol{\theta}$, the population score, $\boldsymbol{U}_{\rm pop}(\boldsymbol{\theta})$, is just a vector of population totals. Thus, since estimating population totals is something that traditional survey theory does well, we can estimate $\boldsymbol{U}_{pop}(\boldsymbol{\theta})$ from our sample.

Setting $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}})$ equal to $\mathbf{0}$, where

$$\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \sum_{sample} w_i \mathbf{U}_i = \sum_{sample} w_i \frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \tag{2}$$

is the Horvitz-Thompson estimator of $\boldsymbol{U}_{pop}(\boldsymbol{\theta})$, then gives us the estimator $\widehat{\boldsymbol{\theta}}$ (sometimes called the pseudo-MLE).

It is well-known that weighting can be very inefficient if the weights vary widely. However, unweighted estimates are biased unless the weights are uncorrelated with the residuals

from the fitted model, as would be the case if all the design variables used in forming the selection probabilities are included as covariates in the regression model. Some authors have suggested that we could ensure this by including the vector of weights as extra covariates. However, this may distort the meaning of the coefficients of importance. Using weighted estimates seems the most foolproof default setting for a general purpose program, while including more flexible options for more sophisticated users.

We shall assume the asymptotic setting and regularity conditions of Th 1.3.9 in Fuller (2009). Then it follows that

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

where $\mathbf{V}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}^{-1} Cov\{\widehat{\boldsymbol{U}}\} \boldsymbol{\mathcal{I}}^{-1}$ with $\boldsymbol{\mathcal{I}} = E\left\{\widehat{\boldsymbol{\mathcal{J}}}\right\}$. Here $\widehat{\boldsymbol{\mathcal{J}}}$ is defined by

$$\widehat{\boldsymbol{\mathcal{J}}} = \widehat{\boldsymbol{\mathcal{J}}}(\boldsymbol{\theta}) = -\frac{\partial \widehat{\mathbf{U}}}{\partial \boldsymbol{\theta}^T} = -\sum_{sample} w_i \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$

the analogue of the observed information matrix.

We can estimate $\mathbf{V}(\boldsymbol{\theta})$ by $\widehat{\mathbf{V}} = \widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})^{-1} \widehat{\mathbf{V}}_U(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})^{-1}$ where $\widehat{\mathbf{V}}_U(\boldsymbol{\theta})$ is an estimate of $Cov\left\{\widehat{\boldsymbol{U}}\right\}$. The form of $\widehat{\mathbf{V}}_U(\boldsymbol{\theta})$ will depend on the framework of inference being adopted. The default option in survey (and most other packages) is based on the pretence that PSUs are sampled independently within strata, recognising that this inevitably still gives an underestimate of the true mean squared error when differences among populations are taken into account. Different possibilities are available as options for people who have different (clear) ideas about what they want.

Most statistical packages now include procedures for fitting standard statistical models to survey data using this approach, which was first suggested by Wayne Fuller (Fuller, 1975) for linear regression and David Binder (Binder, 1983) for more general regression models. All the major packages have survey routines for linear and logistic regression and for fitting log-linear models to contingency tables. Survey and Stata can handle arbitrary Generalized Linear Models.

So what more needs to be done?

If we compare `svy:glm` with `glm` in Stata or `svyglm` with `glm` in the R package survey, we see that they are very similar, apart from the need to set up a frame containing all the design information at the beginning of any analysis involving survey data. The main exceptions are outputs related to likelihoods − likelihood-ratio tests, deviances, AIC, BIC, etc. We look at analogues of the likelihood-ratio test in the next section.

As a cautionary note, we observe that some programs (`proc surveylogist` in SAS, for example) do have entries labelled "likelihood-ratio", "AIC", "BIC", etc but these are simply artifacts of converting the equivalent non-survey program and have no valid statistical meaning.

4. Pseudo Likelihood-Ratio Tests

It is easy to extend the Fuller-Binder approach to test hypotheses about a p-dimensional subvector, say $\boldsymbol{\theta}_1$, of the parameters, or to produce confidence regions for $\boldsymbol{\theta}_1$, via the Wald statistic,

$$W = (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)^T \widehat{\mathbf{V}}_1^{-1} (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1),$$

which has an asymptotic χ^2_p distribution under the assumptions of the previous section.

This has the usual problems associated with the Wald test. For example, the statistic is not invariant under nonlinear transformations of the parameter, the tests often have poor small sample behavior, the confidence regions may contain invalid values of the parameter, etc. There is an additional, potentially more serious, problem with survey data: even in very large surveys, the degrees of freedom for the estimated covariance matrix are often very small, depending on the number of primary sampling units rather than the number of observations. For example, a two-year cycle of the current continuous NHANES survey has approximately 15 degrees of freedom. The estimated covariance matrix can be singular or nearly so, if p is large. Even in less extreme cases, the estimate often has high variance and its inverse tends to be very unstable.

Ideally we would prefer to use a likelihood ratio test which is invariant and usually has better small sample properties. Although there is no natural likelihood function for survey data, it is possible to construct a pseudo likelihood that has many of the same properties.

Write $\boldsymbol{\theta}$ in the form $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$ and suppose that we are interested in testing the hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$. Let $\widehat{\boldsymbol{\theta}}_0$ be the solution of $\widehat{\boldsymbol{U}}_2(\boldsymbol{\theta}_0) = \mathbf{0}$, where $\boldsymbol{\theta}_0 = \begin{pmatrix} \boldsymbol{\theta}_{10} \\ \boldsymbol{\theta}_2 \end{pmatrix}$ and

$$\widehat{\boldsymbol{U}}_2(\boldsymbol{\theta}) = \sum_{sample} w_i \frac{\partial \ell_i}{\partial \boldsymbol{\theta}_2}. \tag{3}$$

Then our pseudo likelihood-ratio test statistic is given by

$$\Lambda = 2\left\{\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}_0)\right\}$$

with $\widehat{\ell}(\boldsymbol{\theta}) = \sum_{samp} w_i \ell_i(\boldsymbol{\theta})$.

The asymptotic distribution of Λ is given in the theorem below. The main steps in the derivation, which are very similar to those for establishing the asymptotic distribution of likelihood-ratio statistic in the classical i.i.d. setting, are outlined in the appendix

Theorem:
If the regularity conditions of Th 1.3.9 in Fuller (2009) are satisfied then, under $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$,

$$\Lambda = 2\{\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}_0)\} \sim \sum_1^p \lambda_i Z_i^2,$$

where $Z_1, \ldots, Z_p$ are independent $N(0,1)$ random variables and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{D} = \left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}\right)\mathbf{V}_1$ with $\mathbf{V}_1 = ACov\left\{\widehat{\boldsymbol{\theta}}_1\right\}$ and

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \boldsymbol{\mathcal{I}}_{21} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix}.$$

□

Recall that, if our sample had been a random sample from a superpopulation, then $ACov\{\widehat{\boldsymbol{\theta}}\}$ would be equal to $\boldsymbol{\mathcal{I}}^{-1}$. Using the standard form for the inverse of a partitioned matrix, it follows that $ACov\{\widehat{\boldsymbol{\theta}_1}\}$ would be equal to

$$\left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}^{-1}\right)^{-1} = \mathbf{V_{01}}, \text{ say.}$$

Thus we can write the matrix $\mathbf{D}$ in the form $\mathbf{D} = \mathbf{V_{01}}^{-1}\mathbf{V_1}$. By analogy with the simple scalar case, we call $\mathbf{D}$ the "design-effect matrix" and the eigenvalues, $\lambda_1, \ldots, \lambda_p$, "generalised design effects".

In the special case of log-linear models for contingency tables, the results are well-known (Rao & Scott, 1981, 1984) and already included in most of the major packages. The results above show that the Rao-Scott results apply almost unchanged to arbitrary Generalized Linear Models and, more generally, to any model fitted via a census estimating equation of the form (1).

5. Evaluating percentage points of the null distribution

If we knew the values of $\lambda_1, \ldots, \lambda_p$, there are good routines for getting accurate percentage points of the asymptotic distribution and hence calculate p-values, confidence limits, etc). The usual approximation in survey statistics is a Satterthwaite approximation that matches the mean and variance of the distribution to a scaled χ^2 distribution. In many large surveys, this gives an adequate approximation as long as we do not venture too far out into the tails. More accurate approximations can be obtained by numerically integrating the characteristic function (Davies, 1990) or summing an infinite series of F percentiles (Farebrother, 1984). A saddlepoint approximation (Kuonen, 1999), which gives higher accuracy than the Satterthwaite approximation with easier implementation than the integration and infinite-series approaches, is also available.

Of course, we do not know the values of $\lambda_1, \ldots, \lambda_p$ but, since an estimate of $\mathbf{V}_1 = ACov\left\{\widehat{\boldsymbol{\theta}}_1\right\}$ is available routinely, we can obtain estimates of $\lambda_1, \ldots, \lambda_p$. Unfortunately, as we have noted already, estimates of $\mathbf{V}_1$, and hence of quantities desived from such estimates, tend to be extremely variable even in very large surveys and we need to take this into account. Treating $\Lambda / \sum_1^p \widehat{\lambda}_i$ as an F-statistic with $\nu_1 = p/(1+c^2)$, where c is the CV of the $\lambda_i s$, and $\nu_2 = k\nu_1$ d.f, where k is the d.f. of the variance estimate, has been shown to work well in the special case of log-linear models for contingency tables (see Rao & Thomas, 2003) and should work equally well in more general cases. Similar approximations are straightforward for the saddlepoint and characteristic function approaches, since a linear combination of F variables can be transformed to a linear combination of χ^2, as long as negative multipliers are allowed. Simulations in Lumley & Scott (2012) suggest that these work well for designs like NHANES.

6. Conclusion

The development of natural analogues of likelihood ratio tests in this paper applies to arbitrary regression models and, more generally, to tests for any parameter defined through a census estimating equation of the form in (1). We can also use very similar methods to develop tests for other situations such as partial likelihood ratio tests for proportional hazards models fitted to survey data. Details can be found in Lumley & Scott (2012).

An important gap still left unfilled is the development of analogues of AIC and BIC for survey data. Some results for BIC are given in Fabrizi & Lahiri (2004) but much work still needs to be done.

APPENDIX

Sketch of proof of Theorem

As in Section 3, we assume that our sequence of sampling designs is such that $\mathbf{V}(\boldsymbol{\theta})^{-\frac{1}{2}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I})$ and $\widehat{\boldsymbol{\mathcal{J}}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) + o_p(1)$ as $n, N \to \infty$. We need a preliminary result on the relationship between the full pseudo-MLE $\widehat{\boldsymbol{\theta}}$ and the restricted estimator $\widehat{\boldsymbol{\theta}}_0$.

Lemma

If $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ then $\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0 = \mathbf{A}(\widehat{\theta}_1 - \boldsymbol{\theta}_{10}) + o_p(n^{-1/2})$ with $\mathbf{A} = \begin{pmatrix} \mathbf{I_p} \\ \boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21} \end{pmatrix}$, where $\mathbf{I_p}$ is the $p \times p$ identity matrix.

Proof. We first expand $\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \partial\widehat{\ell}/\partial\boldsymbol{\theta}$ about $\boldsymbol{\theta}_0$, recalling that $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ and that $-\partial\widehat{\boldsymbol{U}}/\partial\boldsymbol{\theta}^T = \widehat{\boldsymbol{\mathcal{J}}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) + o_p(n)$:

$$\mathbf{0} = \widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{U}}(\boldsymbol{\theta}_0) - \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(n^{1/2}).$$

In particular, if $\boldsymbol{U}_2 = \partial\widehat{\ell}/\partial\boldsymbol{\theta}_2$ is the second component of $\widehat{\boldsymbol{U}}$, then

$$\boldsymbol{U}_2(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{I}}_{22}(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) + \boldsymbol{\mathcal{I}}_{21}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) + o_p(n^{1/2})$$

Similarly, expanding $\boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}_0)$ about $\boldsymbol{\theta}_0$ and assuming that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$:

$$\mathbf{0} = \boldsymbol{U}_2(\widehat{\boldsymbol{\theta}}_0) = \boldsymbol{U}_2(\boldsymbol{\theta}_0) - \boldsymbol{\mathcal{I}}_{22}(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2) + o_p(n^{1/2}).$$

Combining the two expressions for $\boldsymbol{U}_2(\boldsymbol{\theta}_0)$ leads to $\widehat{\boldsymbol{\theta}}_2 - \widehat{\boldsymbol{\theta}}_{20} = \boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}(\widehat{\theta}_1 - \boldsymbol{\theta}_{10}) + o_p(n^{-1/2})$ when $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$. The lemma then follows immediately. □

Theorem

If $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$, then $\Lambda = -2\left[\widehat{\ell}(\widehat{\boldsymbol{\theta}}_0) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})\right] \sim \sum_1^p \lambda_i Z_i^2$, where $Z_1^2, \ldots, Z_p^2$ are independent χ_1^2 random variables and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\left(\boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}\right)\mathbf{V}_{11}$

Proof. Expand $\widehat{\ell}(\widehat{\boldsymbol{\theta}}_0)$ about $\widehat{\boldsymbol{\theta}}$, noting that $\partial\widehat{\ell}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$:

$$\widehat{\ell}(\widehat{\boldsymbol{\theta}}_0) = \widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}})^T\widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}) + o_p(n||(\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}})||^2)$$

Thus, using the results of the lemma,

$$D_W = -2[\widehat{\ell}(\widehat{\boldsymbol{\theta}}_0) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})] = (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^T\mathbf{A}^T\boldsymbol{\mathcal{I}}\mathbf{A}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) + o_p(1).$$

Note that $\mathbf{A}^T\boldsymbol{\mathcal{I}}\mathbf{A} = \boldsymbol{\mathcal{I}}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}$. The theorem then follows from standard results on quadratic forms of asymptotically normal random variables. □

References

[1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

[2] Davies, R.B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J. Roy. Statist. Soc, C,* **29**, 323–333.

[3] Fabrizi, E. and Lahiri, P. (2004) A design based approximation to BIC in finite population sampling. University of Maryland Working Paper.

[4] Farebrother, R.W. (1984). Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. *J. Roy. Statist. Soc, C,* **33**, 332–339.

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

[6] Fuller, W. (2009). *Sampling Statistics*. New York: Wiley.

[7] Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *J. R. Statist. Soc. B*, **36**, 1–37.

[8] Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, **86**, 929–935.

[9] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis esing generalized linear models. *Biometrika*, **73**, 13–22.

[10] Lumley, T.S. (2010). *Surveys: A Guide to Analysis Using R.* New York: Wiley.

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

[12] Pan, W. Louis, T.A. and Connett, J.E. (2000). A note on marginal linear regression with correlated response data. *The American statistician*, **54**, 191–195.

[13] Pepe, M.S. and Anderson (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics:Simulation and Communication* **23**, 939–951

[14] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 317–337.

[15] 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.

[16] 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.

[17] Thomas, D.R. and Rao, J.N.K.(1987), Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling *Journal of the American Statistical Association*, **82**, 630–636.

[18] Yusuf, S., Collins, R and Peto, R. (1984). Why do we need some large simple randomized trials? *Statistics in Medicine*, **3**, 409-420.