Journal of Survey Statistics and Methodology

AIC AND BIC FOR MODELING WITH COMPLEX SURVEY DATA

THOMAS LUMLEY*
ALASTAIR SCOTT

Model-selection criteria such as AIC and BIC are widely used in applied statistics. In recent years, there has been a huge increase in modeling data from large complex surveys, and a resulting demand for versions of AIC and BIC that are valid under complex sampling. In this paper, we show how both criteria can be modified to handle complex samples. We illustrate with two examples, the first using data from NHANES and the second using data from a case–control study.

KEY WORDS: AIC; BIC; Kullback–Leibler divergence; Model selection; Pseudo-likelihood.

1. INTRODUCTION

The analysis of survey data has expanded enormously 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. Researchers analyzing such data sets usually have a clear idea of the questions they want answered and would be able to carry out an appropriate analysis if the data had been selected through a simple random sample. There are problems with the technical details of the analysis when the data are collected via a complex survey with varying selection probabilities and multistage sampling. However, the underlying population, and what researchers want to know about it, are not changed by the method of data collection. Moreover, most researchers still want to use the same techniques that they would use with a random sample to answer these questions and, in our experience, they want to implement them using programs that mimic familiar software as closely as possible.

Thomas Lumley is Chair in Biostatistics, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Alastair Scott is Professor, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand.
*Address correspondence to Alastair Scott; E-mail: a.scott@auckland.ac.nz.

doi: 10.1093/jssam/smu021

After a lot of work by many people over the past 25 years or so, much of this is now possible. All the main statistics packages have survey versions for implementing standard techniques such as linear or logistic regression, and some can handle arbitrary generalized linear models. There are still some widely used quantities missing from these packages, however. Among the most notable of these are standard criteria for model selection such as AIC (Akaike 1974) and BIC (Schwarz 1978). We note that there are a large number of published analyses of survey data quoting so-called AIC and BIC values—several thousand just for NHANES or the British Household Panel Survey, for example. As far as we are aware, there is nothing in the current literature to justify any of these. However, the existence of so much literature does suggest that there is a strong desire from subject-matter researchers for versions of standard model-selection criteria that could be used correctly with survey data.

In this paper, we develop principled survey analogues of AIC and BIC for fixed-effects regression models fitted using pseudo-likelihood methods. More specifically, we show that, following the approach of Takeuchi (Takeuchi 1976; Claeskens and Hjort 2008) for possibly misspecified models, AIC can be modified by inflating the penalty term by a design effect related to the Rao–Scott correction for log-linear models (Rao and Scott 1984). We also show how BIC can be extended using a Bayesian coarsening argument, where the point estimates under complex sampling are treated as the data available for Bayesian modeling. In the special case of choosing between submodels of a given regression model, the Laplace approximation argument used to construct the usual expression for BIC leads to a natural survey analogue. We conclude with two examples illustrating the end result with two very different sampling designs. In the first example, we investigate the effect of sodium consumption on hypertension using data from NHANES. In the second example, we look at the effects of alcohol and tobacco use on esophageal cancer using data from a well-known case–control study. We compare the results with those obtained with two ad hoc methods that are in fairly common use.

2. BASIC SETUP

We adopt a now-standard pseudo-likelihood approach. Our development follows that given in Lumley and Scott (2014), where a more extensive rationale for the approach is given. 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 an arbitrary probability sampling design. Let w_i be the weight associated with the ith unit with this design. (In many cases, the weights will be the inverse selection probabilities, perhaps adjusted to compensate for non-response and frame errors by calibration to known population totals; other choices are possible, as in Pfefferman and Sverchkov [1999] and Hernán, Brumback, and Robins [2000], for example). We shall 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 measuring extra variables, such as latitude and longitude, for example, and sorting on them. A more detailed discussion is given in Lumley and Scott (2013).

Suppose that, after plotting the data and carrying out other preliminary investigations, we decide that we want to fit a parametric model, $\{f_{\boldsymbol{\theta}}(y|\boldsymbol{x}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \Re^p\}$, for the conditional density of y given $\boldsymbol{x}$. We do not assume that this parametric family necessarily contains the true model g. The Kullback–Leibler divergence between an arbitrary member of the family, f_θ, and $g(y/\boldsymbol{x})$ is

$$KL(f_{\boldsymbol{\theta}}, g) = E_g\left[\log\left\{\frac{g(y|\boldsymbol{x})}{f_\theta(y|\boldsymbol{x})}\right\}\right] = E_g[\log g(y|\boldsymbol{x})] - \ell(\boldsymbol{\theta}), \qquad (1)$$

where $\ell(\boldsymbol{\theta}) = E_g[\log f_{\boldsymbol{\theta}}(y|\boldsymbol{x})]$ is the expected population log-likelihood. The first term does not involve $\boldsymbol{\theta}$, so that the best-fitting model in our class, in the sense of minimizing the Kullback–Leibler divergence between it and the superpopulation model $g(\cdot)$, is obtained by maximizing the expected log-likelihood $\ell(\boldsymbol{\theta})$. We shall assume that the maximum is attained at a unique value of $\boldsymbol{\theta}$, which we shall denote by $\boldsymbol{\theta}^*$. For standard regression models with normal errors, choosing the model with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is equivalent to choosing the model that minimizes the mean squared prediction error of a new observation drawn from the superpopulation.

Since for any fixed value of $\boldsymbol{\theta}$, $\ell(\boldsymbol{\theta})$ is just a population mean, we can estimate it from our sample, for example by the weighted estimator

$$\widehat{\ell}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i \in s} w_i \ell_i(\boldsymbol{\theta}), \qquad (2)$$

where $\ell_i(\boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$. (We shall assume that the weights are scaled so that $\sum_{i \in s} w_i = N$.) Let $\widehat{\theta}$ be the value we obtain by maximizing $\widehat{\ell}(\boldsymbol{\theta})$. Under suitable regularity conditions (see section1.3 in Fuller [2009], for example), $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of θ^* 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 fitting regression models to survey data and the one that we shall adopt here.

We shall also adopt the asymptotic setting and regularity conditions of Theorem 1.3.9 in Fuller (2009). We have a sequence of finite populations assumed to be random samples from a fixed superpopulation. 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 and associated weights (a central limit theorem for weighted estimators), and on the parametric family $\{f_\theta\}$ (continuous second derivatives). Under these conditions, $\boldsymbol{\theta}^*$ satisfies the population score equation $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ and $\widehat{\boldsymbol{\theta}}$ satisfies the pseudo-score equation $\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \mathbf{0}$, where $\boldsymbol{U}(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \partial\widehat{\ell}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

Then, it follows from the theorem above that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}^*)) \quad \text{as} \quad n \to \infty. \tag{3}$$

We can estimate $\boldsymbol{V}(\boldsymbol{\theta}^*)$, the asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\theta}}$, consistently by

$$\widehat{V} = \widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1}\,\widehat{V}_U(\widehat{\boldsymbol{\theta}})\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1},$$

where $\widehat{\mathcal{J}}(\boldsymbol{\theta})$ is the analogue of the observed information matrix defined by

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

and $\widehat{V}_U(\boldsymbol{\theta})$ is a consistent estimator of $Cov[\sqrt{n}\widehat{\boldsymbol{U}}(\boldsymbol{\theta})]$. (Since $\widehat{\boldsymbol{U}}(\boldsymbol{\theta})$ is a vector of population totals, it is reasonable to assume that such an estimator is available routinely.)

Note that $\boldsymbol{\theta}^*$ is a superpopulation parameter in our treatment and the distributions involved are those generated by the combined operation of choosing a finite population from the superpopulation and then selecting a sample using the sampling design. It would also be possible to work within a strict finite population framework with $\boldsymbol{\theta}^*$ defined as the solution of the finite population score equations. In large populations, the results are almost identical.

In the next section, we build on all this to construct useful analogues of AIC and BIC for use with survey data. More specifically, we assume that the regularity conditions underlying (3) are satisfied and that we have a program that calculates the vector of estimated regression coefficients, $\widehat{\boldsymbol{\theta}}$, along with the estimated covariance matrix, $\widehat{V}$. This will need either information on cluster and stratum membership for every sample unit, or a set of replicate weights.

3. Model-selection criteria

3.1 AIC

Our development follows that for independent sampling in Claeskens and Hjort (2008). From (1), the appropriately weighted Kullback–Leibler divergence of our fitted model, $f_{\hat{\boldsymbol{\theta}}}$, from the true model is

$$KL(f_{\hat{\boldsymbol{\theta}}}, g) = E_g[\log g(y|\boldsymbol{x})] - \ell(\hat{\boldsymbol{\theta}}),$$

where $\ell(\boldsymbol{\theta}) = E_g[\log f_\theta(y|\boldsymbol{x})]$ is the expected population log-likelihood. If we are comparing a number of candidate models, then we are interested in maximizing $\ell(\widehat{\boldsymbol{\theta}})$, since the first term is the same for all models. Now $\widehat{\boldsymbol{\theta}}$, and hence $\ell(\widehat{\boldsymbol{\theta}})$, is a random variable. The AIC strategy is to estimate $E_g[\ell(\widehat{\boldsymbol{\theta}})] = Q_n$, say, for each candidate model, and then select the model with the largest estimated value of Q_n. This is equivalent to searching for the model with the smallest estimated average Kullback–Leibler divergence from the true model.

A naive first estimate of Q_n would be $\widehat{\ell}(\widehat{\boldsymbol{\theta}})$. This turns out to be an overestimate. More precisely,

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

where $\boldsymbol{\Delta} = \mathcal{I}(\boldsymbol{\theta}^{*})\mathbf{V}(\boldsymbol{\theta}^{*})$ with $\boldsymbol{V}(\boldsymbol{\theta}^{*})$ denoting the asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and

$$\mathcal{I}(\boldsymbol{\theta}) = E\{\widehat{\mathcal{J}}(\boldsymbol{\theta})\} = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

Note that the expectation here is with respect to both the sampling design and the superpopulation distribution. A sketch of the proof is given in the appendix. This result leads to $\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - p\bar{\delta}/n$, where p is the dimension of $\boldsymbol{\theta}$ and $\bar{\delta} = tr\{\boldsymbol{\Delta}\}/p$, as a bias-corrected estimate of Q_n. We can estimate $\bar{\delta}$ by $\widehat{\bar{\delta}} = tr\{n\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1}\widehat{V}_U(\widehat{\boldsymbol{\theta}})\} = n\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})\widehat{V}(\widehat{\boldsymbol{\theta}})$. Both $\widehat{\mathcal{J}}$ and $\widehat{V}$ are computed routinely in the course of calculating $\widehat{\boldsymbol{\theta}}$. For consistency with the standard expression for AIC under random sampling, we multiply by $-2n$ to obtain

$$d\text{AIC} = -2n\widehat{\ell}(\widehat{\boldsymbol{\theta}}) + 2p\widehat{\bar{\delta}},$$

as our modified design-based version of AIC for survey data. We want to make this as small as possible.

If we had a simple random sample and if the true model $g(y|\boldsymbol{x})$ belonged to our parametric family $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, then $\boldsymbol{V}(\boldsymbol{\theta}^{*})$ would be equal to $\mathcal{I}(\boldsymbol{\theta}^{*})^{-1}$. Thus, it is natural to call $\boldsymbol{\Delta} = \mathcal{I}(\boldsymbol{\theta}^{*})\mathbf{V}(\boldsymbol{\theta}^{*})$ the "design effect matrix," as in Rao and Scott (1984). To calculate dAIC, we simply have to inflate the usual penalty term, $2p$, by the average estimated design effect, $\widehat{\bar{\delta}}$.

Under simple random sampling, where the weights are constant, dAIC reduces to TIC, the robust version of AIC developed by Takeuchi (1976). If, in addition, our parametric family contains the true model, then $\boldsymbol{\Delta}$ reduces to the $p \times p$ identity matrix, so that $\bar{\delta} = 1$ and we get the conventional expression for AIC. For overdispersed generalized linear models, dAIC is very similar to the modified version of AIC suggested by Claeskens and Hjort (2008, section 2.7), with $\bar{\delta}$ acting as an overdispersion parameter. If $\bar{\delta}$ had the same value for all models under consideration, we could divide by this common value and use $-2n\widehat{\ell}(\widehat{\boldsymbol{\theta}})/\bar{\delta} + 2p$, which is essentially the QAIC criterion suggested in Lebreton et al. (1992) for overdispersed count data. In the examples in section 4, as in most surveys, $\bar{\delta}$ depends on the particular model being fitted and the dAIC version should be used.

In standard statistical theory, there is a close relationship between AIC, leave-one-out cross-validation, and the jackknife (Stone 1977). Similar results can be demonstrated in the survey context. A natural cross-validation estimator of $\ell(\widehat{\boldsymbol{\theta}})$ in single-stage sampling would be

$$\widehat{\ell}_{cv} = \frac{1}{N}\sum\nolimits_{i\in s} w_i \ell_i(\widehat{\boldsymbol{\theta}}_{(i)}),$$

where $\widehat{\boldsymbol{\theta}}_{(i)}$ is the estimator computed from the reduced sample obtained by omitting the *i*th unit. If the replication weights for the *i*th subsample are given by $\widetilde{w}_j^{(i)} = k_i w_j$ with $k_i = N/(N - w_i)$, then we can show that

$$\begin{aligned}\widehat{\ell}_{cv} &= \frac{1}{N}\sum_{i\in s} w_i \ell_i(\widehat{\boldsymbol{\theta}}) - tr[\widehat{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n)\,\widehat{V}_J] + o_p(n^{-1}) \\ &= \widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \mathrm{tr}[\widehat{\mathcal{J}}_n(\widehat{\boldsymbol{\theta}}_n)\,\widehat{V}_J] + o_p(n^{-1}),\end{aligned}$$

where $\widehat{V}_J = \frac{n-1}{n}\sum_i(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})^2$ is the jackknife estimator of $Cov\{\widehat{\boldsymbol{\theta}}\}$. A sketch of the proof is given in the appendix. A similar result holds to $o_p(c^{-1})$, where c is the number of primary sampling units in the sample, for multistage sampling when replicates are formed by omitting one primary sampling unit at a time. Thus, minimizing the cross-validation estimate of the prediction error of a new observation drawn from the superpopulation distribution would be asymptotically equivalent to minimizing dAIC for any design where the jackknife provides a valid variance estimator. This connection is perhaps not surprising, given the well-known bias-correcting property of the jackknife (Quenouille 1949).

3.2 BIC

The Bayesian Information Criterion (BIC) (Schwarz 1978) is based on an asymptotic Bayesian argument. Suppose that there are a finite number of models under consideration. In conventional random sampling theory, a Laplace approximation to the marginal log-likelihood log λ_m of model m gives

$$\log \lambda_m = \log L_m(\widehat{\boldsymbol{\theta}}_m) - \frac{p_m}{2}\log n + O_p(1)$$

where $L_m(\boldsymbol{\theta})$ is the likelihood, $\widehat{\boldsymbol{\theta}}_m$ the maximum likelihood estimator of $\boldsymbol{\theta}$, and $p_m = \dim(\boldsymbol{\theta})$ under model m. In large samples, choosing the model with the highest marginal log-likelihood thus corresponds to minimizing

$$\mathrm{BIC} = -2\log L(\widehat{\boldsymbol{\theta}}_m) + p_m \log n.$$

If one of the models under consideration is true, BIC will select it with probability converging to one. More realistically, if none of the models are true, the closest model in the Kullback–Leibler sense will have posterior probability converging to one (see Shalizi 2009), and so will be selected by BIC.

In general, BIC does not adapt as neatly as AIC to complex sampling, since a full Bayesian analysis would require a joint probability model for the sampling process and all the measured variables, including all design variables. In his 2007 Wald lectures, Berger used such a full-model approach to develop an extension of BIC to account for the variation in effective sample size between parameters in models fitted to clustered data.

It is possible to construct a natural analogue of BIC for the more limited purpose of selecting between submodels of a given regression model. More specifically, if all the models under consideration are submodels of a maximal model M, then model selection reduces to deciding which of the constraints defining the submodels are satisfied by $\boldsymbol{\theta}_M$. Without loss of generality, we can consider just the case where submodels are defined by zeros in certain components of $\boldsymbol{\theta}_M$. In this situation, it follows from the asymptotic equivalence of the Wald and likelihood-ratio tests (Seber and Wild 2003, section 12.4) that, ignoring terms of $O_p(1)$, the usual expression for BIC is asymptotically equivalent to

$$BIC^* = W_m - (p_M - p_m)\log n + \text{const},$$

where p_m is the number of non-zero components of $\boldsymbol{\theta}_m$ and W_m is the Wald statistic for testing that model m is true (i.e., that the appropriate components of $\boldsymbol{\theta}_m$ are zero). In this form, BIC generalizes immediately to survey inference.

We consider a coarsened Bayesian approach in which $\widehat{\boldsymbol{\theta}}_M$, the pseudo-likelihood point estimate for the full model, is regarded as the data available to the analyst, and base its likelihood on the approximate Gaussian distribution $\widehat{\boldsymbol{\theta}}_M \sim N(\boldsymbol{\theta}, n^{-1}\widehat{\mathbf{V}}_M)$. The Laplace approximation underlying the usual BIC approximates the likelihood of the data by a Gaussian likelihood for the maximum likelihood estimator. The same approximation can be used for design-based inference. The approximate marginal likelihood of model m is given by

$$\lambda_m = \int \left(\frac{n}{2\pi}\right)^{p/2} |\widehat{\mathbf{V}}_M|^{-1/2} e^{-\frac{n}{2}(\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta})^T \hat{V}_M^{-1} (\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta})} \pi_m(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta},$$

where $\pi_m(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$ given model m. Let $\widehat{\boldsymbol{\theta}}_{(m)}$ denote the subvector of $\widehat{\boldsymbol{\theta}}_M$ containing those components that are set to zero under model m and $\widehat{\mathbf{V}}_{(m)}$ the corresponding submatrix of $\widehat{\mathbf{V}}_M$. Then, the same approximation used to construct BIC now gives

$$2\log\lambda_m = (p_M - p_m)\log n_m^* - W_m + O_p(1),$$

where W_m is now the design-based Wald statistic, $W_m = n\,\widehat{\boldsymbol{\theta}}_{(m)}^T\,\widehat{\mathbf{V}}_{(m)}^{-1}\,\widehat{\boldsymbol{\theta}}_{(m)}$, for testing the hypothesis that $\boldsymbol{\theta}_{(m)} = \mathbf{0}$ and n_m^* is an effective sample size. More specifically, $n_m^* = n/\overline{d}_{(m)}$, where $\overline{d}_{(m)}$ is the geometric mean of the eigenvalues of the design effect matrix, $\mathbf{D}_{(m)} = \mathcal{I}^{(m)-1}\,\widehat{\mathbf{V}}_{(m)}$, with $\mathcal{I}^{(m)}$ denoting the appropriate submatrix of $\mathcal{I}^{-1}$. (Recall that $Cov[\widehat{\boldsymbol{\theta}}_{(m)}]$ would be equal to $\mathcal{I}^{(m)}$ under simple random sampling.) The geometric means arise because of the presence of $\det[\widehat{\mathbf{V}}_{(m)}]$ in the Gaussian likelihood. A sketch of the proof is given in the appendix. This leads to

$$\mathrm{dBIC}_m = W_m - (p_M - p_m)\log n_m^* + \mathrm{const}$$

as our design-based version of BIC. We want to make dBIC as small as possible. As with the usual BIC, the first term, W_m, penalizes oversimplification, while the second, $-(p_M - p_m)\log n_m^*$, penalizes complexity. We can set the constant (which is equal to dBIC_M for the full model) to any value we like, since this does not affect the relative likelihood and hence the posterior probability of model m. For simplicity, we set it equal to zero in the examples, so the values presented really represent dBIC_m–dBIC_M. Note that dBIC is the BIC value for the reduced Gaussian likelihood and hence, provided $\bar{d}_{(m)}$ is bounded, inherits all the standard BIC properties.

Note also that $\mathbf{D}_{(m)}$, although a design effect matrix, is quite different from the design effect matrix $\boldsymbol{\Delta}$ appearing in the expression for dAIC. Basically, $\mathbf{D}_{(m)}$ measures the effect of the design on $\widehat{\boldsymbol{\theta}}_{(m)}$, corresponding to the zero components of $\boldsymbol{\theta}_m$, while $\boldsymbol{\Delta}$ measures the effect on the estimates of the complementary non-zero components. Details are given in the appendix.

For the special case of $p = 1$, a criterion equivalent to dBIC was proposed by Fabrizi and Lahiri (2007) based on slightly different reasoning. They give two versions, with and without the $\bar{d}_{(m)}$ term (i.e., with $n_{(m)}^*$ and with $n_{(m)}^*$ replaced by n). The difference will be of smaller order than the other terms in dBIC if design effects for the parameters are bounded but can still be important, even with moderate sample sizes, when design effects are large. Using n_m^* ensures that dBIC has the useful property of being invariant under artificially increasing the sample size by duplicating data. They conduct a simulation study with observations generated from a beta-binomial distribution under a number of parameter settings. A full likelihood analysis is possible in this setting, and selection based on the dBIC criterion gives almost identical model choices to those based on the true BIC in their simulations.

Another alternative version of BIC, $BIC_n = -2\,\widehat{\ell}_m(\widehat{\boldsymbol{\theta}}_m) + p_m \log n$, has been proposed recently by Xu, Chen, and Mantell (2013), who give a non-Bayesian justification. If we write this in the form $\mathrm{BIC}_n = \Lambda_m - (p_M - p_m)\log n + \mathrm{const}$, where $\Lambda_m = 2n[\widehat{\ell}_M(\widehat{\boldsymbol{\theta}}_M) - \widehat{\ell}_m(\widehat{\boldsymbol{\theta}}_m)]$ is the pseudo likelihood-ratio statistic for testing the hypothesis that $\boldsymbol{\theta}_{(m)} = \mathbf{0}$, we see immediately that BIC_n is equal to dBIC with the Wald statistic, W_m, replaced by the pseudo likelihood-ratio statistic for the same hypothesis and n_m^* by n. Note that W_m is exactly equal to the likelihood ratio statistic for the approximate Gaussian likelihood. Roughly speaking, Λ_m acts like $(n/n^*)W_m$ (see Lumley and Scott 2014), so that BIC_n behaves like dBIC with n^* replaced by n. Thus, we might expect dBIC and BIC_n to lead to broadly similar models if design effects are not too far from one. This is what happens in Example 1 in the next section, where design effects are all less than two. If design effects are large, BIC_n overestimates the amount of information in the sample and chooses a more complex model. On the other hand, if design effects are less than one, as can happen with an effective stratification, BIC_n underestimates the amount of information and prefers simpler models. Example 2 illustrates an extreme case of this.

Xu, Chen, and Mantell (2013) show that, if one or more of the models is true, then the probability that the most parsimonious true model is selected using the BIC_n criterion converges to one as $n \to \infty$. We note that the same argument can be applied immediately to dBIC: the key feature of the derivation is to show that the likelihood-ratio test statistic Λ_m is $O_p(1)$ if model m is true and of order n if some components of $\boldsymbol{\theta}_{(m)}$ are not zero. These properties follow directly for W_m.

4. EXAMPLES

We present two examples from different extremes of sampling. The first uses data from the National Health and Nutrition Examination Survey (NHANES), which is a series of stratified multistage surveys with a large sample size but very few primary sampling units. The second is a well-studied case–control sample that investigated risk factors for esophageal cancer in Brittany. The case–control sample has no clustering, but has extremely variable weights, with big differences between those for cases and controls. It has the advantage of being one of the rare survey designs where a full maximum likelihood analysis is available for comparison.

4.1 Hypertension in NHANES

We examine the association between sodium intake and hypertension using data from NHANES, which is a multistage probability sample of the civilian non-institutionalized population of the United States, with data released in two-year waves. Each two-year wave samples approximately 10,000 people, from approximately 60 clusters (cities or counties) in 30 strata. We use data from the 2003–4 and 2005–6 waves (Centers for Disease Control and Prevention [CDC] 2005, 2007).

After restricting consideration to the clinical-examination sample and removing missing data, the sample size is $n = 13{,}057$ and the estimate of the corresponding non-missing population size is $N = 2.5 \times 10^8$. The impact of the sampling design varies considerably among variables depending on their geographic clustering and their relationship to the design criteria. For example, the design effect is 2.1 for the proportion of women, 8.6 for mean age, and it varies from 7.8 for "Other" to 38 for "Non-Hispanic Black" for proportions in race/ethnicity categories. Design effects are high because the clusters are large, due to the cost of moving the big mobile examination centers needed for the detailed clinical examinations and blood samples.

We fit logistic regression models for hypertension prevalence, defined as systolic blood pressure above 140 mmHg or diastolic blood pressure above 90 mmHg, consider adjustment for age, race/ethnicity, and gender, and present model-selection criteria for five nested models: a natural cubic spline in age, then adding race/ethnicity, then gender, then a gender by age-spline interaction, and finally sodium intake estimated from a food-frequency questionnaire. A similar analysis using systolic blood pressure instead of hypertension is given by Lumley (2010, Ch. 4). A sketch of the exploratory analyses that led to our class of models is given there. A more realistic model would be more complicated with random-effects terms for the geographical clusters, for example. However, a family of marginal models might still be appropriate for a researcher developing a model that could be used to predict outcomes for new patients not in one of the sample clusters.

Table 1 shows values $p, \overline{\delta}, \mathrm{dAIC}, n^*$ and dBIC_m (centered about dBIC_M) as we add terms to the models. Just as with their random-sampling equivalents, dBIC penalizes complexity more severely than dAIC and usually leads to simpler models. In this example, however, both criteria lead to the same model, namely the one containing all the adjustment variables except sodium. Standard Wald tests suggest that the effect of sodium is not significant ($p = 0.57$) while the gender:age interaction is very significant ($p = 5 \times 10^{8}$), providing alternative support for this model choice.

4.2 Case–control study of esophageal cancer

The classical case–control design is one of the most widely used examples of unequal-probability sampling in medical research, and is one of the few important complex-sampling designs where a full probability model is easily available. When a logistic regression model is fitted to case–control data, the full maximum likelihood estimator for all parameters except the intercept is

Table 1. Models for Hypertension, Using AIC for Selection: Spline in Age, Race/Ethnicity, Gender, Gender: Age Interaction, Sodium Intake

Model	p	$\overline{\delta}$	dAIC	n*	dBIC
Age spline	4	1.83	−196.4	10454	293.2
+ Race/ethnicity	8	1.84	−239.6	9957	73.4
+ Gender	9	1.78	−238.0	8444	82.6
+ Gender:age	12	1.69	−422.0	4211	−8.0
+ Sodium	13	1.82	−416.8	–	–

obtained by unweighted logistic regression (Prentice and Pyke 1979). The AIC and BIC values produced by the program are also valid, provided the model contains an intercept term. The weighted likelihood estimator can be substantially less efficient than the maximum likelihood estimator, but need not be, and no simple rule of thumb is available to predict its relative efficiency.

The combination of a design effect very different from unity (here smaller than one because the stratified case–control design is much more efficient than simple random sampling), very large variation in sampling weights, and the availability of a full maximum likelihood estimator make logistic regression in the case–control design a valuable test case for our proposed model-selection criteria.

We will use data from a case–control study of esophageal cancer in Brittany (Tuyns, Péquignot, and Jensen 1977), which has been previously analyzed by Breslow and Day (1980), Lumley (2010), and others, and is available on Professor Norm Breslow's website at http://faculty.washington.edu/norm/datasets.html. The data consist of alcohol consumption, tobacco consumption, and age for 200 men with esophageal cancer and 975 controls. Cases were sampled with certainty, and controls had a selection probability of approximately 1/441. This example is interesting because, despite the large variability in weights and strong covariate effects, weighted and unweighted logistic regression models give very similar point estimates, so it is meaningful to compare weighted and unweighted model choice.

We fit seven models, not all nested. The null model has an intercept only, then we add a quadratic in age. The third model includes linear main effects for alcohol and tobacco consumption, and we then add a linear-by-linear interaction term. The fifth model has terms in log alcohol and tobacco consumption, and we then add an interaction between them. The final subsuming model includes all the terms in all the other models (and thus contains both logged and unlogged alcohol and tobacco consumption). Table 2 shows the

Table 2. Models for Alcohol and Tobacco Risk in Esophageal Cancer: Age Alone, Main Effects of Risk Factors, Linear Interaction Terms, or Discrete Interaction Terms

Model	p	dAIC	dBIC	AIC	BIC
Null	1	9.62	48.2	991.9	261.3
Age	3	9.19	48.7	863.2	142.3
Main effects	5	8.65	8.5	740.1	29.0
+Interaction	6	8.39	−27.8	701.1	−10.1
Log main effects	5	8.67	17.8	741.9	35.6
+ Log interaction	6	8.41	−16.2	703.0	−3.3
Full	9	8.37	–	691.3	–

corresponding values of dAIC and dBIC, along with the values of AIC and BIC (centered at the value for the full model) obtained from the full likelihood. Model selection based on dAIC or AIC turns out to rank the models in exactly the same order, with a slight preference for the full model over that with linear main effects and interaction. Ranking based on dBIC or BIC also gives exactly the same ordering of models, now showing a clear preference for the model with linear main effects and interaction.

This particular design is extremely efficient even for a case–control study, with average design effects around 0.005 and equivalent sample sizes of about 250,000. There are approximately 430,000 people in the population, so the case–control sample, with a sampling fraction of less than 0.003, recovers most of the information that we would get with the full cohort. With so much information, it is not surprising that AIC-based methods, aimed at minimizing prediction errors, prefer the most complex model available. For scientific understanding, we might well prefer the BIC-based choice.

4.3 Alternatives

As we noted in the introduction, there are a large number of published papers that quote AIC or BIC values in the course of analyzing data from complex sample surveys. It is not always clear how the quoted values were obtained, but many use one or the other of the ad hoc criteria obtained by replacing the log-likelihood in the standard expressions for AIC and BIC formulas by either the estimated census log-likelihood, $N\widehat{\ell}$, or the weighted log-likelihood scaled to the sample size, $n\widehat{\ell}$. The BIC_n method of Xu, Chen, and Mantell (2013) corresponds to this latter approach. Statistics based on $N\widehat{\ell}$ are displayed by SAS PROC SURVEYLOGISTIC (version 9.3), for example, and scaling the weights to sum to n and otherwise ignoring the design is a simple ad hoc approach to adopt when accurate design-based methods are not available and has been a common practice historically. For AIC, the two ad hoc criteria are equivalent to using our dAIC criterion with $\overline{\delta}$ replaced by the sampling fraction, $f = n/N$, in the first case, and by 1 in the second. The relationships are slightly more complicated for BIC, but the effects are similar.

In our first example, where N is very large and f is very small, the census likelihood, $N\widehat{\ell}$, completely dominates the penalty term and criteria based on it suggest unreasonably strong support for the most complex model considered. Using $n\widehat{\ell}$, which is equivalent to replacing values of $\overline{\delta}$ (mostly just under 2 here) by 1, also results in a reduction in the effect of the penalty term, but a much smaller one, and gives more reasonable results. Here, the corresponding AIC and BIC criteria would both lead to the model containing all adjustment variables except sodium, the model selected using dAIC or dBIC.

In the second example, the sampling fraction is $f = 0.0027$, while values of $\overline{\delta}$ are also very small, ranging around 0.005. Again, we might expect analyses based on $N\widehat{\ell}$ to favor more complex models, and both AIC_N and BIC_N do indeed select the most complex possible model. In this example, however, this is not unreasonable, as this is the model selected by dAIC and by the likelihood-based AIC. On the other hand, the criteria based on $n\widehat{\ell}$ are extremely conservative, underestimating the amount of information in the sample and thus overpenalizing complexity. Both prefer the null model, even though age and both exposure variables all have very significant effects. Case–control sampling for a rare disease is used precisely because the information content of a case–control sample is almost the same as if a large fraction of the source population had been sampled. Thus, it should not be surprising that the evidence for a more complex model scales approximately as N rather than n in this example.

Our examples show that both of the ad hoc criteria can break down in some circumstances. Any good rule of thumb for deciding when such breakdowns might occur must require some idea of the actual $\overline{\delta}$ values, and hence almost as much computation as our design-based criteria.

5. DISCUSSION

The existence of so much literature using invalid versions of standard model-selection criteria suggests that there is a strong desire from subject-matter researchers for versions that could be used correctly with survey data. The dAIC and dBIC criteria that we have developed here provide such versions. They give the same model choices as the standard AIC and BIC criteria in Example 2 and in the simulations in Fabrizi and Lahiri (2007), where valid likelihoods are available. More generally, their rationale is exactly the same as that underlying the standard AIC and BIC and they have exactly the same strengths and weaknesses. A good account of these is given in Burnham and Anderson (2002), for example. Values of dAIC and dBIC are both included in version 3.29-9 of the survey package (Lumley 2013) for R (R Core Team 2013), and implementation should be straightforward for other software that already provides the Rao–Scott tests for contingency tables.

Similar arguments to those used here can be used to develop design-based analogues of AIC and BIC in other situations where a parameter estimate is defined as the solution of a weighted estimating equation, even if that estimating equation does not have the simple linear form of $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. An important example is fitting Cox models to survey survival data (Binder 1992; Lin 2000). Versions of AIC and BIC based on the partial log-likelihood can be developed using similar arguments to those used in Lumley and Scott (2013) to develop analogues of partial likelihood-ratio tests. Taniguchi, Hirukawa and Tamaki (2010, Ch. 6) also use similar techniques in time-series models to develop what they call the Generalized Takeuchi Criterion. One important situation in which this approach does not work is in fitting random effects or mixed models, and more work is needed here.

Appendix A: Asymptotics

Our asymptotic results are obtained by supposing that we have a sequence of finite populations indexed by ν and a sequence of samples of size n_ν drawn from the N_ν units in the νth population using some well-defined probability-sampling scheme. We assume that $n_\nu, N_\nu \to \infty$ with $\limsup n_\nu/N_\nu < 1$ as $\nu \to \infty$. We adopt the regularity conditions of Theorem 1.3.9 in Fuller (2009) under which the following results are established as $\nu \to \infty$:

→ **R1.** $\sqrt{n_\nu}(\widehat{\boldsymbol{\theta}}_\nu - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}^*))$:

→ **R2.** If $\{\boldsymbol{\theta}_\nu\}$ is a sequence of consistent estimators of $\boldsymbol{\theta}$* so that $\boldsymbol{\theta}_\nu \xrightarrow{p} \boldsymbol{\theta}^*$, then $\widehat{\mathcal{J}}_\nu(\boldsymbol{\theta}_\nu) \xrightarrow{p} \mathcal{I}(\boldsymbol{\theta}^*)$ as $\nu \to \infty$.

It follows that $\widehat{\boldsymbol{\theta}}_\nu = \boldsymbol{\theta}^* + O_p(n_\nu^{-1/2})$ and $\widehat{\mathcal{J}}_\nu(\widehat{\boldsymbol{\theta}}_\nu) = \mathcal{I}(\boldsymbol{\theta}^*) + o_p(1)$. We shall use both these results repeatedly.

To avoid the notation getting too cumbersome, we shall omit the subscript ν in most of what follows.

A.1 AIC

Lemma 1. Let $B(\boldsymbol{\theta}) = \widehat{\ell}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})$. Then

$$B(\widehat{\boldsymbol{\theta}}) = B(\boldsymbol{\theta}^*) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(n^{-1}).$$

Proof. We have

$$\frac{\partial B(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = \widehat{\boldsymbol{U}}(\boldsymbol{\theta}^*) - \boldsymbol{U}(\boldsymbol{\theta}^*) = \widehat{\boldsymbol{U}}(\boldsymbol{\theta}^*), \text{ since } \boldsymbol{U}(\boldsymbol{\theta}^*) = \mathbf{0},$$

$$\text{and } \frac{\partial^2 B(\boldsymbol{\theta}^*)}{\partial^2 \boldsymbol{\theta}} = -\widehat{\mathcal{J}}(\boldsymbol{\theta}^*) + \mathcal{I}(\boldsymbol{\theta}^*) = o_p(1).$$

Expanding $B(\widehat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}$*, and recalling that $\boldsymbol{U}(\boldsymbol{\theta}^*) = 0$, then leads to

$$B(\widehat{\boldsymbol{\theta}}) = B(\boldsymbol{\theta}^*) + \widehat{\boldsymbol{U}}(\boldsymbol{\theta}^*)^T (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(n^{-1}).$$

In addition, expanding $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}$* gives

$$\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{U}}(\boldsymbol{\theta}^*) - \widehat{\mathcal{J}}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(n^{-1/2}).$$

Setting $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = 0$ then leads to

$$\widehat{\boldsymbol{U}}(\boldsymbol{\theta}^*) = \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(n^{-1/2}),$$

and the Lemma follows.

The theorem below follows immediately from Lemma 1 and R2 on taking expectations and noting that $E_g[B(\boldsymbol{\theta}^*)] = \mathbf{0}$.

Theorem 1.

$$E_g[\widehat{\ell}(\widehat{\boldsymbol{\theta}})] = E_g[\ell(\widehat{\boldsymbol{\theta}})] + \frac{1}{n} tr[\boldsymbol{\Delta}] + o_p(n^{-1}),$$

where $\boldsymbol{\Delta} = \mathcal{I}(\boldsymbol{\theta}^*)\mathbf{V}(\boldsymbol{\theta}^*)$.

Now consider $\widehat{\boldsymbol{\theta}}_{(i)}$, the estimate computed from the replicate sample obtained by omitting the *i*th unit. Recall that $\widetilde{w}_j^{(i)} = k_i w_j$ with $k_i = N/(N-w_i)$. Since $\widehat{\boldsymbol{\theta}}_{(i)}$ satisfies $\widetilde{\boldsymbol{U}}_{(i)}(\widehat{\boldsymbol{\theta}}_{(i)}) = \mathbf{0}$, where

$$\widetilde{\boldsymbol{U}}_{(i)}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{j\neq i} \widetilde{w}_j^{(i)}\, \boldsymbol{U}_j(\boldsymbol{\theta}) = k_i[\widehat{\boldsymbol{U}} - w_i\boldsymbol{U}_i/N]$$

with $\boldsymbol{U}_i = \partial\ell_i/\partial\boldsymbol{\theta}$, it follows that $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}_{(i)}) = w_i\boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}_{(i)})/N$.

Lemma 2.

$$\boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}) = -\tfrac{N}{w_i}\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}) + o_p(n^{-1}).$$

Proof. Expanding $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}_{(i)})$ about $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ gives

$$\begin{aligned}\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}_{(i)}) &= \widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) - \widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}) + o_p(n^{-1})\\ &= -\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}) + o_p(n^{-1}),\end{aligned}$$

since $\widetilde{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. The result follows on setting

$$\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}_{(i)}) = \frac{w_i\boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}_{(i)})}{N},$$

and noting that $\boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}) = \boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}_{(i)}) + o_p(1))$.

Theorem 2.

$$\widehat{\ell}_{cv} = \frac{1}{N}\sum_{i\in s} w_i\ell_i(\widehat{\boldsymbol{\theta}}_{(i)}) = \widehat{\ell}(\widehat{\boldsymbol{\theta}}) - tr\{\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})\,\widehat{\mathbf{V}}_J\} + o_p(n^{-1}),$$

where $\widehat{\mathbf{V}}_J = \frac{n-1}{n}\sum_{i\in s}(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})^T$.

Proof. Expanding $\ell_i(\widehat{\boldsymbol{\theta}}_{(i)})$ about $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ gives

$$\begin{aligned}\ell_i(\widehat{\boldsymbol{\theta}}_{(i)}) &= \ell_i(\widehat{\boldsymbol{\theta}}) + \boldsymbol{U}_i(\widehat{\boldsymbol{\theta}})^T(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}) + o_p(n^{-1})\\ &= \ell_i(\widehat{\boldsymbol{\theta}}) - \frac{N}{w_i}(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})^T\mathcal{J}(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}) + o_p(n^{-1}),\end{aligned}$$

using Lemma 2. Thus, discarding the term of $o_p(n^{-1})$,

$$\begin{aligned}\frac{1}{N}\sum_{i\in s} w_i\ell_i(\widehat{\boldsymbol{\theta}}_{(i)}) - \frac{1}{N}\sum_{i\in s} w_i\ell_i(\widehat{\boldsymbol{\theta}}) & \\ &\approx -\sum_{i\in s}(\widehat{\boldsymbol{\theta}}_{(i)}-\widehat{\boldsymbol{\theta}})^T\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)}-\widehat{\boldsymbol{\theta}}) \\ &= -\frac{n}{n-1}tr\{\widehat{\mathcal{J}}(\widehat{\boldsymbol{\theta}})\,\widehat{\mathbf{V}}_J\},\end{aligned}$$

where $\widehat{\mathbf{V}}_J = \frac{n}{n-1}\sum_{i\in s}(\widehat{\boldsymbol{\theta}}_{(i)}-\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{(i)}-\widehat{\boldsymbol{\theta}})^T$.

A similar result holds to $o_p(c^{-1})$ for multistage sampling with PSUs omitted one at a time.

A.2 BIC

In this section, we assume that there is a maximal model, M, say, with parameter $\boldsymbol{\theta}_M$ that all the models under consideration correspond to setting some of the components of $\boldsymbol{\theta}_M$ equal to zero. We want to evaluate the marginal likelihood, λ_m, of model m. Reorder the components of $\boldsymbol{\theta}$ so that the first p_m components are unrestricted and the last $(p-p_m)$ are equal to zero under model m, and write $\boldsymbol{\theta}$ and $\widehat{\mathbf{V}}$ in the partitioned form

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{pmatrix} \text{and}\, \widehat{\mathbf{V}} = \begin{pmatrix}\widehat{\mathbf{V}}_{11} & \widehat{\mathbf{V}}_{12}\\ \widehat{\mathbf{V}}_{21} & \widehat{\mathbf{V}}_{22}\end{pmatrix}.$$

(Thus, $\widehat{\boldsymbol{\theta}}_2$ corresponds to $\widehat{\boldsymbol{\theta}}_{(m)}$ and $\widehat{\mathbf{V}}_{22}$ to $\widehat{\mathbf{V}}_{(m)}$ in the notation of section 6.2.) Then

$$\lambda_m = c\int e^{-\frac{n}{2}Q(\boldsymbol{\theta}_1)}\pi_m(\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1,$$

where, using standard results for the multivariate normal distribution (Seber 2009), $Q(\boldsymbol{\theta}_1)$ can be written in the form

$$Q(\boldsymbol{\theta}_1) = \widehat{\boldsymbol{\theta}}_2^T\,\widehat{\mathbf{V}}_{22}^{-1}\,\widehat{\boldsymbol{\theta}}_2 + (\widehat{\boldsymbol{\theta}}_{10}-\boldsymbol{\theta}_1)^T[\widehat{\mathbf{V}}_{11}-\widehat{\mathbf{V}}_{12}\,\widehat{\mathbf{V}}_{22}^{-1}\,\widehat{\mathbf{V}}_{21}]^{-1}(\widehat{\boldsymbol{\theta}}_{10}-\boldsymbol{\theta}_1),$$

with $\widehat{\boldsymbol{\theta}}_{10} = \widehat{\boldsymbol{\theta}}_1 - \widehat{\mathbf{V}}_{12}\,\widehat{\mathbf{V}}_{22}^{-1}\,\widehat{\boldsymbol{\theta}}_2$ and

$$\begin{aligned}c &= \left(\frac{n}{2\pi}\right)^{\frac{p}{2}}|\widehat{\mathbf{V}}|^{-1/2}\\ &= \left(\frac{n}{2\pi}\right)^{(p-p_m)/2}|\widehat{\mathbf{V}}_{22}|^{-1/2}\left(\frac{n}{2\pi}\right)^{p_m/2}|\widehat{\mathbf{V}}_{11}-\widehat{\mathbf{V}}_{12}\widehat{\mathbf{V}}_{22}^{-1}\widehat{\mathbf{V}}_{21}|^{-1/2}.\end{aligned}$$

The Laplace approximation in this case is equivalent to simply taking the first two terms in the expansion of $\pi_m(\boldsymbol{\theta}_1)$ about $\boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_{10}$. Ignoring terms that do not involve m and terms of order $O_p(1)$, this leads to

$$\log \lambda_m = \frac{1}{2}[(p - p_m)\log n - n\,\widehat{\boldsymbol{\theta}}_2^T\,\widehat{\mathbf{V}}_{22}^{-1}\,\widehat{\boldsymbol{\theta}}_2 - \log|\,\widehat{\mathbf{V}}_{22}\,| \\ + p_m\log(2\pi)] + \log \pi_m(\widehat{\boldsymbol{\theta}}_{10}),$$

assuming that $\pi_m(\boldsymbol{\theta}_1)$ has a bounded second derivative in a neighborhood of $\underline{\boldsymbol{\theta}}_{10}$.

If we had taken a random sample, then $\mathbf{V}_{22}$ would be equal to $\mathbf{V}_{22}^* = (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1}$. If we add and subtract $\frac{1}{2}\log|\,\widehat{\mathbf{V}}_{22}^*\,|$ to the expression for $\log \lambda_m$ we get

$$2\log \lambda_m = (p - p_m)\log n - n\,\widehat{\boldsymbol{\theta}}_1^T\,\widehat{\mathbf{V}}_{22}^{-1}\,\widehat{\boldsymbol{\theta}} + \log|\mathbf{D}_{(m)}| + T_m,$$

with $\mathbf{D}_{(m)} = \mathbf{V}_{22}^{*-1}\mathbf{V}_{22}$ and $T_m = p_m\log(2\pi) + 2\log \pi_m(\widehat{\boldsymbol{\theta}}_{10}) - \log|\,\widehat{\mathbf{V}}_{22}^*\,|$. The final term, T_m, is $O_p(1)$ and is omitted, as it is in the conventional development of BIC. The term $\log|\mathbf{D}_{(m)}|$ is also $O_p(1)$ but can be important with small to moderate sample sizes if design effects are large. Moreover, retaining this term ensures that the expression is not affected if the sample size is artificially augmented by replicating observations.

Note that $\mathbf{D}_{(m)}$, which is the design effect matrix associated with the likelihood-ratio test of $H_0{:}\boldsymbol{\theta}_2 = 0$ (see Lumley and Scott 2014), is quite different from the matrix $\boldsymbol{\Delta}$ appearing in the expression for dAIC. In the notation of this section, $\boldsymbol{\Delta} = \mathcal{I}_{11}\mathbf{V}_{11}$ while $\mathbf{D}_{(m)} = (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})\mathbf{V}_{22}$.

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contrl*, AC-19(6), 716–723.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.

Binder, D. A. (1992), "Fitting Cox's Proportional Hazards Models from Survey Data," *Biometrika*, 79, 139–147.

Breslow, N. E., and N. E. Day (1980), *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case–Control Studies*, World Health Organization [Distribution and Sales Service].

Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New York: Springer.

Centers for Disease Control and Prevention (CDC) (2005), *National Health and Nutrition Examination Survey Data 2003–2004*, Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (CDC) (2007), *National Health and Nutrition Examination Survey Data 2005–2006*, Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention.

Claeskens, G., and N. L. Hjort (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

Fabrizi, E., and P. Lahiri (2007), A Design-Based Approximation to the BIC in Finite Population Sampling, Technical Report 4, Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Università degli Studi di Bergamo.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā C*, 37, 117–132.

Fuller, W. A. (2009), *Sampling Statistics*, Hoboken, NJ: John Wiley and Sons.

Hernán, M. A., B. Brumback, and J. M. Robins (2000), "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men," *Epidemiology*, 11, 561–570.

Lebreton, J.-D., K. P. Burnham, J. Clobert, and D. R. Anderson (1992), "Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies," *Ecological Monographs*, 62, 67–118.

Lin, D. Y. (2000), "On Fitting Cox's Proportional Hazards Models to Survey Data," *Biometrika*, 87(1), 37–47.

Lumley, T. (2010), *Complex Surveys: A Guide to Analysis Using R*, Hoboken, NJ: John Wiley and Sons.

Lumley, T. (2013), *Survey: Analysis of Complex Survey Samples*. R package version 3.29-4.

Lumley, T., and A. Scott (2014), "Tests for Regression Models Fitted to Survey Data," *Australian and New Zealand Journal of Statistics*, 56, 1–14.

Lumley, T., and A. J. Scott (2013), "Partial Likelihood Ratio Tests for the Cox Model Under Complex Sampling," *Statistics in Medicine*, 32(1), 110–123.

Pfefferman, D., and M. Sverchkov (1999), "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data," *Sankhyā, Series B*, 61, 166–186.

Prentice, R. L., and R. Pyke (1979), "Logistic Disease Incidence Models and Case–Control Studies," *Biometrika*, 66, 403–412.

Quenouille, M. H. (1949), "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society, Series B*, 11, 68–84.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. N. K., and A. J. Scott (1984), "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data," *Annals of Statistics*, 12(1), 46–60.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Seber, G. A. F. (2009), *Multivariate Observations*, Hoboken, NJ: John Wiley & Sons.

Seber, G. A. F., and C. J. Wild (2003), *Nonlinear Regression*, Hoboken, NJ: John Wiley & Sons.

Shalizi, C. R. (2009), "Dynamics of Bayesian Updating with Dependent Data and Misspecified Models," *Electronic Journal of Statistics*, 3, 1039–1074.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Crossvalidation and Akaike's Criterion," *Journal of the Royal Statistical Society, Series B*, 39, 44–47.

Takeuchi, K. (1976), "Distribution of Information Statistics and Criteria for Adequacy Of Models (in Japanese)," *Suri-Kagaku (Mathematical Sciences)*, 153, 12–18.

Taniguchi, M., J. Hirukawa, and K. Tamaki (2010), *Optimal Statistical Inference in Financial Engineering*, Boca Raton, FL: CRC Press.

Tuyns, A. J., G. Péquignot, and O. M. Jensen (1977), "Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. Des risques qui se multiplient," *Bulletin du Cancer*, 64, 45–60.

Xu, C., J. Chen, and H. Mantell (2013), "Pseudo-Likelihood-Based Bayesian Information Criterion for Variable Selection in Survey Data," *Survey Methodology*, 39, 303–322.