AIC and BIC for Survey Data

Thomas Lumley & Alastair Scott*

Department of Statistics
University of Auckland

t.lumley@auckland.ac.nz, a.scott@auckland.ac.nz

Analysing survey data

Analyzing survey data has become big business — driven in particular by public access to the results of large medical surveys such as NHANES:

- GoogleScholar lists more than 37,000 papers with both "NHANES" and "regression" in the abstract.
- Results from hundreds of similar (albeit smaller) studies are being analyzed around the world.

What do researchers analysing data sets like NHANES want?

- They usually have a clear idea of the questions they want answered and could carry out an appropriate analysis with data from a simple random sample.

- With data from a complex survey, there are technical problems. However, the population, and what researchers want to know about it, remains unchanged — nothing to do with the sampling method. So:

  - they want to fit the same models, estimate the same parameters, as with a simple random sample
  - and they want to do this with software that mimics familiar standard software as closely as possible.

What do researchers want?

After a lot of work over the last 30 years or so, most of this has become routine.

- All the major packages now have routines for analyzing survey data by fitting regression models (linear, logistic, Cox), analyzing contingency talbles, fitting Kaplan-Meier curves.

- **Stata** and Thomas's **R** package **survey** can handle arbitrary generalized linear models.

What else is needed?

- If we compare programs like **svy:stglm** with **stglm** in **Stata** or **svyglm** with **glm** in the **R** package **survey** we see that the outputs are very similar.

- The big exceptions are quantities related to the likelihood:
  - likelihood-ratio tests
  - deviances
  - **AIC**, **BIC**, etc.

One exception

Almost all programs for fitting log-linear models to categorical survey data contain a version of the pseudo likelihood-ratio test developed by Rao & Scott (1981, 1984).

It turns out to be straightforward to extend the Rao-Scott aproach to tests for regression models in general (Lumley & Scott, 2012, 2013).

Implemented for generalized linear models in **survey** (Lumley 2004, 2013).

That leaves **AIC** and **BIC**.

Here we build on the work on likelihood-ratio tests and show how it can be used to develop natural survey analogues of **AIC** and **BIC**

Basic set-up

We have observations $\{(y_i, \boldsymbol{x}_i); i \in s\}$ on a response variable, y, and a vector of possible explanatory variables, $\boldsymbol{x}$.

The observations come from a sample, s, of n units drawn from a finite population or cohort of N units using some probability sampling design.

Let w_i be the design weight associated with the ith unit.
We assume that $\sum_{i \in s} w_i = N$.

Basic set-up

We assume that the finite population values are generated independently from some distribution with joint 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).

Basic set-up

After plotting the data etc, we decide 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 this parametric family necessarily contains the true model.

It follows from standard work on **AIC** that the best fitting model in our class, in the sense of minimizing the average Kullback-Leibler distance between it and the super-population model $g(.)$, is obtained by setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ satisfies the population score equation

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

Fitting models to survey data

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

$$\widehat{\boldsymbol{U}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{sample} w_i \mathbf{U}_i(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i \in s} w_i \frac{\partial \ell_i}{\partial \boldsymbol{\theta}}, \tag{1}$$

with $\ell_i = \log f(y_i \mid \boldsymbol{x}_i : \boldsymbol{\theta})$, be the Horvitz-Thompson estimator of $\boldsymbol{U}(\boldsymbol{\theta})$ and let $\widehat{\boldsymbol{\theta}}$ be the value we obtain by setting $\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}})$ equal to $\mathbf{0}$.

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.

Fitting models to survey data

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);
- on the estimating functions $\mathbf{U}_i$ (continuous second derivatives).

Fitting models to survey data

It follows from the theorem above that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \stackrel{\mathcal{L}}{\to} N\big(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}^*)\big) \text{ as } \mathbf{n}, \mathbf{N} \to \infty,$$

We can estimate $\mathbf{V}(\boldsymbol{\theta}^*)$ by $\widehat{\mathbf{V}} = n\widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})^{-1}\widehat{\mathbf{V}}_U(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})^{-1}$ where $\widehat{\mathbf{V}}_U(\boldsymbol{\theta})$ is an estimator of $Cov\left\{\widehat{\boldsymbol{U}}(\boldsymbol{\theta})\right\}$ (we assume that this is available routinely) and $\widehat{\boldsymbol{\mathcal{J}}}$ is the observed information matrix:

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

Pseudo likelihood-ratio tests

We can use this set-up to construct an analogue of the likelihood ratio test based on

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

with many of the properties of an ordinary likelihood ratio test.

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 $\widehat{\boldsymbol{U}}_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i\in s} w_i \partial \ell_i / \partial \boldsymbol{\theta}_1$.

Pseudo likelihood-ratio tests

Then our pseudo likelihood-ratio test statistic is given by

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

with $\widetilde{\ell}(\boldsymbol{\theta}) = n\widehat{\ell}(\boldsymbol{\theta})$.

We multiply by n to get the same value as we would with a standard regression program when we have a simple random sample with weights $w_i = N/n$

– or we could just scale the weights so that they sum to the sample size n rather than the population size N.

Pseudo likelihood-ratio tests

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

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

where $Z_1, \ldots, Z_q$ are independent $N(0,1)$ random variables and $\delta_1, \ldots, \delta_q$ are the eigenvalues of $\boldsymbol{\Delta} = \left(\boldsymbol{\mathcal{I}}_{22} - \boldsymbol{\mathcal{I}}_{21}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\right)\mathbf{V}_{22}$ where

$$\boldsymbol{\mathcal{I}} = E\left\{\widehat{\boldsymbol{\mathcal{J}}}\right\} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \boldsymbol{\mathcal{I}}_{21} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}.$$

Pseudo likelihood-ratio tests

With a random sample, $\mathbf{V}$ would be equal to $\boldsymbol{\mathcal{I}}^{-1}$.

Using the standard form for the inverse of a partitioned matrix, $\mathbf{V}_{22}$ would then be equal to $\left(\boldsymbol{\mathcal{I}}_{22} - \boldsymbol{\mathcal{I}}_{21}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{12}\right)^{-1} = \mathbf{V}_{22}^{*}$, say.

Thus we can write the matrix $\boldsymbol{\Delta}$ in the form $\boldsymbol{\Delta} = \mathbf{V}_{22}^{*-1}\mathbf{V}_{\mathbf{22}}$.

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).

Pseudo likelihood-ratio tests

To make the likelihood-ratio test statistic comparable to the Wald statistic for the same hypothesis, we suggest displaying the first-order correction,

$$\widetilde{\Lambda} = \Lambda/\widehat{\bar{\delta}},$$

in the output from a computer program.

Calculating the null distribution of $\widetilde{\Lambda}$ is discussed in Lumley & Scott (2013). The F-approximation of Thomas & Rao (1987) works well for most purposes.

$\boldsymbol{\Delta}$ and $\bar{\delta}$ play an important role in developing analogues of **AIC** and **BIC**.

AIC

The average Kullback–Leibler distance of $f(\widehat{\boldsymbol{\theta}})$ from the true model is

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

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 $\ell(\widehat{\boldsymbol{\theta}})$, which is a random variable through its dependence on $\widehat{\boldsymbol{\theta}}$.

The **AIC** strategy is to estimate $Q_n = E_g\{\ell(\widehat{\boldsymbol{\theta}})\}$ for each candidate model and then select the model with the largest value of Q_n.

AIC

A naive first estimator of Q_n would be $\widehat{\ell}(\widehat{\boldsymbol{\theta}})$ where, as before, $\widehat{\ell}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i\in s} w_i \ell_i(\boldsymbol{\theta})$. This turns out to be an overestimate.

More precisely,

$$E_g\{\widehat{\ell}(\widehat{\boldsymbol{\theta}})\} = Q_n + \frac{1}{n} tr\{\boldsymbol{\Delta}\} + o_p(n^{-1}), \tag{3}$$

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

We can estimate $\boldsymbol{\Delta}$ by $\widehat{\boldsymbol{\Delta}} = n\widehat{\boldsymbol{\mathcal{J}}}(\widehat{\boldsymbol{\theta}})^{-1}\widehat{\mathbf{V}}_U(\widehat{\boldsymbol{\theta}})$.

AIC

This leads to $\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - tr\{\boldsymbol{\Delta}\}/n$ as a bias-corrected estimate of Q_n. For consistency with the standard result for random sampling, we multiply by $-2n$ to obtain

$$dAIC = -2n\widehat{\ell}(\widehat{\boldsymbol{\theta}}_n) + 2tr\{\boldsymbol{\Delta}\} = -2\widetilde{\ell}(\widehat{\boldsymbol{\theta}}) + 2p\bar{\delta}, \quad \text{say},$$

as our modified design-based version of **AIC** for survey data.

We simply inflate the usual penalty term by the average design effect.

(Note that $\bar{\delta}$ depends on the particular model being fitted.)

AIC

Under simple random sampling, **dAIC** reduces to **TIC**, the robust version of **AIC** developed by Takeuchi (1976).

If, in addition, our parametric family $f(y \mid \boldsymbol{x} : \boldsymbol{\theta})$ contains the true model $g(y \mid \boldsymbol{x})$, then $\boldsymbol{\Delta}$ is the $p \times p$ identity matrix, so that $\bar{\delta} = 1$ and we get the conventional expression for AIC.

dAIC is very similar to the modification suggested in Claeskens & Hjort (2008) for handling overdispersion in GLMs. Here the penalty p is replaced by $p(1 + d)$, where d is the overdispersion parameter.

AIC

There is a close relationship between the usual **AIC**, the jackknife, and cross-validation.

Similar results hold for **dAIC**.

The cross-validation estimate of $\ell_i(\widehat{\boldsymbol{\theta}})$ would be $\ell_i(\widehat{\boldsymbol{\theta}}_{(i)})$, where $\widehat{\boldsymbol{\theta}}_{(i)}$ is the estimator computed from the reduced sample obtained by omitting the ith unit. Combining these gives

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

as the natural cross validation estimate of $\ell(\widehat{\boldsymbol{\theta}})$.

AIC

It turns out that

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

where $\widehat{\mathbf{V}}_J = \frac{n-1}{n}\sum_i (\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})^2$ is the jacknife estimator of $Cov\{\widehat{\boldsymbol{\theta}}\}$.

We get same result to $o_p(c^{-1})$, where c is the number of primary sampling units in the sample, with multi-stage sampling when replicates are formed by omitting PSUs one at a time.

AIC

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 our bias-corrected **dAIC** for any design where the jackknife provides a valid variance estimator.

This connection is not surprising when we recall that the jackkife was originally developed by Quenouille as a tool for bias reduction.

Example

We illustrate using data from the 2003–4 and 2005–6 waves of NHANES, examining the association between hypertension and dietary sodium and potassium intake, as well as the associations with age and race/ethnicity.

A simple logistic regression on sodium intake shows a negative correlation, possibly due to confounding by age: younger people, with less hypertension, tend to consume more sodium.

Univariate exploratory analyses show that gender and age, with nonlinear effects, are important factors. To represent the nonlinear effects of age, we used an interaction between gender and a cubic natural regression spline with 3 degrees of freedom. Our full model had $p = 14$ parameters.

Output from the SAS SURVEYLOGISTIC Procedure

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	201153424	159489290
SC	201153431	159489396
-2 Log L	201153422	159489262

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41664159.2	13	<.0001
Score	38579687.9	13	<.0001
Wald	1344.6	13	<.0001

Example

Notice that the output contains values for quantities labelled **AIC**, **SC** (aka **BIC**) and **Likelihood Ratio**.

These mean very little as they stand. However, we can adapt them to produce something useful.

Part of the problem is that we have used the published weights, summing to the population size $N = 246.75 \times 10^6$.

We get more reasonable values if we rescale to the sample size $n = 13,957$:

Output from PROC SURVEYLOGISTIC

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12800.7	10173.8
SC	12807.7	10281.8
-2 Log L	12798.7	10147.8

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2356.7	13	<.0001
Score	2182.2	13	<.0001
Wald	1344.6	13	<.0001

Example

We still need to correct for the design effects.

- We have talked about the corrections needed for the Likelihood ratio statistic and for **AIC**;
- Our modification for **dBIC** is given in Fabrizi & Lahiri (2007);
- The Score Test has the same asymptotic distribution as the LR test and a similar design effect correction applies (see Rao, Scott & Skinner, 1998).

An appropriately modified output might look something like this:

Modified output

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
dAIC	12807.0	10209.9
dBIC	1356.8	128.8
-2 Log L	12798.7	10147.8
DEFF	4.16	2.22

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1294.2	13	<.0001
Score	1198.6	13	<.0001
Wald	1344.6	13	<.0001

Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Sq	Pr > ChiSq
Intercept	1	-20.76	4.98	17.37	<.0001
sodium	1	0.05	0.04	1.50	0.2209
potassium	1	-0.07	0.05	2.46	0.1171
age1	1	10.89	2.91	14.00	0.0002
age2	1	38.64	9.56	16.32	<.0001
age3	1	8.66	2.33	13.81	0.0002
GENDR	1	7.60	2.97	6.57	0.0104
RETH1	1	-0.16	0.14	1.42	0.2343
RETH2	1	-0.09	0.24	0.14	0.7047
RETH3	1	-0.27	0.08	11.45	0.0007
RETH4	1	0.29	0.09	9.29	0.0023
age1*GENDR	1	-2.69	1.80	2.22	0.1360
age2*GENDR	1	-17.24	5.69	9.19	0.0024
age3*GENDR	1	-0.97	1.40	0.48	0.4872

BIC

BIC comes from a Laplace approximation to the log posterior probability of a model, so it needs a model.

- The Laplace approximation replaces the likelihood of the data by the (asymptotic) Gaussian likelihood of the parameter estimates
- Under complex sampling, for nested models, the parameter estimates are still asymptotically Gaussian
- Build a 'coarsened' Bayesian model using the likelihood of $\widehat{\boldsymbol{\theta}}$ rather than of the full data
- Standard **BIC** in this Bayesian model is our **dBIC** proposal

dBIC

Because the asymptotic likelihood of $\widehat{\boldsymbol{\theta}}$ is Gaussian, the end result looks just like penalized Wald statistic:

$$dBIC = (\widehat{\boldsymbol{\theta}}^{(M)} - \widehat{\boldsymbol{\theta}}^{(m)})^T \widehat{V}^{-1} (\widehat{\boldsymbol{\theta}}^{(M)} - \widehat{\boldsymbol{\theta}}^{(m)}) + p \log n + \log |\widehat{\Delta}|.$$

- Last term is of same order as some terms already neglected, but can be large for inefficient surveys
- Essentially the same criterion as proposed by Fabrizi & Lahiri (2007) on completely different grounds.