Empirical likelihood

Regression, 2016

Niels Richard Hansen
March 15, 2016

Estimating equations, the plug-in principle and NPMLE

A univariate parameter of interest is generally a function of a probability measure. That is, with $\mathcal{M}_0$ denoting a set of probability measures, a functional

$$\gamma : \mathcal{M}_0 \to \mathbb{R}$$

defines a parameter of interest as $\gamma(\rho)$ for $\rho \in \mathcal{M}_0$.

If $\mathcal{M}_0 = \{\rho_\beta \mid \beta \in \mathbb{R}^p\}$ denotes a parametrized family of probability measures we can define parameters of interest directly in terms of the parameter vector β. We can regard such a parameter function, $\beta \mapsto \gamma(\beta)$, as a function of ρ_β *as long as* $\rho_\beta = \rho_{\beta'}$ *implies that* $\gamma(\beta) = \gamma(\beta')$. This is trivially the case if β is identifiable, that is, if the map $\beta \mapsto \rho_\beta$ is one-to-one. If the property is not fulfilled, the parameter function is actually ill-defined as it depends on aspects of the parameter that we cannot decode from the distribution of the observations. Hence it is always a good idea to ensure that the parameter of interest really is a function of the probability distribution.

In a regression context ρ will have to be the joint distribution of (Y, X), from which we can compute the conditional distribution, $\rho(X)$, of $Y \mid X$ for any value of X.

The plug-in principle gives us an estimate of γ by plugging in an estimate of ρ, that is,

$$\hat{\gamma} = \gamma(\hat{\rho}).$$

We use this principle a lot for parametrized models. For a generalized linear model with $\hat{\beta}$ the MLE of the regression coefficients we use the plug-in principle to compute

an estimate of the linear predictor $\hat{\eta} = X^T\hat{\beta}$ (for a given predictor vector X) as well as of the mean $\hat{\mu} = \mu(\hat{\eta})$. According to the discussion above, this is a well defined parameter function of the probability distribution if β is identifiable from the joint distribution of (Y, X).

More generally, one can define β via an equation

$$Em(Y, X, \beta) = \int m(y, x, \beta)\rho(dy, dx) = 0$$

for a map m taking values in $\mathbb{R}^p$, and for (Y, X) having any distribution ρ as long as the expectation is well defined. If m denotes the score function for a generalized linear model and if GA1 holds, then β is the regression parameter vector. In general, with m the score function you can think of the β that solves the equation as giving the in mean best approximation to $\rho(X)$ within the class of generalized linear models considered. It is nontrivial in general to determine if there is a β solving the equation and whether it is unique. We will not pursue this discussion, but pretend that existence and uniqueness is not a problem. Thereby effectively restricting our attention to those distributions, where there is a well defined unique β solving the equation.

By the plug-in principle we estimate β by solving the estimating equation

$$\sum_{i=1}^{n} m(Y_i, X_i, \beta) = 0,$$

where the NPMLE of ρ is plugged in for the distribution of (Y, X) in the equation that defines β. The solution, $\hat{\beta}(\hat{\rho})$, is thus the plug-in NPMLE of β as a function of the NPMLE $\hat{\rho}$ of ρ, and any parameter function $\gamma(\hat{\beta}(\hat{\rho}))$ is likewise a plug-in NPMLE of γ.

Profiling

As it stands in the notes we have to carry out a non-trivial profiling computation of the empirical likelihood to compute likelihood based interval estimates (pp. 213-214).

In the case where γ is given in terms of a parameter $\beta \in \mathbb{R}^p$, which is defined as a solution of $Em(Y, X, \beta) = 0$, the set that we have to take the supremum over is

$$\Big\{\rho \in \mathcal{S} \,\Big|\, \sum_i \rho_i m(Y_i, X_i, \beta) = 0,\ \gamma(\beta) = \gamma\Big\}.$$

In this case

$$\ell(\gamma) = \sup_{\beta:\gamma(\beta)=\gamma} \underbrace{\sup_{\rho\in\mathcal{S}:\sum_i \rho_i m(Y_i,X_i,\beta)=0} \ell(\rho)}_{\ell(\beta)}.$$

It can be shown – provided $\hat{\beta}$ exists and is in the interior of the parameter space – that $\ell(\beta)$ has a quadratic expansion around $\hat{\beta}$ such that

$$2(\ell_{\max} - \ell(\beta)) \simeq (\hat{\beta} - \beta)^T \hat{H} (\hat{\beta} - \beta).$$

Here $\hat{H} = \hat{J}^T \hat{C}^{-1} \hat{J}$ with

$$\hat{J} = \sum_{i=1}^{n} D_\beta m(Y_i, X_i, \hat{\beta})$$

and

$$\hat{C} = \sum_{i=1}^{n} m(Y_i, X_i, \hat{\beta}) m(Y_i, X_i, \hat{\beta})^T.$$

The details are not entirely trivial, and they are covered in Art Owen's book *Empirical likelihood*.

In the case where the estimating equation is a score equation, $\hat{J}$ is the observed Fisher information (and it is symmetric as a second derivative), and thus an estimate of the Fisher information $\mathcal{J}$. The matrix $\hat{C}/n$ is an estimate of the variance matrix of $m(Y, X, \beta)$. When using the score equation with a correct model this variance matrix equals $\mathcal{J}/n$, in which case $\hat{C}$ and $\hat{J}$ attempt to estimate the same quantity. So for a score equation and using a correct model we will expect that $\hat{C}^{-1}\hat{J} \simeq I$, and that $\hat{H} \simeq \hat{J}$. Using $\hat{J}$ in place of $\hat{H}$ the quadratic approximation above coincides with the usual parametric quadratic approximation of $2(\ell_{\max} - \ell(\beta))$, cf. p. 198. However, the above quadratic approximation works for general estimating equations and without reference to whether the model is correctly specified or not. The difference is that it is a quadratic approximation of the combinant $2(\ell_{\max} - \ell(\beta))$ based on the *empirical likelihood* instead of a parametric likelihood. It does, however, have a likelihood interpretation, and it can be used for the construction of (empirical) likelihood intervals. Lemma 9.6 can be used for profiling the quadratic approximation.

The estimating equation for a generalized linear model with dispersion parameter $\psi = 1$ is given by

$$m(Y, X, \beta) = \theta'(X^T\beta)(Y - \mu(X^T\beta))X.$$

The observed Fisher information is

$$\hat{J} = \mathcal{J}^{\text{obs}} = \mathbf{X}^T \mathbf{W}^{\text{obs}} \mathbf{X},$$

and we also find that

$$\hat{C} = \sum_{i=1}^{n} \theta'(X_i^T\hat{\beta})^2 (Y_i - \mu(X_i^T\hat{\beta})^2 X_i X_i^T = \mathbf{X}^T \widetilde{\mathbf{W}} \mathbf{X},$$

where $\widetilde{\mathbf{W}}$ is diagonal with entries

$$\tilde{w}_{ii} = (\theta'(X_i^T\hat{\beta})(Y_i - \mu(X_i^T\hat{\beta}))^2.$$

From this it follows that

$$\hat{H} = \mathbf{X}^T\mathbf{W}^{\text{obs}}\mathbf{X}(\mathbf{X}^T\widetilde{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{\text{obs}}\mathbf{X}.$$

We will in practice typically replace $\mathbf{W}^{\text{obs}}$ by $\mathbf{W}$ when computing $\hat{H}$. For a general dispersion parameter $\psi \neq 1$ we have to multiply $\hat{H}$ by $\hat{\psi}^{-1}$.

Note how $\tilde{w}_{ii}$ is related to the variance of Y_i. If GA1 and GA2 hold the expectation of the weight $\tilde{w}_{ii}$ (conditionally on $\mathbf{X}$) is approximately equal to w_{ii}. However, if either or both of GA1 or GA2 are wrong, the weights $\tilde{w}_{ii}$ can be interpreted as adjustments for the misspecification of the model.

Combinants and sampling distributions

When we want to compute an interval estimate of a parameter of interest we have a number of options. If we focus on likelihood and approximate likelihood intervals we can choose from the following list.

- Use the profile combinant $2(\ell_{\max} - \ell(\gamma))$ based on the parametric likelihood.
- If $\gamma = a^T\beta$ we can use the combinant $(\hat{\gamma} - \gamma)^2/a^T\mathcal{J}^{-1}a$ as a quadratic approximation of the profile combinant.
- Use the profile combinant $2(\ell_{\max} - \ell(\gamma))$ based on the empirical likelihood.
- If $\gamma = a^T\beta$ we can use the combinant $(\hat{\gamma} - \gamma)^2/a^T\hat{H}^{-1}a$ as a quadratic approximation of the profile combinant.

Parametric profiling is implemented in R when using `glm` if the parameter of interest is a coordinate of β. The profile combinant is computationally more difficult to work with, but will give exact likelihood intervals. The quadratic approximations are often used in practice for convenience, but they require that the MLE exists.

For any of the four choices above we can choose different methods for calibrating the likelihood interval to achieve a certain coverage, and thus to become a confidence interval. The two possibilities covered in the course are: either we use the theoretical χ^2_1-approximation; or we use bootstrapping for approximating the distribution of the combinant. It should be noted that if we estimate a dispersion parameter by the moment estimator, deviances from the glm should be divided by $\hat{\psi}$ to obtain $2(\ell_{\max} - \ell(\gamma))$.

If the parametric model is (approximately) correct, all four combinants in combination with any of the two calibration methods should give approximately the same results. The profile combinants may give more appropriate intervals in cases where the quadractic approximation is poor, and the bootstrap calibration may be more

appropriate when the dispersion parameter is estimated – in particular for moderately small samples. For very small sample sizes nonparametric bootstrapping may not give an appropriate approximation of the distribution of the combinant.

If the parametric model is wrong, the combinants based on the parametric model will not follow a χ^2_1-distribution. Nonparametric bootstrap calibration can achieve approximately correct coverage, but the combinants based on the empirical likelihood may be more appropriate. They will, in particular, be more pivotal.