Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Output of the estimation

We explain here the various outputs from the maximum likelihood estimation procedure.

Solution of the maximum likelihood estimation

The main outputs of the maximum likelihood estimation procedure are

- the parameter estimates $\widehat{\beta}$,
- the value of the log likelihood function at the parameter estimates $\mathcal{L}(\widehat{\beta})$.

Most estimation software packages provide additional information after the estimation, in order to help appreciating the quality of the results. We summarize the most common ones here.

Variance-covariance matrix of the estimates

In addition to play a role in the optimization algorithm, the second derivatives matrix of the log likelihood function $\nabla^2\mathcal{L}(\beta)$ is also used to compute an estimate of the variance-covariance matrix of the parameter estimates, from which standard errors, t statistics and p values are generated.

Under relatively general conditions, the asymptotic variance-covariance matrix of the maximum likelihood estimates is given by the Cramer-Rao bound

$$-\operatorname{E}\left[\nabla^2\mathcal{L}(\beta)\right]^{-1} = \left\{-\operatorname{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta\partial\beta^T}\right]\right\}^{-1}. \qquad (1)$$

From the second order optimality conditions, this matrix is negative definite if the local maximum is unique, which is the algebraic equivalent of the local strict concavity of the log likelihood function.

Since we do not know the actual values of the parameters at which to evaluate the second derivatives, or the distribution of x_{in} and x_{jn} over which to take their expected value, we estimate the variance-covariance matrix by evaluating the second derivatives at the estimated parameters $\hat{\beta}$ and the sample distribution of x_{in} and x_{jn} instead of their true distribution. Thus we use

$$\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta_k \partial\beta_m}\right] \approx \sum_{n=1}^{N} \left[\frac{\partial^2 \left(y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)\right)}{\partial\beta_k \partial\beta_m}\right]_{\beta=\hat{\beta}}, \tag{2}$$

as a consistent estimator of the matrix of second derivatives. Denote this matrix as $\hat{A}$. Therefore, an estimate of the Cramer-Rao bound (1) is given by

$$\widehat{\Sigma}_{\beta}^{\mathrm{CR}} = -\hat{A}^{-1}. \tag{3}$$

If the matrix $\hat{A}$ is negative definite then $-\hat{A}$ is invertible and the Cramer-Rao bound is positive definite. Note that this may not always be the case, as it depends on the model and the sample.

Another consistent estimator of the (negative of the) second derivatives matrix can be obtained by the matrix of the cross-products of first derivatives as follows:

$$-E\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta \partial\beta^T}\right] \approx \sum_{n=1}^{n} \nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^T = \hat{B}, \tag{4}$$

where

$$\nabla L_n(\hat{\beta}) = \nabla(y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)) \tag{5}$$

is the gradient vector of the log likelihood of observation n. As the gradient $\nabla L_n(\hat{\beta})$ is a column vector of dimension $K \times 1$, and its transpose $\nabla L_n(\hat{\beta})^T$ is a row vector of size $1 \times K$, the product $\nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^T$ appearing for each observation n in (4) is a rank one matrix of size $K \times K$. The approximation $\hat{B}$ is employed by the BHHH algorithm (Berndt et al., 1974). It can also provide an estimate of the variance-covariance matrix:

$$\widehat{\Sigma}_{\beta}^{\mathrm{BHHH}} = \hat{B}^{-1}, \tag{6}$$

although this estimate is rarely used. Instead, $\hat{B}$ is used to derive a third consistent estimator of the variance-covariance matrix of the parameters, defined as

$$\widehat{\Sigma}^{\mathrm{R}}_{\beta} = (-\hat{A})^{-1} \; \widehat{B} \; (-\hat{A})^{-1} = \widehat{\Sigma}^{\mathrm{CR}}_{\beta} \; (\widehat{\Sigma}^{\mathrm{BHHH}}_{\beta})^{-1} \; \widehat{\Sigma}^{\mathrm{CR}}_{\beta}. \tag{7}$$

It is called the *robust* estimator, or sometimes the *sandwich* estimator, due to the form of equation (7).

When the true likelihood function is maximized, these estimators are asymptotically equivalent, and the Cramer-Rao bound (1) should be preferred (Kauermann and Carroll, 2001). When other consistent estimators are used, different from the maximum likelihood, the robust estimator (7) must be used (White, 1982).

Standard errors

Consider an estimate $\widehat{\beta}_k$ of the parameter β_k, and consider $\widehat{\Sigma}_\beta$ an estimate of the variance-covariance matrix of the estimates (typically, the Rao-Cramer bound or the robust estimator, as described above). The standard error of the parameter is defined as

$$\sigma_k = \sqrt{\widehat{\Sigma}_\beta(k,k)}, \tag{8}$$

where $\widehat{\Sigma}_\beta(k,k)$ is the kth entry of the diagonal of the matrix $\widehat{\Sigma}_\beta$.

t statistics

Consider an estimate $\widehat{\beta}_k$ of the parameter β_k, and σ_k its standard error. Its t statistic is defined as

$$t_k = \frac{\widehat{\beta}_k}{\sigma_k}. \tag{9}$$

It is typically used to test the null hypothesis that the true value of the parameter is zero. This hypothesis can be rejected with 95% of confidence if

$$|t_k| \geq 1.96. \tag{10}$$

p value

Consider an estimate $\widehat{\beta}_k$ of the parameter β_k, and t_k its t statistic. The p value is calculated as

$$p_k = 2(1 - \Phi(t_k)), \tag{11}$$

where $\Phi(\cdot)$ is the cumulative density function of the univariate standard normal distribution.

It conveys the exact same information as the t statistic, presented in a different way. It is the probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis that $\beta_k = 0$. The null hypothesis can be rejected with level of confidence $1 - p_k$.

Goodness of fit

Unlike linear regression, there are several measures of goodness of fit. None of them can be used in an absolute way. They can only be used to compare two models.

Clearly, an obvious measure is the log likelihood itself. It is common to compare it with a benchmark model. For instance, consider a trivial model with no parameter, associating a probability of 50% with each of the two alternatives:

$$P_n(i) = P_n(j) = \frac{1}{2}.$$

The log likelihood of the sample is therefore

$$\mathcal{L}(0) = \log(\frac{1}{2^N}) = -N \log(2),$$

where N is the number of observations. It can be used to calculate the likelihood ratio statistic:

$$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})).$$

It is called as such because it is the logarithm of the ratio of the respective likelihood values.

The statistic is used to test the null hypothesis H_0 that the estimated model is equivalent to the equal probability model. Under H_0, $-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta}))$ is asymptotically distributed as χ^2 with K degrees of freedom.

It can also be used to compute a normalized measure of goodness of fit:

$$\rho^2 = 1 - \frac{\mathcal{L}(\widehat{\beta})}{\mathcal{L}(0)}. \tag{12}$$

Such a measure has been derived to somehow mimic the R^2 in linear regression. However, in this case, it is not the square of anything. If the estimated model has the same log likelihood as the equal probability model, $\rho^2 = 0$. If the estimated model perfectly fits the data, that is if $\mathcal{L}(\widehat{\beta}) = 0$, then $\rho^2 = 1$. As mentioned above, the value itself cannot be interpreted, and it must be used only to compare two models. In particular, unlike linear regression, it is possible to have a good model with a low value of ρ^2, and a bad model with a high value.

An important limitation of this goodness of fit measure is that it is monotonic in the number of parameters of the model. It means that ρ^2 mechanically increases each time an additional variable is added to the model, even if this variable does not explain anything. Therefore, the following corrected measure is often preferred:

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\widehat{\beta}) - K}{\mathcal{L}(0)}.$$

References

Berndt, E. K., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**: 653–665.

Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation, *Journal of the American Statistical Association* **96**(456).

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.