Methods of Estimation; Ancillarity

BS2 Statistical Inference, Lecture 4
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; October 22, 2004

Method of Least Squares

Oldest *general* method of estimation, due to Legendre and Gauss, about 1810. Model determined as

$$Y = X\beta + \epsilon,$$

where X is an $n \times p$ matrix (of rank p), β is an unknown $p \times 1$ vector and ϵ an $n \times 1$ random vector with covariance matrix $\sigma^2 W^{-1}$, where W is an $n \times n$ *known* positive definite *weight matrix.* σ^2 might be known or unknown.

Simple special case occurs when W is the identity matrix, sometimes referred to as *unweighted least squares* or just least squares.

Variants relax assumption of regularity and rank condition.

Normal equations

The (weighted) least squares estimator $\hat{\beta}$ is

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||_W^2 = \arg\min_{\beta} (Y - X\beta)^\top W (Y - X\beta).$$

The estimate $\hat{\beta}$ is unique solution to the *normal equations*

$$X^\top W X \beta = X^\top W Y$$

so that

$$\hat{\beta} = (X^\top W X)^{-1} X^\top W Y.$$

If σ^2 is unknown, this is traditionally estimated as

$$\hat{\sigma}^2 = ||Y - X\hat{\beta}||_W^2 / (n - p) = SSD/(n - p),$$

where $d = n - p$ are the *degrees of freedom* of the equations.

Properties of least squares

$\hat{\beta}$ is *unbiased.*

In fact, $\hat{\beta}$ is *BLUE*, the *B*est *L*inear *U*nbiased *E*stimator of β in the sense that it has minimum variance among all unbiased estimators which are *linear* in Y.

Also, $\hat{\sigma}^2 = SSD/d$ is *unbiased* for σ^2.

These facts were established by Gauss about 1815.

If errors are multivariate Gaussian, $\hat{\beta}$ and $\hat{\sigma}^2$ are also MVUE, because then $(\hat{\beta}, \hat{\sigma}^2)$ are sufficient and complete.

Method of moments

Not a method with same generality as LS, but was used e.g. by Thiele (1889) to deal with skew distributions.

Consider $X = (X_1, \ldots, X_n)$ independent and identically distributed with individual densities $f(x_i; \theta)$ where $\theta \in \Theta$ is unknown.

Then estimate the first p moments of the distribution with corresponding empirical moments:

$$\mu_k(\theta) = \int x^k f(x; \theta)\, dx, \quad m_k = \frac{1}{n} \sum x_i^k, quadk = 1, \ldots, p.$$

Using central moments

Equivalently, use *central* moments for $k \geq 2$ where $\mu = \mu_1(\theta) = \bar{\mu}_1(\theta)$:

$$\bar{\mu}_k(\theta) = \int (x - \mu)^k f(x; \theta)\, dx, \quad \bar{m}_k = \frac{1}{n} \sum (x_i - \bar{x})^k$$

for $k = 1, \ldots, p$. Use as many moments as necessary to ensure the equations

$$\bar{\mu}_k(\theta) = \bar{m}_k, \quad k = 1, 2, \ldots, p$$

have a unique solution for θ. Moment estimators are *unbiased for moments* (but not for central moments).

The method of moments can sometimes give "quick and dirty" estimates but is generally not very good.

Example

Consider a sample $X = (X_1, \ldots, X_n)$ from a Gamma distribution with individual densities

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta}, \alpha > 0, \beta > 0.$$

For $\theta = (\alpha, \beta)$ we have

$$\mu_1(\theta) = \mathbf{E}(X) = \alpha\beta, \quad \bar{\mu}_2(\theta) = \mathbf{V}(X) = \alpha\beta^2$$

so the moment estimators become

$$\tilde{\alpha} = \frac{(\sum X_i)^2}{n\sum(X_i - \bar{X})^2}, \quad \tilde{\beta} = \frac{\sum(X_i - \bar{X})^2}{\sum x_i}.$$

Method of maximum likelihood

The method of maximum likelihood chooses the estimate of θ which make the observations most probable

$$\hat{\theta} = \hat{\theta}(x) = \arg\max_{\theta} L_x(\theta) = \arg\max_{\theta} f(x;\theta)$$

provided a unique maximum exists.

As a general method, this is due to R. A. Fisher about 1920. When errors are Gaussian in the general linear model, the maximum likelihood estimator (MLE) is equal to the least squares estimator.

Strictly speaking this is not true for σ^2, where the MLE is SSD/n rather than SSD/d. The latter is then an example of a *bias-corrected* MLE.

Invariance of the MLE

The method of maximum likelihood generally leads to estimators with very good properties, although there are exceptions. For example, the MLE behaves well under reparametrizations:

If g is a one-to-one function, and $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

This is trivially true as if we let $\theta = g^{-1}(\mu)$ then $f\{x; g^{-1}(\mu)\}$ is maximized in μ exactly when $\mu = g(\hat{\theta})$.

When g is not one-to-one the discussion becomes more subtle, but we simply choose to define

$$\hat{g}_{\mathrm{MLE}}(\theta) = g(\hat{\theta}.)$$

Sufficiency and MLE

If $\hat{\theta}$ is the unique MLE and $T = t(X)$ is sufficient, then $\hat{\theta}$ is a function of t.

This follows from the factorization theorem as

$$\arg\max_{\theta} h(x)k\{t(x);\theta\} = h(x)\arg\max_{\theta} k\{t(x);\theta\}$$

and the latter clearly is a function of $t(x)$.

In particular this implies

If the MLE is itself sufficient, it is minimal sufficient.

Note also that *Rao–Blackwellization is never needed for the MLE,* since it already is a function of any sufficient statistic.

Ancillarity

A statistic $A = a(X)$ is *ancillary* if the distribution of A does not depend on θ. Intuitively A is then *uninformative* about the unknown parameter.

Notion of ancillarity is also due to Fisher, but its role is less clear than that of sufficiency.

If $\hat{\theta}$ is not itself sufficient, it is often possible to find an ancillary statistic so that $(\hat{\theta}, A)$ is jointly sufficient. Then

$$f(x \mid A = a; \theta) \propto h(x)k\{\hat{\theta}(x), a; \theta\}$$

so $\hat{\theta}$ is sufficient when considering the conditional distribution given the ancillary A. Since the distribution of A carries no information about θ, it is tempting to insist on conditioning in this way.

Example

Consider an experiment with two instruments available. One produces measurements $\mathcal{N}(\theta, 1)$ whereas the other produces measurements which are $\mathcal{N}(\theta, 100)$.

Toss a fair coin and let $A = i, i = 1, 2$ denote that the instrument i is chosen. Perform then the measurement to obtain X. The joint distribution of (X, A) is determined as

$$f(x, a; \theta) = \phi(x - \theta)1_{\{1\}}(a)/2 + \phi\{(x - \theta)/10\}1_{\{2\}}(a)/2$$

so $\hat{\theta} = x$ is not sufficient. But why should we not consider $A = a$ fixed and condition on the actual instrument used?

This example in convincing, but in general there is no unique ancillary statistic to choose from, so then it is not all that clear.

Basu's Theorem

Sometimes it does not matter, whether we condition on A or not: *If $T = t(X)$ is complete and sufficient and A is ancillary, then T and A are independent.*

The proof is surprisingly simple:

Let g be an arbitrary bounded function of a and let $m = \mathbf{E}_\theta\{g(A)\}$. Note m does not depend on θ as A was ancillary. Now let

$$h(t(x)) = \mathbf{E}_\theta[\{g(A) - m\} \,|\, T = t(x)]$$

which also does not depend on θ because T was sufficient. Iterating expectations and using the definition of m yields

$$0 = \mathbf{E}_\theta\{h(T)\} = \mathbf{E}_\theta \mathbf{E}_\theta[g\{a(X)\} - m \,|\, T] \text{ for all } \theta$$

and completeness therefore yields

$$\mathbf{E}_\theta\{g(A)\,|\,T = t(x)\} = \mathbf{E}\{g(A)\},$$

thus that A and T are independent.