4

Moment Estimators

The method of moments determines estimators by comparing sample and theoretical moments. Moment estimators are useful for their simplicity, although not always optimal. Maximum likelihood estimators for full exponential families are moment estimators, and their asymptotic normality can be proved by treating them as such.

4.1 Method of Moments

Let $X_1, \dots, X_n$ be a sample from a distribution P_θ that depends on a parameter θ, ranging over some set Θ. The method of moments consists of estimating θ by the solution of a system of equations

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = E_\theta f_j(X), \qquad j = 1, \dots, k,$$

for given functions $f_1, \dots, f_k$. Thus the parameter is chosen such that the sample moments (on the left side) match the theoretical moments. If the parameter is k-dimensional one usually tries to match k moments in this manner. The choices $f_j(x) = x^j$ lead to the method of moments in its simplest form.

Moment estimators are not necessarily the best estimators, but under reasonable conditions they have convergence rate $\sqrt{n}$ and are asymptotically normal. This is a consequence of the delta method. Write the given functions in the vector notation $f = (f_1, \dots, f_k)$, and let $e : \Theta \mapsto \mathbb{R}^k$ be the vector-valued expectation $e(\theta) = P_\theta f$. Then the moment estimator $\hat{\theta}_n$ solves the system of equations

$$\mathbb{P}_n f \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i) = e(\theta) \equiv P_\theta f.$$

For existence of the moment estimator, it is necessary that the vector $\mathbb{P}_n f$ be in the range of the function e. If e is one-to-one, then the moment estimator is uniquely determined as $\hat{\theta}_n = e^{-1}(\mathbb{P}_n f)$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\big(e^{-1}(\mathbb{P}_n f) - e^{-1}(P_{\theta_0} f)\big).$$

If $\mathbb{P}_n f$ is asymptotically normal and e^{-1} is differentiable, then the right side is asymptotically normal by the delta method.

The derivative of e^{-1} at $e(\theta_0)$ is the inverse $e'^{-1}_{\theta_0}$ of the derivative of e at θ_0. Because the function e^{-1} is often not explicit, it is convenient to ascertain its differentiability from the differentiability of e. This is possible by the inverse function theorem. According to this theorem a map that is (continuously) differentiable throughout an open set with nonsingular derivatives is locally one-to-one, is of full rank, and has a differentiable inverse. Thus we obtain the following theorem.

4.1 ***Theorem.*** *Suppose that $e(\theta) = P_\theta f$ is one-to-one on an open set $\Theta \subset \mathbb{R}^k$ and continuously differentiable at θ_0 with nonsingular derivative e'_{θ_0}. Moreover, assume that $P_{\theta_0}\|f\|^2 < \infty$. Then moment estimators $\hat\theta_n$ exist with probability tending to one and satisfy*

$$\sqrt{n}(\hat\theta_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} N\Big(0, e'^{-1}_{\theta_0} P_{\theta_0} f f^T \big(e'^{-1}_{\theta_0}\big)^T\Big).$$

Proof. Continuous differentiability at θ_0 presumes differentiability in a neighborhood and the continuity of $\theta \mapsto e'_\theta$ and nonsingularity of e'_{θ_0} imply nonsingularity in a neighborhood. Therefore, by the inverse function theorem there exist open neighborhoods U of θ_0 and V of $P_{\theta_0} f$ such that $e: U \mapsto V$ is a differentiable bijection with a differentiable inverse $e^{-1}: V \mapsto U$. Moment estimators $\hat\theta_n = e^{-1}(\mathbb{P}_n f)$ exist as soon as $\mathbb{P}_n f \in V$, which happens with probability tending to 1 by the law of large numbers.

The central limit theorem guarantees asymptotic normality of the sequence $\sqrt{n}(\mathbb{P}_n f - P_{\theta_0} f)$. Next use Theorem 3.1 on the display preceding the statement of the theorem. ■

For completeness, the following two lemmas constitute, if combined, a proof of the inverse function theorem. If necessary the preceding theorem can be strengthened somewhat by applying the lemmas directly. Furthermore, the first lemma can be easily generalized to infinite-dimensional parameters, such as used in the semiparametric models discussed in Chapter 25.

4.2 ***Lemma.*** *Let $\Theta \subset \mathbb{R}^k$ be arbitrary and let $e: \Theta \mapsto \mathbb{R}^k$ be one-to-one and differentiable at a point θ_0 with a nonsingular derivative. Then the inverse e^{-1} (defined on the range of e) is differentiable at $e(\theta_0)$ provided it is continuous at $e(\theta_0)$.*

Proof. Write $\eta = e(\theta_0)$ and $\Delta h = e^{-1}(\eta + h) - e^{-1}(\eta)$. Because e^{-1} is continuous at η, we have that $\Delta h \mapsto 0$ as $h \mapsto 0$. Thus

$$\eta + h = e\,e^{-1}(\eta + h) = e(\Delta h + \theta_0) = e(\theta_0) + e'_{\theta_0}(\Delta h) + o\big(\|\Delta h\|\big),$$

as $h \mapsto 0$, where the last step follows from differentiability of e. The displayed equation can be rewritten as $e'_{\theta_0}(\Delta h) = h + o\big(\|\Delta h\|\big)$. By continuity of the inverse of e'_{θ_0}, this implies that

$$\Delta h = e'^{-1}_{\theta_0}(h) + o\big(\|\Delta h\|\big).$$

In particular, $\|\Delta h\|\big(1 + o(1)\big) \leq \big\|e'^{-1}_{\theta_0}(h)\big\| = O\big(\|h\|\big)$. Insert this in the displayed equation to obtain the desired result that $\Delta h = e'^{-1}_{\theta_0}(h) + o\big(\|h\|\big)$. ■

4.3 ***Lemma.*** *Let $e: \Theta \mapsto \mathbb{R}^k$ be defined and differentiable in a neighborhood of a point θ_0 and continuously differentiable at θ_0 with a nonsingular derivative. Then e maps every* *sufficiently small open neighborhood U of θ_0 onto an open set V and $e^{-1}: V \mapsto U$ is well defined and continuous.*

Proof. By assumption, $e'_\theta \to A^{-1} := e'_{\theta_0}$ as $\theta \mapsto \theta_0$. Thus $\|I - Ae'_\theta\| \leq \frac{1}{2}$ for every θ in a sufficiently small neighborhood U of θ_0. Fix an arbitrary point $\eta_1 = e(\theta_1)$ from $V = e(U)$ (where $\theta_1 \in U$). Next find an $\varepsilon > 0$ such that $\overline{\text{ball}}(\theta_1, \varepsilon) \subset U$, and fix an arbitrary point η with $\|\eta - \eta_1\| < \delta := \frac{1}{2}\|A\|^{-1}\varepsilon$. It will be shown that $\eta = e(\theta)$ for some point $\theta \in \overline{\text{ball}}(\theta_1, \varepsilon)$. Hence every $\eta \in \text{ball}(\eta_1, \delta)$ has an original in $\overline{\text{ball}}(\theta_1, \varepsilon)$. If e is one-to-one on U, so that the original is unique, then it follows that V is open and that e^{-1} is continuous at η_1.

Define a function $\phi(\theta) = \theta + A\big(\eta - e(\theta)\big)$. Because the norm of the derivative $\phi'_\theta = I - Ae'_\theta$ is bounded by $\frac{1}{2}$ throughout U, the map ϕ is a contraction on U. Furthermore, if $\|\theta - \theta_1\| \leq \varepsilon$,

$$\big\|\phi(\theta) - \theta_1\big\| \leq \big\|\phi(\theta) - \phi(\theta_1)\big\| + \big\|\phi(\theta_1) - \theta_1\big\| \leq \frac{1}{2}\|\theta - \theta_1\| + \|A\|\,\|\eta - \eta_1\| < \varepsilon.$$

Consequently, ϕ maps $\overline{\text{ball}}(\theta_1, \varepsilon)$ into itself. Because ϕ is a contraction, it has a fixed point $\theta \in \overline{\text{ball}}(\theta_1, \varepsilon)$: a point with $\phi(\theta) = \theta$. By definition of ϕ this satisfies $e(\theta) = \eta$.

Any other $\tilde{\theta}$ with $e(\tilde{\theta}) = \eta$ is also a fixed point of ϕ. In that case the difference $\tilde{\theta} - \theta = \phi(\tilde{\theta}) - \phi(\theta)$ has norm bounded by $\frac{1}{2}\|\tilde{\theta} - \theta\|$. This can only happen if $\tilde{\theta} = \theta$. Hence e is one-to-one throughout U. ■

4.4 ***Example.*** Let $X_1, \ldots, X_n$ be a random sample from the beta-distribution: The common density is equal to

$$x \mapsto \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} 1_{0<x<1}.$$

The moment estimator for (α, β) is the solution of the system of equations

$$\overline{X}_n = E_{\alpha,\beta} X_1 = \frac{\alpha}{\alpha + \beta},$$
$$\overline{X_n^2} = E_{\alpha,\beta} X_1^2 = \frac{(\alpha + 1)\alpha}{(\alpha + \beta + 1)(\alpha + \beta)}.$$

The righthand side is a smooth and regular function of (α, β), and the equations can be solved explicitly. Hence, the moment estimators exist and are asymptotically normal. □

*4.2 Exponential Families

Maximum likelihood estimators in full exponential families are moment estimators. This can be exploited to show their asymptotic normality. Actually, as shown in Chapter 5, maximum likelihood estimators in smoothly parametrized models are asymptotically normal in great generality. Therefore the present section is included for the benefit of the simple proof, rather than as an explanation of the limit properties.

Let $X_1, \ldots, X_n$ be a sample from the k-dimensional *exponential family* with density

$$p_\theta(x) = c(\theta)\, h(x)\, e^{\theta^T t(x)}.$$

Thus h and $t = (t_1, \dots, t_k)$ are known functions on the sample space, and the family is given in its natural parametrization. The parameter set Θ must be contained in the *natural parameter space* for the family. This is the set of θ for which p_θ can define a probability density. If μ is the dominating measure, then this is the right side in

$$\Theta \subset \left\{ \theta \in \mathbb{R}^k : c(\theta)^{-1} \equiv \int h(x)\, e^{\theta^T t(x)}\, d\mu(x) < \infty \right\}.$$

It is a standard result (and not hard to see) that the natural parameter space is convex. It is usually open, in which case the family is called "regular." In any case, we assume that the true parameter is an inner point of Θ. Another standard result concerns the smoothness of the function $\theta \mapsto c(\theta)$, or rather of its inverse. (For a proof of the following lemma, see [100, p. 59] or [17, p. 39].)

4.5 Lemma. *The function $\theta \mapsto \int h(x)\, e^{\theta^T t(x)}\, d\mu(x)$ is analytic on the set $\{\theta \in \mathbb{C}^k : \operatorname{Re} \theta \in \overset{\circ}{\Theta}\}$. Its derivatives can be found by differentiating (repeatedly) under the integral sign:*

$$\frac{\partial^p \int h(x)\, e^{\theta^T t(x)}\, d\mu(x)}{\partial\theta_1^{i_1} \cdots \partial\theta_k^{i_k}} = \int h(x)\, t_1(x)^{i_1} \cdots t_k(x)^{i_k}\, e^{\theta^T t(x)}\, d\mu(x),$$

for any natural numbers p and $i_1 + \cdots + i_k = p$.

The lemma implies that the log likelihood $\ell_\theta(x) = \log p_\theta(x)$ can be differentiated (infinitely often) with respect to θ. The vector of partial derivatives (the score function) satisfies

$$\dot{\ell}_\theta(x) = \frac{\dot{c}}{c}(\theta) + t(x) = t(x) - E_\theta t(X).$$

Here the second equality is an example of the general rule that score functions have zero means. It can formally be established by differentiating the identity $\int p_\theta\, d\mu \equiv 1$ under the integral sign: Combine the lemma and the Leibniz rule to see that

$$\frac{\partial}{\partial\theta_i} \int p_\theta\, d\mu = \int \frac{\partial c(\theta)}{\partial\theta_i} h(x)\, e^{\theta^T t(x)}\, d\mu(x) + \int c(\theta)\, h(x)\, t_i(x)\, e^{\theta^T t(x)}\, d\mu(x).$$

The left side is zero and the equation can be rewritten as $0 = \dot{c}/c(\theta) + E_\theta t(X)$.

It follows that the likelihood equations $\sum \dot{\ell}_\theta(X_i) = 0$ reduce to the system of k equations

$$\frac{1}{n} \sum_{i=1}^n t(X_i) = E_\theta t(X).$$

Thus, the maximum likelihood estimators are moment estimators. Their asymptotic properties depend on the function $e(\theta) = E_\theta t(X)$, which is very well behaved on the interior of the natural parameter set. By differentiating $E_\theta t(X)$ under the expectation sign (which is justified by the lemma), we see that its derivative matrices are given by

$$e'_\theta = \operatorname{Cov}_\theta t(X).$$

The exponential family is said to be of *full rank* if no linear combination $\sum_{j=1}^k \lambda_j t_j(X)$ is constant with probability 1; equivalently, if the covariance matrix of $t(X)$ is nonsingular. In view of the preceding display, this ensures that the derivative e'_θ is strictly positive-definite throughout the interior of the natural parameter set. Then e is one-to-one, so that there exists at most one solution to the moment equations. (Cf. Problem 4.6.) In view of the expression for $\dot\ell_\theta$, the matrix $-ne'_\theta$ is the second-derivative matrix (Hessian) of the log likelihood $\sum_{i=1}^n \ell_\theta(X_i)$. Thus, a solution to the moment equations must be a point of maximum of the log likelihood.

A solution can be shown to exist (within the natural parameter space) with probability 1 if the exponential family is "regular," or more generally "steep" (see [17]); it is then a point of absolute maximum of the likelihood. If the true parameter is in the interior of the parameter set, then a (unique) solution $\hat\theta_n$ exists with probability tending to 1 as $n \mapsto \infty$, in any case, by Theorem 4.1. Moreover, this theorem shows that the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal with covariance matrix

$$e'^{-1}_{\theta_0} \operatorname{Cov}_{\theta_0} t(X) \left(e'^{-1}_{\theta_0}\right)^T = \left(\operatorname{Cov}_{\theta_0} t(X)\right)^{-1}.$$

So far we have considered an exponential family in standard form. Many examples arise in the form

$$p_\theta(x) = d(\theta)\, h(x)\, e^{Q(\theta)^T t(x)}, \tag{4.6}$$

where $Q = (Q_1, \ldots, Q_k)$ is a vector-valued function. If Q is one-to-one and a maximum likelihood estimator $\hat\theta_n$ exists, then by the invariance of maximum likelihood estimators under transformations, $Q(\hat\theta_n)$ is the maximum likelihood estimator for the natural parameter $Q(\theta)$ as considered before. If the range of Q contains an open ball around $Q(\theta_0)$, then the preceding discussion shows that the sequence $\sqrt{n}\big(Q(\hat\theta_n) - Q(\theta_0)\big)$ is asymptotically normal. It requires another application of the delta method to obtain the limit distribution of $\sqrt{n}(\hat\theta_n - \theta_0)$. As is typical of maximum likelihood estimators, the asymptotic covariance matrix is the inverse of the *Fisher information matrix*

$$I_\theta = \mathrm{E}_\theta \dot\ell_\theta(X) \dot\ell_\theta(X)^T.$$

4.6 ***Theorem.*** *Let $\Theta \subset \mathbb{R}^k$ be open and let $Q : \Theta \mapsto \mathbb{R}^k$ be one-to-one and continuously differentiable throughout Θ with nonsingular derivatives. Let the (exponential) family of densities p_θ be given by (4.6) and be of full rank. Then the likelihood equations have a unique solution $\hat\theta_n$ with probability tending to 1 and $\sqrt{n}(\hat\theta_n - \theta) \overset{\theta}{\rightsquigarrow} N(0, I_\theta^{-1})$ for every θ.*

Proof. According to the inverse function theorem, the range of Q is open and the inverse map Q^{-1} is differentiable throughout this range. Thus, as discussed previously, the delta method ensures the asymptotic normality. It suffices to calculate the asymptotic covariance matrix. By the preceding discussion this is equal to

$$Q'^{-1}_\theta \left(\operatorname{Cov}_\theta t(X)\right)^{-1} \left(Q'^{-1}_\theta\right)^T.$$

By direct calculation, the score function for the model is equal to $\dot\ell_\theta(x) = \dot d / d(\theta) + (Q'_\theta)^T t(x)$. As before, the score function has mean zero, so that this can be rewritten as $\dot\ell_\theta(x) = (Q'_\theta)^T\big(t(x) - E_\theta t(X)\big)$. Thus, the Fisher information matrix equals $I_\theta = (Q'_\theta)^T \operatorname{Cov}_\theta t(X) Q'_\theta$. This is the inverse of the asymptotic covariance matrix given in the preceding display. ■

Not all exponential families satisfy the conditions of the theorem. For instance, the normal $N(\theta, \theta^2)$ family is an example of a "curved exponential family." The map $Q(\theta) = (\theta^{-2}, \theta^{-1})$ (with $t(x) = (-x^2/2, x)$) does not fill up the natural parameter space of the normal location-scale family but only traces out a one-dimensional curve. In such cases the result of the theorem may still hold. In fact, the result is true for most models with "smooth parametrizations," as is seen in Chapter 5. However, the "easy" proof of this section is not valid.

PROBLEMS

1. Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $[-\theta, \theta]$. Find the moment estimator of θ based on $\overline{X^2}$. Is it asymptotically normal? Can you think of an estimator for θ that converges faster to the parameter?
2. Let $X_1, \ldots, X_n$ be a sample from a density p_θ and f a function such that $e(\theta) = E_\theta f(X)$ is differentiable with $e'(\theta) = E_\theta \dot{\ell}_\theta(X) f(X)$ for $\ell_\theta = \log p_\theta$.
   (i) Show that the asymptotic variance of the moment estimator based on f equals $\mathrm{var}_\theta(f)/\mathrm{cov}_\theta(f, \dot{\ell}_\theta)^2$.
   (ii) Show that this is bigger than I_θ^{-1} with equality for all θ if and only if the moment estimator is the maximum likelihood estimator.
   (iii) Show that the latter happens only for exponential family members.
3. To what extent does the result of Theorem 4.1 require that the observations are i.i.d.?
4. Let the observations be a sample of size n from the $N(\mu, \sigma^2)$ distribution. Calculate the Fisher information matrix for the parameter $\theta = (\mu, \sigma^2)$ and its inverse. Check directly that the maximum likelihood estimator is asymptotically normal with zero mean and covariance matrix I_θ^{-1}.
5. Establish the formula $e'_\theta = \mathrm{Cov}_\theta\, t(X)$ by differentiating $e(\theta) = E_\theta t(X)$ under the integral sign. (Differentiating under the integral sign is justified by Lemma 4.5, because $E_\theta t(X)$ is the first derivative of $c(\theta)^{-1}$.)
6. Suppose a function $e : \Theta \mapsto \mathbb{R}^k$ is defined and continuously differentiable on a convex subset $\Theta \subset \mathbb{R}^k$ with strictly positive-definite derivative matrix. Then e has at most one zero in Θ. (Consider the function $g(\lambda) = (\theta_1 - \theta_2)^T e\bigl(\lambda\theta_1 + (1-\lambda)\theta_2\bigr)$ for given $\theta_1 \neq \theta_2$ and $0 \leq \lambda \leq 1$. If $g(0) = g(1) = 0$, then there exists a point λ_0 with $g'(\lambda_0) = 0$ by the mean-value theorem.)