Mathematical statistics
Week 4: General methods to prepare estimators

Péter Elek and Ádám Reiff

October 8 and 10, 2013

1. Method of Moments (MM) Estimator

Distribution moments.

- $m_k = E\left(X^k\right)$ is the k-th moment of the distribution
  - First moment is $E\left(X\right)$, the expected value.
  - Second moment is $E\left(X^2\right)$, not the variance!
- $E\left[(X - E(X))^k\right]$ is the k-th central moment of the distribution
  - Variance: $\sigma^2 = E\left[(X - E(X))^2\right]$ (second central moment)
  - Skewness = $E\left[(X - E(X))^3\right] / \sigma^3$ (standardized third central moment)
  - Kurtosis = $E\left[(X - E(X))^4\right] / \sigma^4$ (standardized fourth central moment)

Sample moments.

- The sample moment $\widehat{m}_k = \frac{\sum_{i=1}^n X_i^k}{n}$ is an unbiased and consistent estimator of m_k.

Method of Moments (MM) estimator.

- Let us denote the estimated parameters by $(\theta_1, \ldots, \theta_k)$.
- Suppose they are some function of the distribution moments: $\theta_j = g_j\left(m_1, m_2, \ldots, m_k\right)$.
- Then the method of moments estimator of θ_j is defined as the same function of the sample moments: $\widehat{\theta}_j = g_j\left(\widehat{m}_1, \widehat{m}_2, \ldots, \widehat{m}_k\right)$.

- (Can be used not only to distribution moments, but other characteristics of distributions: e.g. sample median is the MM-estimator of the population median.)

Calculation of the MM-estimator.

- In practice, the MM-estimator is calculated in two steps.
- First, a few distribution moments are obtained as functions of the parameters:

$$\begin{aligned} m_1 &= h_1(\theta_1, \ldots, \theta_k) \\ m_2 &= h_2(\theta_1, \ldots, \theta_k) \\ m_k &= h_k(\theta_1, \ldots, \theta_k). \end{aligned}$$

- Then, the sample moments are made equal to these functions and the resulting system of equations is solved:

$$\begin{aligned} \widehat{m}_1 &= h_1(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) \\ \widehat{m}_2 &= h_2(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) \\ \widehat{m}_k &= h_k(\widehat{\theta}_1, \ldots, \widehat{\theta}_k). \end{aligned}$$

MM-estimator of population mean and variance.

- $\overline{X}$ is the MM-estimator of the population mean μ.
  - It holds by definition.
- s^2 is the MM-estimator of the population variance.
  - It "almost" holds by definition.
  - We know that $m_1 = \mu$ and $m_2 = \mu^2 + \sigma^2$.
  - Hence $\widehat{\sigma^2}_{MM} = \widehat{m}_2 - \left(\overline{X}\right)^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \left(\frac{\sum_{i=1}^n X_i}{n}\right)^2 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{n}$.
- Examples:
  - If $X \sim N(\mu, \sigma^2)$, then $\widehat{\mu}_{MM} = \overline{X}$ and $\widehat{\sigma^2}_{MM} = s^2$.
  - Also, if $X \sim Bernoulli(p)$, then $\widehat{p}_{MM} = \overline{X}$.

MM-estimator of covariance.

- (Theoretical) covariance is defined as $\sigma_{XY} = E\left[(X - E(X))\,(Y - E(Y))\right]$.
- The MM-estimator is $\widehat{\sigma}_{XY} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{n}$.
- It is biased and consistent.
- (It would be unbiased if we divided by $n - 1$.)

MM-estimator of correlation.

- (Theoretical) correlation is defined as:
$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{E\left[(X-E(X))^2\right]}\sqrt{E\left[(Y-E(Y))^2\right]}}$.
- The MM-estimator is $\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$,
- which is biased and consistent.
- There is NO unbiased estimator of r_{XY} (a famous result in statistics).

Example: MM-estimator for θ in an exponential sample.

- Assume $X \sim Exp(\theta)$ and $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of size n. Find the method of moments estimator of θ.
- Solution 1: in this case we have $E(X) = \theta$ (can be proven by integration), so $\widehat{\theta}_{MM} = \overline{X}$.
- Solution 2: we also have $\sqrt{Var(X)} = \theta$, so an alternative estimator is $\widehat{\theta}_{MM} = s$. But in practice, we prefer estimators calculated from lower moments.

General properties of MM-estimators.

- If $g(\cdot)$ is a continuous function, then $\widehat{\theta} = g(\widehat{m}_1, \widehat{m}_2, \ldots, \widehat{m}_k)$ is a consistent estimator of $\theta = g(m_1, m_2, \ldots, m_k)$.
- However, in general, MM-estimators are not necessarily unbiased.
  - But sometimes yes: If $g(\cdot)$ is linear function, then e.g. $g\left(\overline{X}\right)$ is an unbiased estimator of $g(\mu)$.
- MM-estimators only use the information in some moments (and not in the whole distribution), so they are not always efficient.
- But often they are easy to calculate and robust to misspecifications.

2. Maximum Likelihood (ML) Estimators

Maximum Likelihood (ML) estimators.

- Assume that the true p.d.f. of the population is $f(x;\theta)$ with some parameter (vector) θ.

- If θ is fixed, then the "probability" (or likelihood) of drawing any particular i.i.d. random sample $x_1, x_2, \ldots, x_n$ from this population is $f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta)$.
- This is called the *likelihood function*: $\mathcal{L}(x_1, x_2, \ldots, x_n; \theta)$.
- The ML-estimator of θ is the value for which $\mathcal{L}(x_1, x_2, \ldots, x_n; \theta)$ attains its maximum: $\widehat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(x_1, x_2, \ldots, x_n; \theta)$.
- Intuitively, we choose a θ so that the likelihood of the sample that we have should be maximized.

Example: ML-estimator for the population proportion p.

- Assume $X \sim Bernoulli(p)$ and $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of size n. Find the ML-estimator of p.

Solution.

- $\Pr(X = x_i) = p$ if $x_i = 1$ and $\Pr(X = x_i) = 1 - p$ if $x_i = 0$.
- So in the likelihood function the number of p-s is equal to the number of 1-s (or $\sum_{i=1}^{n} x_i$) and the number of $(1-p)$-s is equal to the number of 0-s (or $n - \sum_{i=1}^{n} x_i$): $\mathcal{L}(x_1, x_2, \ldots, x_n; p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$.
- To obtain $\widehat{p}_{ML}$, we have to maximize $\mathcal{L}(x_1, x_2, \ldots, x_n; p)$ with respect to p.
- It is equivalent to maximizing $\log \mathcal{L}$ with respect to p: $\log \mathcal{L} = \sum_{i=1}^{n} x_i \log p + (n - \sum_{i=1}^{n} x_i) \log(1-p)$.
- $$\frac{\partial \log \mathcal{L}}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0,$$
- and the solution of this is $\widehat{p}_{ML} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$. (It can be shown that it is indeed a maximum.)

Example: ML-estimators for the parameters of the normal distribution.

- Assume $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of size n. Find the ML-estimator of μ and σ^2.

Solution.

- $f\left(x_i;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$, so
-

$$\begin{aligned}\mathcal{L}\left(x_1,x_2,\ldots,x_n,\mu,\sigma^2\right) &= f\left(x_1\right)\cdot f\left(x_2\right)\cdot\ldots\cdot f\left(x_n\right)\\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}e^{-\sum_{i=1}^n\frac{(x_i-\mu)^2}{2\sigma^2}},\end{aligned}$$

which should be maximized with respect to μ and σ^2.

- It is equivalent to maximizing the log-likelihood:

$$\log\mathcal{L}\left(x_1,x_2,\ldots,x_n,\mu,\sigma^2\right)=\sum_{i=1}^n\log f\left(x_i\right)$$

$$=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\sum_{i=1}^n\frac{\left(x_i-\mu\right)^2}{2\sigma^2}.$$

Solution (cont.).

- Maximize this with respect to μ and σ^2:

$$\begin{aligned}\frac{\partial\log\mathcal{L}}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n\left(x_i-\mu\right)=0\\ \frac{\partial\log\mathcal{L}}{\partial\left(\sigma^2\right)} &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^n\left(x_i-\mu\right)^2=0,\end{aligned}$$

- from which we have $\widehat{\mu}_{ML}=\overline{X}$ and $\widehat{\sigma^2}_{ML}=s^2$. (It can be shown that it is indeed a maximum.)

ML-estimator of θ in an exponential sample.

- Assume $X\sim Exp(\theta)$ and $X_1,X_2,\ldots,X_n$ is an i.i.d. sample of size n. Find the maximum likelihood estimator of θ.
- Solution:
  - The likelihood of observing x_i is $f(x_i;\theta)=\frac{1}{\theta}e^{-\frac{x_i}{\theta}}$,
  - so the likelihood function is $\mathcal{L}\left(x_1,x_2,\ldots,x_n;\theta\right)=\frac{1}{\theta^n}e^{-\frac{\sum_{i=1}^n x_i}{\theta}}$.
  - The log-likelihood is $\log\mathcal{L}\left(x_1,x_2,\ldots,x_n;\theta\right)=-n\log\theta-\frac{\sum_{i=1}^n x_i}{\theta}$.

- The first-order condition is

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0,$$

- which gives $\widehat{\theta}_{ML} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$.

General properties of ML-estimators.

- ML-estimators are not necessarily unbiased.
  - Example: $\widehat{\sigma^2}_{ML} = s^2$ for normal distribution.
- Under general "regularity" conditions they are consistent.
- Under general conditions they are asymptotically normally distributed with a computable variance: $\widehat{\theta}_{ML} \overset{A}{\sim} N\left(\theta, -\left[E\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]^{-1}\right)$
- and under these conditions they are the most asymptotically efficient estimators.
- (The above properties hold if the distribution is correctly specified.)

Supplementary material: Cramer-Rao bound.

- *Cramer-Rao theorem*: if $\mathcal{L}(x_1, x_2, \ldots, x_n; \theta)$ is the likelihood function and $\widehat{\theta}$ is an *unbiased* estimator, then under general conditions
  $Var\left(\widehat{\theta}\right) \geq -\left[E\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]^{-1}$.
- (Here $-\left[E\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]^{-1}$ is called the "Cramer-Rao Lower Bound".)
- If $\widehat{\theta}_{ML}$ is unbiased, it often reaches this lower bound and hence it is the best unbiased estimator.
- Examples:
  - If $X \sim N(\mu, \sigma^2)$, then $\overline{X}$ is the best unbiased estimator for μ.
  - If $X \sim Bernoulli(p)$, then $\overline{X}$ is the best unbiased estimator for p.
  - Note: we already know that $\overline{X}$ is always BLUE for the population mean.

3. Least Squares (LS) Estimators

Least Squares Estimators (LS).

- LS-estimator: we choose $\widehat{\theta}$ in such a way that the sum of squared deviations of the sample elements from their expected values should be as small as possible.
- Example: suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of size n. The least squares estimator of $E(X) = \mu$ is obtained by minimizing $\sum_{i=1}^{n} (X_i - \widehat{\mu})^2$, which gives $\widehat{\mu}_{LS} = \overline{X}$. (It was proven earlier.)
  - So for normal distribution, $\widehat{\mu}_{LS} = \overline{X}$.
  - For Bernoulli-distribution, $\widehat{p}_{LS} = \overline{X}$.
  - For exponential distribution, $\widehat{\theta}_{LS} = \overline{X}$.

4. Material

Material.

- W Appendix C.4
- CB 7.2.1-7.2.2 (except for Example 7.2.3, 7.2.9), 10.1.1
  - only to the extent covered in the course