Statistics

Week 4: Maximum likelihood and estimator evaluation

Etienne Wijler

Econometrics and Data Science
Econometrics and Operations Research
Bachelor Program

VRIJE UNIVERSITEIT AMSTERDAM

SCHOOL OF BUSINESS AND ECONOMICS

Course overview: MLE and estimator evaluation

P4: Estimation

Week 1 Probability Recap
Week 2 Statistical Models
Week 3 Data Reduction and MME
Week 4 MLE and Evaluation
Week 5 Estimator Optimality
Week 6 Consistency

P5: Inference

Week 7 Hypothesis testing
Week 8 Mean and Variance testing
Week 9 Finding test statistics
Week 10 Evaluating tests
Week 11 Interval estimation
Week 12 Asymptotic tests

A new perspective

Which value of the parameter in my statistical model would make the observed sample "most likely"?

Before: calculate probabilities of *different events* with a distribution depending on *selected parameters*.

New: calculate the probability density of a *selected event*, i.e. the sample outcome, under *different parameters*.

Mathematically: the difference boils down to treating $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{x}$ (before) or $\boldsymbol{\theta}$ (new).

ML: answering the above question = maximum likelihood!

Building intuition for maximum likelihood (i)

Example

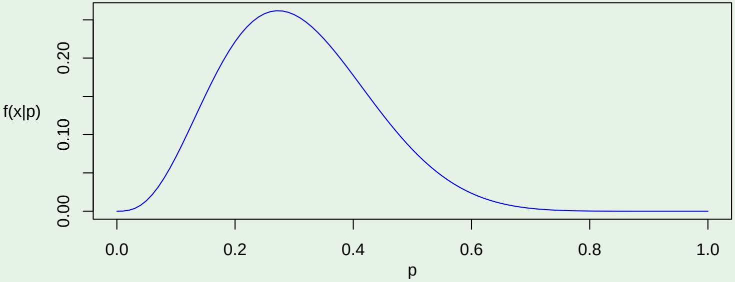
Suppose that we are given a coin that is twice a likely to fall on one side, but we do not know which side. We formulate the statistical model $\{\text{Bernoulli}(p) \mid p \in \{1/3, 2/3\}\}$. Suppose that we get a sample $x_1, \ldots, x_{11}$, of which ten are heads and one tails. What do you think p_0 is? What would be a sensible estimator?

Note: you are implicitly answering the question "which value for p_0 is most likely, given the data?"

Building intuition for maximum likelihood (ii)

Example

Now consider the statistical model $\{\text{Bernoulli}(p) \mid p \in [0,1]\}$ with observed sample $x_1, \ldots, x_{11}$, of which 3 are heads. We plot $\mathbb{P}_p(\sum_{i=1}^{11} X_i = 3)$ as a function of p:



What is the most likely value for p_0, given the data we have observed?

Maximum Likelihood estimator

Definition

The likelihood function for a statistical model is defined as the function

$$L(\theta \mid \boldsymbol{x}) : \theta \to f(\boldsymbol{x} \mid \theta)$$

Definition (7.2.4)

The maximum likelihood estimate $W(\boldsymbol{x})$ is the parameter value in Θ at which $L(\theta \mid \boldsymbol{x})$ attains its maximum:

$$W(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} L(\theta \mid \boldsymbol{x}).$$

The maximum likelihood estimator of θ_0 is the accompanying estimator $W(\boldsymbol{X})$.

Computing the MLE (i)

Problem 1: finding the global maximum of $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ can be difficult sometimes.

Idea: by assuming that $X_1, \ldots, X_n$ are iid, we simplify the likelihood function to

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(x \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} g(x_i \mid \boldsymbol{\theta})$$

Solution 1: use calculus (setting derivative equal to zero) to find the maximizer.

Problem 2: The derivative of $\prod_{i=1}^{n} g(x_i \mid \boldsymbol{\theta})$ will involve a painful amount of chain rules.

Computing the MLE (ii)

Idea: for any monotonically increasing function h, it holds that

$$\arg\max_y f(y) = \arg\max_y h(f(y))$$

Solution 2: taking the logarithm gets rid of the product and does not change the maximizer!

Definition

We define the log likelihood as the function $\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) : \theta \mapsto \log L(\theta \mid \boldsymbol{x})$.

Note: in our setting $\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \log g(x_i \mid \boldsymbol{\theta})$.

Computing the MLE (iii)

Roadmap: If $\ell(\theta \mid \boldsymbol{x})$ is differentiable, then we:

1. Find stationary points $\tilde{\theta} \Rightarrow \frac{d}{d\theta}\ell(\theta \mid \boldsymbol{x}) \stackrel{s}{=} 0$.
2. Check if stationary points are local maxima $\Rightarrow \frac{d^2}{d\theta^2}\ell(\theta \mid \boldsymbol{x})\Big|_{\theta=\tilde{\theta}} < 0$.
3. Evaluate $\ell(\theta \mid \boldsymbol{x})$ at the local maxima and the boundary points of Θ.

Exception: we may omit checking the boundary points if the stationary point is unique and (ii) $\frac{d^2}{d\theta^2}\ell(\theta \mid \boldsymbol{x}) < 0$.

Lemma

Let $\Theta \in \mathbb{R}$ be an interval and let $h : \Theta \to \mathbb{R}$ be a function that is twice differentiable. Suppose that there exists a unique stationary point $\tilde{\theta} \in \Theta$, that is $h'(\tilde{\theta}) = 0$, and that it satisfies the second derivative test $h''(\tilde{\theta}) < 0$. Then $\tilde{\theta}$ is the unique point at which the global maximum is attained.

MLE: examples

Example (Please stay on the line)

To optimise staff allocation at a call-center you are analysing customer waiting times. Let X_i, $i = 1, \ldots, n$, denote the total waiting time of the i-th customer. Your statistical model is $\{\text{Exponential}(\lambda) \mid \lambda > 0\}$. Find the MLE of λ_0.

Example (Too much air in my chips)

To measure whether a potato chips producer operates within legal margins, you investigate the weight of chips of bag. Let X_i, $i = 1, \ldots, n$ denote the weight (in grams) of the i-th bag of chips. Your statistical model is $\{\text{Normal}(\mu, 1) \mid \mu > 0\}$. Find the MLE of μ_0.

MLE: more examples

Example (virus recovery time)

Say that you believe that recovery from a virus takes a minimum number of hours θ and the chance of longer recovery times decreases exponentially fast in the number of hours. The statistical model is $\{g(x \mid \theta) \mid \theta > 0\}$ with

$$g(x \mid \theta) = e^{-(x-\theta)}, \quad x \geq \theta.$$

Note that this function is not continuous in θ over $\mathbb{R}_+$. Find the MLE of θ_0.

Example (More chips!)

Let's revisit our potato chips example. This time, let X_i denote the deviation from the stated content on bag i and let's assume we don't know the variance. Our statistical model is $\{\text{Normal}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 > 0\}$. Find the MLE of $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)$.

Functions of parameters

Suppose we have a statistical model $\{f(\boldsymbol{x} \mid \theta) \mid \theta \in \Theta\}$.

Recall that θ_0 is the unique parameter in Θ such that $f(x \mid \theta) = f(x)$ (the DGP).

However, we are often interested in a certain attribute of $f(x)$, such as $\mathbb{E}(X) = \int x f(x) dx$ or $P(X \leq a) = \int_{-\infty}^{a} f(x) dx$.

Important: these attributes are typically not equal to $\boldsymbol{\theta}_0$, but will be functions of θ_0, say $\eta_0 = \tau(\theta_0)$!

Example (Light bulb)

Suppose we are interested in the expected lifetime of a lightbulb. Lifetimes are commonly modelled via $\{\text{Exponential}(\lambda) \mid \lambda > 0\}$. Letting X denote the unobserved lifetime of light bulb, we are interested in estimating $\mathbb{E}(X) = \frac{1}{\lambda} =: \eta_0$.

Invariance

Intuitively, if we already have derived the MLE of θ_0, say $\hat{\theta}$, then we could aim to estimate η_0 via $\hat{\eta} = \tau(\hat{\theta})$.

Alternatively, we could redefine the likelihood in terms of η_0 and then derive the MLE η_0.

Amazingly: these two approaches are always equivalent!

Theorem (7.2.10)

Let $\hat{\theta}$ be the MLE of θ_0. Then, for any function τ, it holds that $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

Note: while this result is not surprising for one-to-one functions, it holds for any function! Try it out yourself.

Course overview: MLE and estimator evaluation

P4: Estimation

- Week 1 Probability Recap
- Week 2 Statistical Models
- Week 3 Data Reduction and MME
- Week 4 MLE and Evaluation
- Week 5 Estimator Optimality
- Week 6 Consistency

P5: Inference

- Week 7 Hypothesis testing
- Week 8 Mean and Variance testing
- Week 9 Finding test statistics
- Week 10 Evaluating tests
- Week 11 Interval estimation
- Week 12 Asymptotic tests

Choosing between estimators

So far, we discussed methods to find estimators (intuition, MM and ML).

Problem: the estimators may provide very different estimates in any given application. Which one should we believe?

Example (Uniform(0,θ))

We study the statistical model $\{\text{Uniform}(0, \theta) \mid \theta > 0\}$. Common estimators are given by $\hat{\theta}_{MM} = 2\overline{X}$ and $\hat{\theta}_{ML} = X_{(n)}$ (see tutorial exercises). Suppose we know that $\theta_0 = 10$ and that we have two separate experiments. In the first experiment we observe $\boldsymbol{x} = (6, 2, 5, 7, 2)$ and in the second experiment we observe $\boldsymbol{x} = (3, 2, 5, 9, 2)$. Which estimator is best? Do you believe one of these estimators is always optimal?

Note: In practice, we of course don't know θ_0, further complicating the problem.

Performance measures

Fact: a good estimate $W(\boldsymbol{x})$ for $\tau(\theta_0)$ is such that $W(\boldsymbol{x}) - \tau(\theta_0)$ is close to zero.

Problem: The estimation error $W(\boldsymbol{x}) - \tau(\theta_0)$ depends on the realization of the random vector $\boldsymbol{X}$.

Idea: require that $\mathbb{E}\left(W(\boldsymbol{x}) - \tau(\theta_0)\right)$ is close to zero.

Problem 2: the above expectation depends on θ_0, which is unknown!

Solution 2: require that $\mathbb{E}\left(W(\boldsymbol{x}) - \tau(\theta_0)\right)$ is close to zero for all $\theta \in \Theta$.

Unbiased estimators

Definition (7.3.2)

The bias of an estimator $W(\boldsymbol{X})$ of $\tau(\theta_0)$ is defined as the function $\theta \mapsto \mathbb{B}\mathrm{ias}_\theta(W) = \mathbb{E}_\theta(W(\boldsymbol{X}) - \tau(\theta))$. An estimator of $\tau(\theta_0)$ is unbiased if $\mathbb{E}_\theta(W(\boldsymbol{X})) = \tau(\theta)$ for all $\theta \in \Theta$.

Note: Unbiased estimators give the correct answer "on average". Equivalently, they do not consistently over- or underestimate.

Example

Suppose we have the model $\{\text{Bernoulli}(p) \mid p \in [0,1]\}$. The MME of p_0 is given by $\hat{p}_{MM} = \overline{X}$ Is this an unbiased estimator? Let the random variable Y be 0.1 with probability 0.9 and 9.1 with probability 0.1 and Define $Z = Y\overline{X}$. Is Z an unbiased estimator? Which estimator would you intuitively prefer?

Distance measures: MAE and MSE

Problem: extreme positive and negative values may offset each other, resulting in small or zero bias.

Solution: incorporate distance measures to quantify the expected estimation error.

Definition (7.3.1)

The mean absolute error (MAE) of an estimator W of $\tau(\theta_0)$ is the function $\text{MAE}(\theta, W) : \theta \to \mathbb{E}_\theta \|W(\boldsymbol{X}) - \tau(\theta)\|$. The mean squared error (MSE) of an estimator W of $\tau(\theta_0)$ is the function $\text{MSE}(\theta, W) : \theta \to \mathbb{E}_\theta \|W(\boldsymbol{X}) - \tau(\theta)\|^2$.

Note: for any vector $\boldsymbol{y} = (y_1, \ldots, y_n)'$, we define $\|\boldsymbol{y}\| = \sum_{i=1}^n |y_i|$ and $\|\boldsymbol{y}\|^2 = \sum_{i=1}^n y_i^2$.

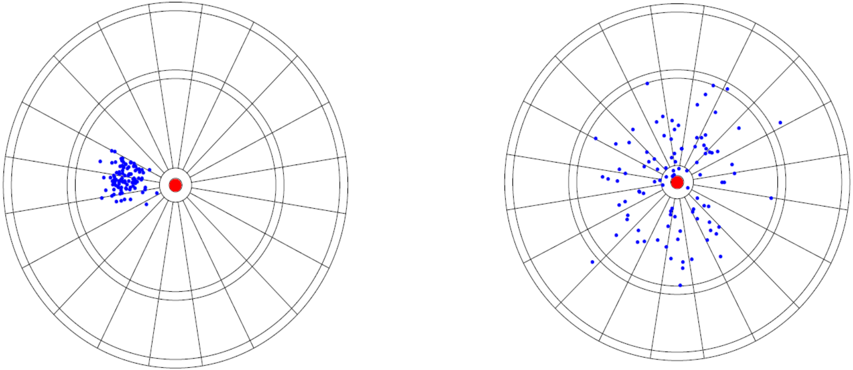
Mean-squared error decomposition

Note: the mean-squared error is the most popular distance measure. It has nice theoretical properties and satisfies an intuitive decomposition in the scalar case.

$$
\begin{aligned}
\mathrm{MSE}(\theta, W) &= \mathbb{E}_\theta \|W(\boldsymbol{X}) - \tau(\theta)\|^2 = \mathbb{E}_\theta (W(\boldsymbol{X}) - \tau(\theta))^2 \\
&= \mathbb{E}_\theta \left(W(\boldsymbol{X}) - \mathbb{E}_\theta(W(\boldsymbol{X})) + \mathbb{E}_\theta(W(\boldsymbol{X})) - \tau(\theta)\right)^2 \\
&= \mathbb{E}_\theta \left(W(\boldsymbol{X}) - \mathbb{E}_\theta(W(\boldsymbol{X}))\right)^2 + 2\, \underbrace{\mathbb{E}_\theta \left((W(\boldsymbol{X}) - \mathbb{E}_\theta(W(\boldsymbol{X}))(\mathbb{E}_\theta(W(\boldsymbol{X})) - \tau(\theta)\right)}_{=0} \\
&\quad + \mathbb{E}_\theta \left(\mathbb{E}_\theta(W(\boldsymbol{X})) - \tau(\theta)\right)^2 \\
&= \mathbb{E}_\theta \left(W(\boldsymbol{X}) - \mathbb{E}_\theta(W(\boldsymbol{X}))\right)^2 + \left(\mathbb{E}_\theta(W(\boldsymbol{X})) - \tau(\theta)\right)^2 \\
&= \mathbb{V}\mathrm{ar}_\theta(W(\boldsymbol{X})) + \mathbb{B}\mathrm{ias}_\theta(W(\boldsymbol{X}))^2
\end{aligned}
$$

Hence, the MSE decomposed into the sum of the variance and the squared bias!

Bias versus variance



(a) Large bias, small variance (b) Small bias, large variance

Figure: Bias versus variance

Uniformly better estimators

Note: The MSE gives us, for each $\theta \in \Theta$, a measure to evaluate how well an estimator of θ_0 performs.

Problem: say we have two estimators W_1, W_2 with $\mathrm{MSE}(\theta_1, W_1) \leq \mathrm{MSE}(\theta_1, W_2)$ and $\mathrm{MSE}(\theta_2, W_1) \geq \mathrm{MSE}(\theta_2, W_2)$ for some $\theta_1, \theta_2 \in \Theta$. We don't know θ_0 so we don't know whether W_1 or W_2 is the best estimator!

Definition

We call an estimator W_1 uniformly better than an estimator W_2 if $\mathrm{MSE}(\theta, W_1) \leq \mathrm{MSE}(\theta, W_2)$ for all $\theta \in \Theta$.

Intuitively: an estimator with a smaller MSE for all possible parameter values is always preferred.

Estimator evaluation: examples

Example (Bernoulli revisited)

We revisit the Bernoulli example with estimators $W_1(\boldsymbol{X}) = \bar{X}$ and $W_2(\boldsymbol{X}) = Y\bar{X}$. Is one estimator uniformly better?

Example (Uniform with bias correction)

Consider the statistical model $\{\text{Uniform}(0, \theta) \mid \theta > 0\}$. The MME of θ_0 is given by $\hat{\theta}_{MM} = 2\overline{X}$ and the MLE by $\hat{\theta}_{ML} = X_{(n)}$.

1. Are these estimators unbiased?
2. Construct an unbiased estimator based on $\hat{\theta}_{ML}$.
3. Evaluate which of these estimator is uniformly better than the others.