Constructing estimators

James Scott (UT-Austin)

Outline

- Method of moments
- Maximum likelihood
- Evaluating estimators

The "facts of life" about estimators

We'll talk about two ways of creating sensible estimators:

- method of moments (MM)
- maximum likelihood

There are tons of others we won't cover: generalized method of moments, Bayes estimators, MAP estimators, shrinkage estimators, penalized likelihood, minimum-variance unbiased estimators, generalized estimating equations, maximum entropy, minimum description length…

A preliminary note

This whole discussion today assumes that we observe IID data $X_1, \ldots, X_N$ arising from a parametric probability model with parameter θ:

$$X_i \overset{iid}{\sim} f(x \mid \theta)$$

And that our goal is to estimate either θ itself or some function of the parameter $g(\theta)$. Whatever we're trying to estimate is called the **estimand.**

Note: θ might be a vector in multi-parameter models. For example, $\theta = (\mu, \sigma^2)$ in a normal model.

Method of moments

The principle behind the method of moments is very simple to state:
choose the parameter θ so that the theoretical moments and sample moments are identical.
(Remember, moments are means, variances, etc.)

Example: suppose we observe $X_i \overset{iid}{\sim} \mathrm{Poisson}(\lambda)$, and λ is unknown.

- Theoretical mean: $E(X_i) = \lambda$
- Sample mean: $\bar{X}_n$
- MoM estimator: equate the two, setting $\hat{\lambda}_n = \bar{X}_n$.

Method of moments

Here's the general principle. Let θ be a K-dimensional parameter. Define the theoretical moments as the following function of θ:

$$\alpha^{(k)}(\theta) = E(X^k \mid \theta)$$

and the sample moments as the follow function of the data:

$$\hat{\alpha}_n^{(k)} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

The law of large numbers says that eventually, $\hat{\alpha}_n^{(k)}$ converges in probability to $\alpha^{(k)}(\theta_0)$, where θ_0 is the true parameter.

Method of moments

The method of moments estimator $\hat{\theta}_{MM}$ solves the following system of K equations:

$$\begin{aligned}
\alpha^{(1)}(\hat{\theta}) &= \hat{\alpha}_n^{(1)} \\
\alpha^{(2)}(\hat{\theta}) &= \hat{\alpha}_n^{(2)} \\
&\vdots \\
\alpha^{(K)}(\hat{\theta}) &= \hat{\alpha}_n^{(K)}
\end{aligned}$$

This is a system of K equations in K unknowns and should therefore (usually!) have a unique solution.

Method of moments

So the general recipe for calculating the method of moments estimator of a K-dimensional parameter is:

Example 1: binomial proportion

Suppose $X_i \sim \mathrm{Bern}(p)$ for $i = 1, \ldots, N$. What is $\hat{p}$ under the method of moments? Note: here $K = 1$.

Example 1: binomial proportion

Step 1: use probability theory to write down expressions for the theoretical moments as a function of p.

We've done this before:

$$\alpha^{(1)}(p) = E(X^1 \mid p) = p$$

Example 1: binomial proportion

Step 2: Calculate the sample moments:

$$\hat{\alpha}_n^{(1)} = \frac{1}{n} \sum_{i=1}^{n} X_i^1 = \bar{X}_n$$

Example 1: binomial proportion

Step 3: Set up the system of K equations, $\alpha^{(k)}(\theta) = \hat{\alpha}_n^{(k)}$ for $k = 1, \ldots, K$.

Here $K = 1$, so it's a system of one equation:

$$\alpha^{(1)}(\theta) = \hat{\alpha}_n^{(1)}$$

So

$$\bar{X}_n = p$$

Step 4 (solve the system for p) is easy: this equation is already solved! The MoM estimator is $\hat{p} = \bar{X}_n$.

Example 2: normal distribution

Suppose we assume that our data $X_1, \ldots, X_n$ comes from a normal distribution: $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

The unknown parameter vector is $\theta = (\mu, \sigma^2)$. Here $K = 2$.

Example 2: normal distribution

Step 1: use probability theory to write down expressions for the theoretical moments as a function of $\theta = (\mu, \sigma^2)$.

Here we need two moments, since $K = 2$:

$$\alpha^{(1)}(\theta) = E(X^1 \mid \theta) = \mu$$
$$\alpha^{(2)}(\theta) = E(X^2 \mid \theta) = \sigma^2 + \mu^2$$

The second equation follows from the fact that $\mathrm{var}(X) = E(X^2) - E(X)^2$.

Example 2: normal distribution

Step 2: Calculate the sample moments

$$\hat{\alpha}_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i^1 = \bar{X}_n$$

$$\hat{\alpha}_n^{(2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = S_X^2$$

Example 2: normal distribution

Step 3: Set up the system of K equations, $\alpha^{(k)}(\theta) = \hat{\alpha}_n^{(k)}$ for $k = 1, \ldots, K$.

Here we have a system of two equations in two unknowns:

$$\mu = \bar{X}_n$$

$$\sigma^2 + \mu^2 = S_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Example 2: normal distribution

Step 4: Solve the system of equations. Clearly the first equation is solved at $\mu = \bar{X}_n$. So the second equation is solved at

$$\sigma^2 + \bar{X}_n^2 = S_X^2$$

Or equivalently,

$$\sigma^2 = S_X^2 - \bar{X}_n^2$$

Thus the method of moments estimator is $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\bar{X}_n, S_X^2 - \bar{X}_n^2)$.

See `gas_method_moments.R`.

Method of moments: summary

In most non-crazy situations, the method of moments estimator converges in probability to the right answer:

$$\hat{\theta}_{MM} \xrightarrow{P} \theta$$

Note: we say that the estimator is *consistent.* Consistency means "converging in probability to the right answer with more data."

Method of moments: summary

Similarly, in most non-crazy situations, the method of moments estimator is asymptotically normal:

$$Z_n = \frac{\hat{\theta}_{MM} - \theta}{\mathrm{se}(\hat{\theta}_{MM})} \rightsquigarrow N(0, 1)\,.$$

Thus we can make approximate probability statements about the estimator using the normal distribution.

Note: there is a more general version of the method of moments, called the "generalized method of moments" (GMM). This is **wildly popular** in econometrics. To understand GMM (not covered here), you have to understand MM.

Maximum likelihood

Outside of econometrics, the most popular way to construct estimators is by the **principle of maximum likelihood.** Suppose, as before, we observe IID data $X_1, \ldots, X_n$ from some unknown parametric model with parameter θ: $X_i \sim f(X \mid \theta)$.

The **likelihood function** is defined as follows:

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \,.$$

Note: if X is discrete, then f refers to the PMF; if X is continuous, then f refers to the PDF.

The likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \,.$$

The likelihood function is a function of the parameter θ:

- you plug in some particular value of $\theta\ldots$
- it spits out some number called the "likelihood" at θ.
- It measures how likely the data is, assuming that the true parameter is equal to θ.
- The product form of the likelihood comes from the assumption of independence.

An example

Suppose we observe Bernoulli trials, $X_i \sim \mathrm{Bern}(p)$.

- The PMF of a Bernoulli random variable is $f(x \mid p) = p^x(1-p)^{1-x}$.
- So the likelihood function is

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} f(X_i \mid p) \\
&= \prod_{i=1}^{n} p_i^X (1-p)^{1-X_i} \\
&= p^Y (1-p)^{(n-Y)}
\end{aligned}
$$

where $Y = \sum_{i=1}^{n} X_i$.

An example

Suppose we observe $Y = 12$ successes (1) out of $n = 20$ Bernoulli trials.

Let's try calculating the likelihood function at two different values:

- $p = 0.4$
- $p = 0.7$

An example

At $p = 0.4$, the likelihood function is

$$L(0.4) = 0.4^{12}(1 - 0.4)^{(20-12)} = 2.82 \times 10^{-7}$$

Interpretation: if $p = 0.4$, the probability of observing this data set X with 12 successes and 8 failures is 2.82×10^{-7}.

An example

At $p = 0.7$, the likelihood function is

$$L(0.7) = 0.7^{12}(1 - 0.7)^{(20-12)} = 9.08 \times 10^{-7}$$

Interpretation: if $p = 0.7$, the probability of observing this data set X with 12 successes and 8 failures is 9.08×10^{-7}.

An example

So it looks like our data would have been more likely to arise if $p = 0.7$, versus $p = 0.4$:

- $L(0.4) = 2.82 \times 10^{-7}$.
- $L(0.7) = 9.08 \times 10^{-7}$.

Conclusion: $p = 0.7$ is a better (higher likelihood) guess for the parameter. If these were your only two choices for p, you'd probably choose $p = 0.7$.

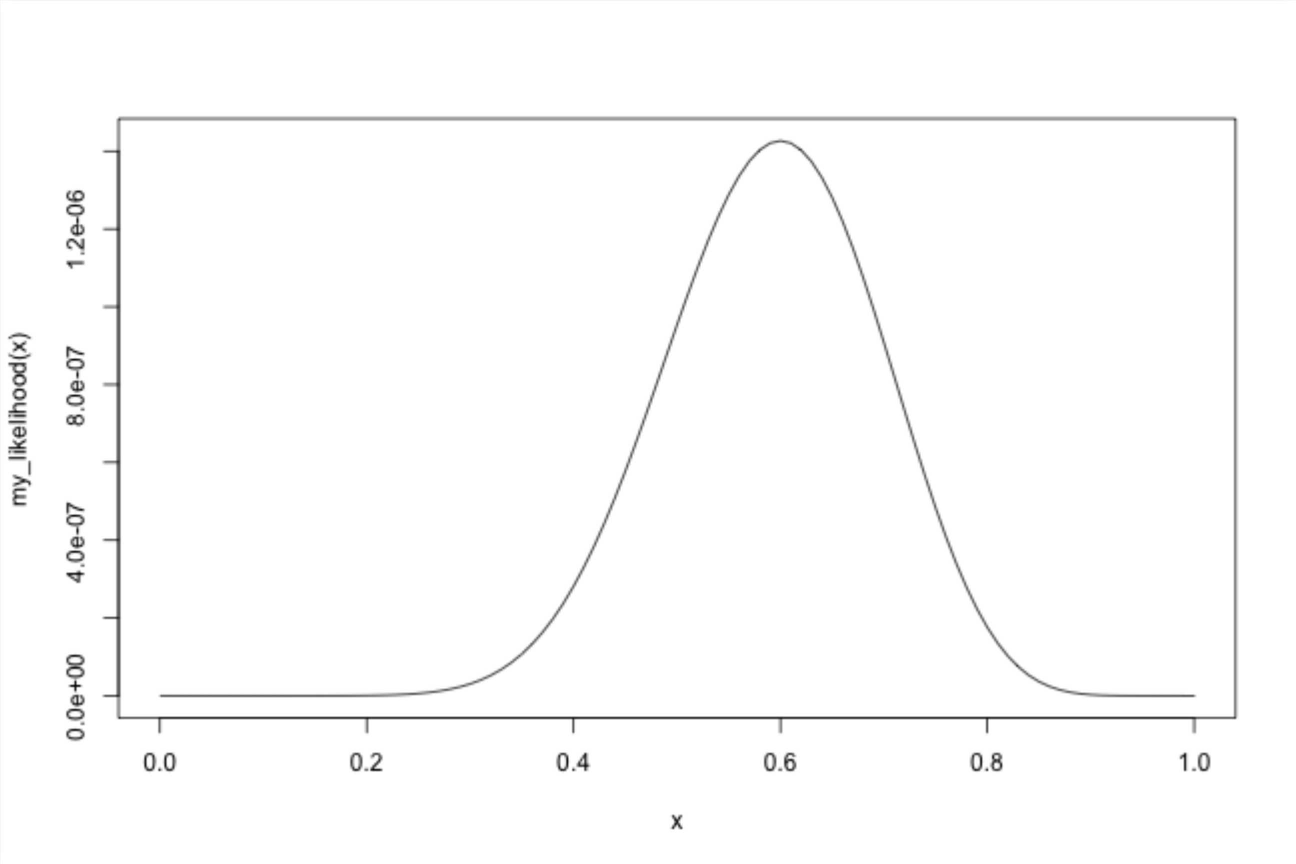
An example

But of course, those aren't the only two choices!

You can guess any probability between 0 and 1.

So let's plot the likelihood as a function of all possible guesses $p = (0, 1)$.

An example



It looks like $p = 0.6$ is the choice of p that makes the data look most likely. It is the **maximum likelihood estimate**, or MLE.

The MLE

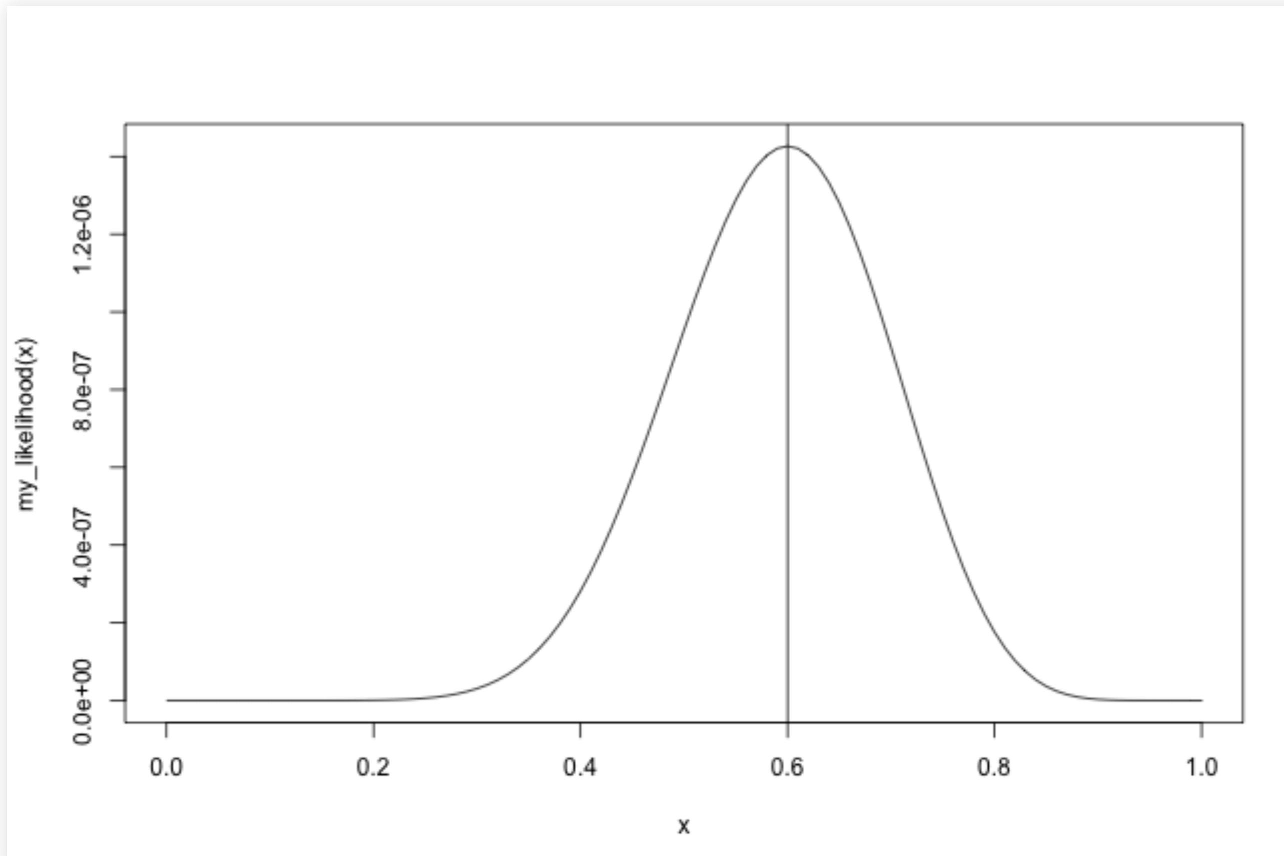
The maximum likelihood estimate (MLE) is the value of θ that maximizes $L(\theta)$, the likelihood function.

Equivalently, the MLE is the value of θ that maximizes the logarithm of the likelihood function,

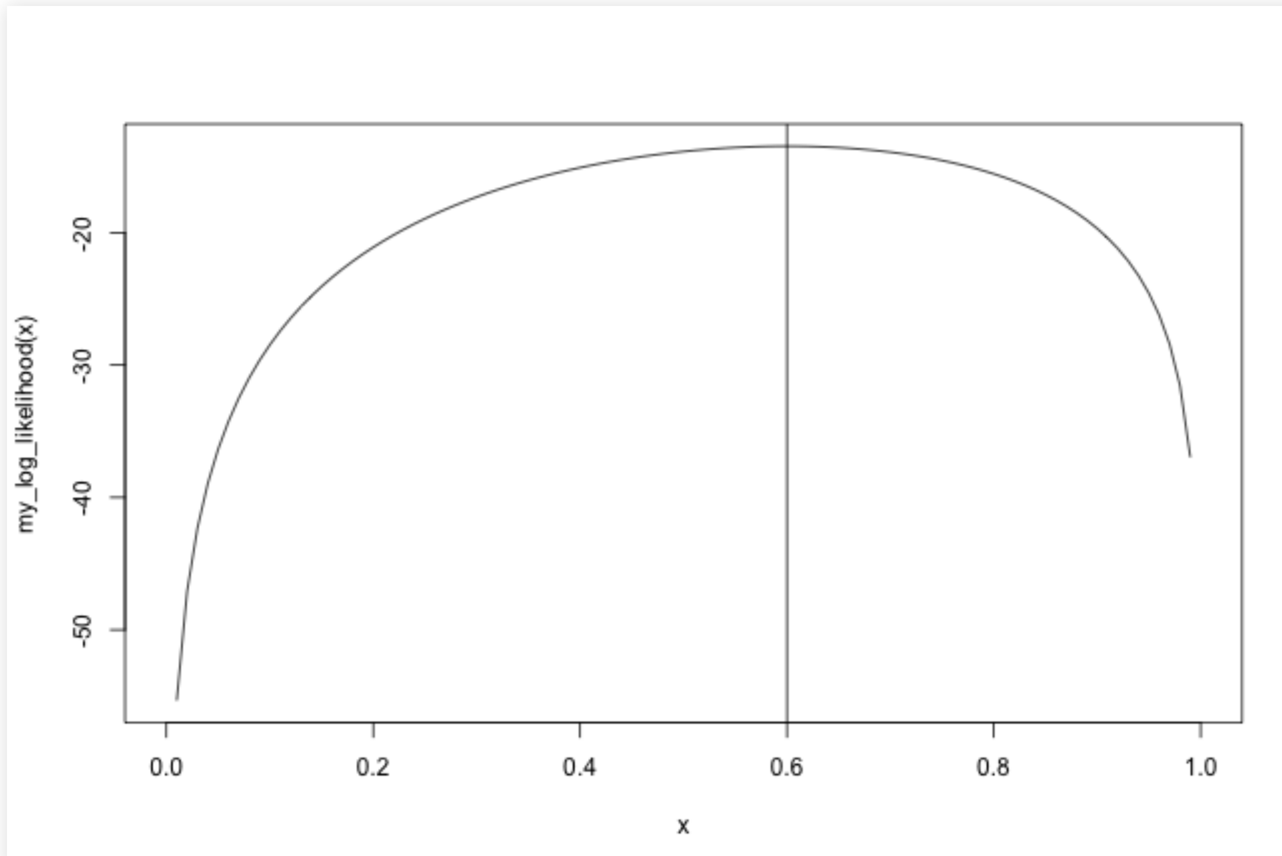
$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

Taking the log doesn't change the answer (since log is a monotonic transformation). But it does avoid the problem of the likelihood becoming so small that it can't be represented using the floating-point numerical system on a computer.

The MLE: original likelihood



The MLE: log likelihood



The MLE: a helpful fact

Fact: if we multiply $L(\theta)$ by any positive constant c, we will not change the MLE. This allows us to be sloppy about ignoring multiplicative constants in the likelihood function.

The MLE: a painful fact

How do we actually calculate the MLE?

Answer: calculus. Take the derivative of the log likelihood function with respect to θ, and set it equal to 0.

The MLE: example 2

Suppose $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. Let's derive the MLE for $\theta = (\mu, \sigma^2)$ together on the board.

Properties of the MLE

- It is consistent: $\hat{\theta}_{MLE} \xrightarrow{P} \theta$.
- It is invariant to transformations: if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- It is asymptotically normal: $(\hat{\theta} - \theta)/\widehat{se} \rightsquigarrow N(0, 1)$.
- It is *asymptotically efficient*: this means, roughly, that the MLE has the smallest variance among all "well-behaved" estimators, at least for large samples.