Lecture 3: Estimation frameworks

Mitch Downey[1]

January 31, 2024

[1] IIES, Stockholm University. Slides heavily based on those developed by Markus Jäntti based on the textbook by Bruce Hansen.

Today

- Key question of Estimation (course part 1):
  What is a “good” guess for some parameter’s value?
- Approaches to generating an estimator:
  - Maximum likelihood
  - Bayesian estimation
  - Method of moments
- What are the properties of these estimators?
- What are the costs and benefits?
- Today’s results are relevant for:
  - both structural work and reduced form work
  - both regressions and non-regressions
  - both descriptive work and casual work

The maximum likelihood estimator

- Consider two iid observations: y_1, y_2
- Because independent: $f(y_1, y_2|\theta_1, \theta_2) = f_1(y_1|\theta_1)f_2(y_2|\theta_2)$ (by factorialization of dist. fnct.)
- Because identically distributed: $f_1(y_1|\theta_1)f_2(y_2|\theta_2) = f(y_1|\theta)f(y_2|\theta)$
- More generally:

$$f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- Example: If y_i is normally distributed with mean μ and variance σ^2 then

$$f(y_1, ..., y_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

- This tells you the “probability” of observing something in the data
- But the data actually happened, so shouldn’t that tell us something about the likelihood the parameters attain certain values

Simple example: Maximizing the likelihood of the data

- I have two jars, each with 10 marbles
    - Jar 1: 8 red, 2 green
    - Jar 2: 2 red, 8 green
- I draw a marble from one jar, but you don't know which
- It's green
- Note: $n = 1, f(x|\theta) = Pr(\text{green}) = \theta, \theta \in \{.2, .8\}$
- Which jar do you think I drew the marble from?
- Two observations:
    - Reliable statistical procedures are usually formalizations of intuitive principals of accumulating knowledge
    - It's not impossible that I took the marble from Jar 1
        - Hypothesis testing (Markus) and asymptotic variance of estimators (later in this lecture) are about saying "Well how sure are you?"
        - Bayesian (later): Why would you only care about which jar is more likely? Wouldn't you also care about how much more likely? What if it's very important to know whether it was Jar 1 (decision theory)?

The maximum likelihood estimator

- The “probability” of observing something in the data:

$$f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- If θ is a fixed number (more controversial than it sounds; see Bayesian statistics) then $f(y_1, ..., y_n|\theta) = f(y_1, ..., y_n$ and $\theta)$
- Given that $y_1, ..., y_n$ actually happened, and that we don’t know θ, it makes sense that we’re more inclined to believe in the values of θ under which it is more likely that $y_1, ..., y_n$ would occur
- Where is that likelihood the highest?
- **The likelihood function** is the likelihood that the parameters take some value:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- Choose the θ that maximizes this: **Maximium likelihood estimation**

The maximum likelihood estimator

- Conveniently, if $g(\cdot)$ is a strictly increasing function, then whatever θ maximizes $\mathcal{L}(\theta)$ also maximizes $g\big(\mathcal{L}(\theta)\big)$
- Thus, it's often useful to work with the log-likelihood function (log is strictly increasing):

$$\ell(\theta) \equiv \ln\left(\mathcal{L}(\theta)\right) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

- Note that $\mathcal{L}(\theta)$ and $\ell(\theta)$ are both random variables, and so $\hat{\theta}_{MLE} \equiv \arg\max_{\theta} \mathcal{L}(\theta)$ is also a random variable
- How to maximize?

The maximum likelihood estimator

- Conveniently, if $g(\cdot)$ is a strictly increasing function, then whatever θ maximizes $\mathcal{L}(\theta)$ also maximizes $g\big(\mathcal{L}(\theta)\big)$
- Thus, it's often useful to work with the log-likelihood function (log is strictly increasing):

$$\ell(\theta) \equiv \ln\big(\mathcal{L}(\theta)\big) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

- Note that $\mathcal{L}(\theta)$ and $\ell(\theta)$ are both random variables, and so $\hat{\theta}_{MLE} \equiv \arg\, \max_{\theta} \mathcal{L}(\theta)$ is also a random variable
- How to maximize? Take first and second derivatives
- The first derivative of the log-likelihood function is called the **score**: $\frac{\partial}{\partial\theta}\ell(\theta)$

Second derivatives

- What about the second derivative?
- Standard approach to maximization: When $\ell'(\theta_0) = 0$ then...
  - ... ensure second derivative is negative
  - ... compare endpoints to $\ell(\theta_0)$
- Statistics: We have a similar thing, but it's more involved and more informative

Mild regularity conditions

- **Mild regularity conditions** are satisfied when a set of assumptions are met such that you can interchange integration and differentiation
- There are many examples of types of differentiation/integration that one might need, and saying "under mild regularity conditions" assumes them all.
- One example:

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}\hat{\theta}(x)f_n(x|\theta)dx=\int_{\mathbb{R}^n}\hat{\theta}(x)\frac{\partial}{\partial\theta}f_n(x|\theta)dx$$

  where $\hat{\theta}(x)$ is an estimator (which is a random variable and function of x, the data)
- Think of these as similar to a well-defined second derivative
- What are these assumptions?

Mild regularity conditions

- **Mild regularity conditions** are satisfied when a set of assumptions are met such that you can interchange integration and differentiation
- There are many examples of types of differentiation/integration that one might need, and saying "under mild regularity conditions" assumes them all.
- One example:

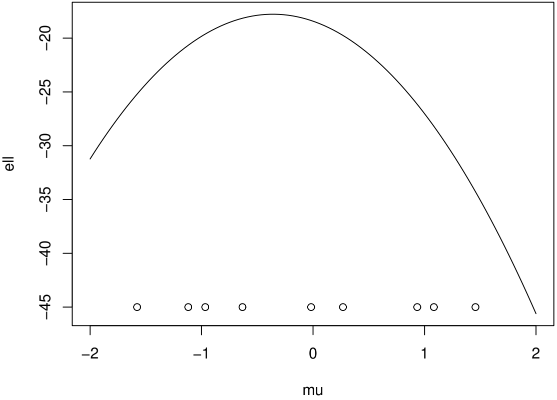
  $$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}\hat{\theta}(x)f_n(x|\theta)dx = \int_{\mathbb{R}^n}\hat{\theta}(x)\frac{\partial}{\partial\theta}f_n(x|\theta)dx$$

  where $\hat{\theta}(x)$ is an estimator (which is a random variable and function of x, the data)
- Think of these as similar to a well-defined second derivative
- What are these assumptions? I don't know...
- When do they fail? Mainly when the parameter determines the parameter space: e.g., $x \sim unif[0,\eta]$
- Note: It makes sense that the second derivative will be inadequate in these cases: We need to check whether $\ell(\theta)$ is maximized at the boundary, and derivatives won't help us
- If the parameter space is unbounded (or bounded by a constant, like zero or one) then this doesn't come up because the likelihood will fall away towards extreme (infinite) values of the space

Normal distribution: Log likelihood

$x \sim N(\mu, \sigma^2)$
$E(x) = \mu$

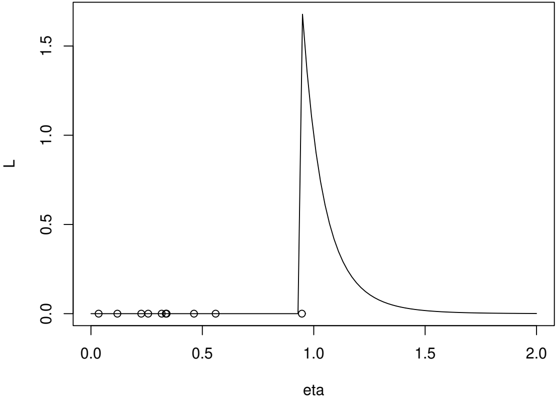
Simulation: $\mu = 0, \sigma^2 = 1$



Uniform distribution: Likelihood

$x \sim U(0, \eta)$
$E(x) = \eta/2$

Simulation: $\eta = 1$



Uniform distribution: Likelihood

$x \sim U(0, \eta)$
$E(x) = \eta/2$

Simulation: $\eta = 1$

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}\hat{\theta}(x) f_n(x|\theta)dx \neq \int_{\mathbb{R}^n}\hat{\theta}(x)\frac{\partial}{\partial\theta} f_n(x|\theta)dx$$

Regularity conditions not satisfied

What to do with mild regularity conditions?

- These are important because they let us talk about the second derivative
- This lets us apply our intuition from calculus!

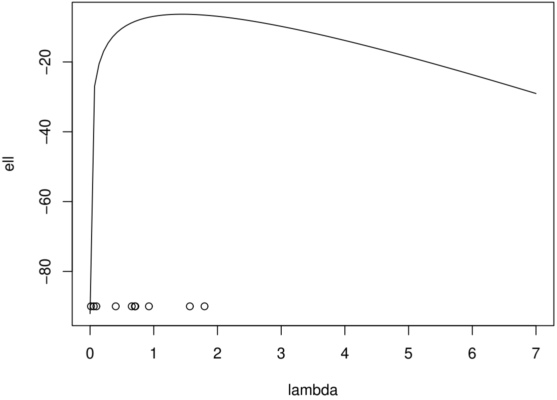
Properties of MLE

- Assume *i*) mild regularity conditions are satisfied, *ii*) the data are iid, and *iii*) the data are drawn from $f(X|\theta)$ (i.e., the model is correctly specified). Then:
  - The score's expected value is zero: $E\big(\partial \ell_n / \partial \theta\big) = 0$
    - Recall the data is a collection of random variables, the likelihood function is a function of the data, and the score is the derivative of the likelihood function, so the score is a random variable
    - This is intuitive and helpful
    - Intuitive: It increases for a while and then decreases for a while, and the range over which it is increasing is probability weighted to be the same as the range over which it is decreasing

Exponential distribution: Log likelihood

$x \sim exp(\lambda)$
$f(x|\lambda) = \lambda e^{-\lambda x}$
$E(x) = 1/\lambda$

Simulation: $\lambda = 1$



Properties of MLE

- Assume *i*) mild regularity conditions are satisfied, *ii*) the data are iid, and *iii*) the data are drawn from $f(X|\theta)$ (i.e., the model is correctly specified). Then:
  - The score's expected value is zero: $E\big(\partial\ell_n/\partial\theta\big) = 0$
    - This is intuitive and helpful
    - Helpful: Knowing that a variable is mean zero is useful when studying its variance
    - Why study the variance?
- The **Fisher Information Matrix** is defined as:

$$H_n(\theta) = E_\theta\left[\left(\frac{\partial\ell_n(\theta)}{\partial\theta}\right)\left(\frac{\partial\ell_n(\theta)}{\partial\theta'}\right)\right]$$

  where we treat $\partial\ell_n(\theta)/\partial\theta$ as a $k \times 1$ vector (when theta includes k parameters) and θ' is the transpose of θ and $E_\theta(\cdot)$ denotes the expectation under the assumption that the true parameter is θ.
  - Note: With one parameter this is a 1×1 "matrix"
  - The Fisher Information Matrix is the expected value of squared slope of the likelihood function
  - Recall that under the assumptions above, the score's expected value is zero and $Var(x) = E(x^2) - \big(E(x)\big)^2$, so $H_n(\theta)$ is $Var(d\ell_n/d\theta)$ (more generally, is the variance-covariance matrix)

Fisher Information Matrix

$$H_n(\theta) \equiv E_\theta\left[\left(\frac{\partial \ell_n(\theta)}{\partial \theta}\right)\left(\frac{\partial \ell_n(\theta)}{\partial \theta'}\right)\right] =^{\text{if iid}} nE_\theta\left[\left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)\left(\frac{\partial \ln f(x|\theta)}{\partial \theta'}\right)\right]$$

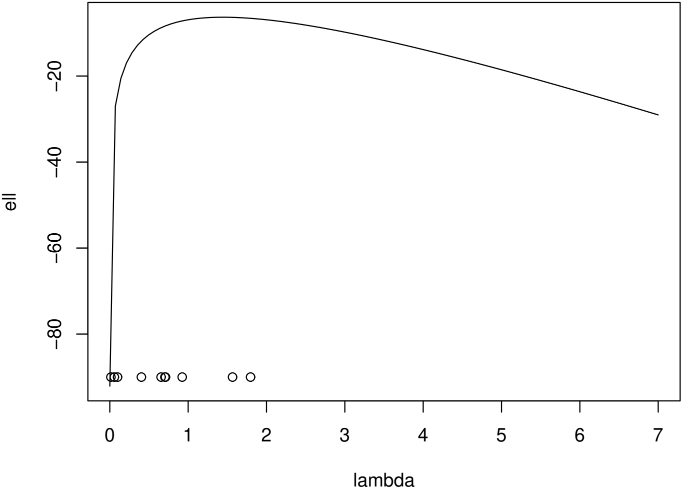
- Assume *i*, *ii*, *iii* from previously (iid, correctly specified, regularity). Then the Fisher Information Matrix can be shown to be:

$$H_n(\theta) = -E_\theta\left(\frac{d^2\ell_n(\theta)}{d\theta d\theta'}\right)$$

- Thus, under those three conditions, the Fisher Information Matrix can be expressed as three equivalent things:
  - (verbally ignoring matrix language)

  1. The variance of the slope of the log likelihood function across the probability space
  2. The negative of the EV of the second derivative of the log likelihood function
  3. The sample size times the square of the slope of the log density

Why is it called an information matrix?



Cramer-Rao inequality

- **Cramer-Rao inequality** Let $x_1, ..., x_n$ be a sample of random variables (not necessarily iid) with joint pdf $f_n(x|\theta)$. Let $W(x)$ be an unbiased estimate of θ. Then
$$Var_\theta(W) \geq \left(E_\theta\left[\left(\frac{\partial \ell_n(\theta)}{\partial \theta}\right)\left(\frac{\partial \ell_n(\theta)}{\partial \theta'}\right)\right]\right)^{-1}$$
(note: This is the inverse of the information matrix)
- Let $x_1, ..., x_n$ be an iid random sample. Then
$$Var_\theta(W) \geq \left(nE_\theta\left[\left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)\left(\frac{\partial \ln f(x|\theta)}{\partial \theta'}\right)\right]\right)^{-1}$$

Properties of MLE

- Suppose regularity conditions are met and the data is drawn iid from $f(x|\theta)$. Let $\hat{\theta}_n$ be the MLE estimate of θ. Then:
  - $\hat{\theta}_n$ is a consistent estimate of θ
  - The sequence of $\hat{\theta}$ satisfies:

    $$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\big(0, nH_n(\theta)^{-1}\big)$$

    Note that $nH_n(\theta)^{-1}$ doesn't depend on n because the n cancels with the $\frac{1}{n}$ in $H_n(\theta)^{-1}$.
- This is remarkable
  - This theorem means that the asymptotic variance of $\hat{\theta}_n$ coincides with the Cramer-Rao lower bound for the variance of unbiased estimators.
  - For this reason, the maximum likelihood estimator is said to be **asymptotically efficient**, meaning it has the lowest variance, asymptotically.
  - This also implies asymptotic normality, which is useful for hypothesis testing
  - Note that this is pretty amazing: The MLE is asymptotically normally distributed regardless of the parameter (e.g., a non-negative estimate of variance)

Bayesian estimation and inference

- Everything so far has been *frequentist* or *classical*
- Now we discuss Bayesian inference
- Four ways to think about why one would do this:
    1. What if your data gives a parameter estimate that you consider implausible. Would you believe it no matter what? Or would you reject it and do something else? If the latter, then you're doing ad hoc stuff to combine data and intuition, and you should be more systematic.
    2. What if you have some information from outside of your analysis (e.g., the results of a pilot experiment, or from someone else's estimates of the same parameter). If the point of your paper is not to estimate some parameter but to use some parameter for something, perhaps you could improve precision by combining your estimate with external information.
    3. What if the likelihood is totally intractable?
    4. Why only focus on the maximum of the likelihood function? Why not focus on the whole function?
- For any of these reasons, you might prefer Bayesian statistics

Bayesian estimation and inference

- Bayes Theorem:

$$\begin{aligned}\text{Fact 1: } P(A \cap B) &= P(B \cap A)\\ \text{Fact 2: } P(A \cap B) &= P(A|B)P(B)\\ P(A|B)P(B) &= P(B|A)P(A)\\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)}\end{aligned}$$

- In practice, letting X_n be a random sample $x_1, ..., x_n$:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X)}$$

  - What is $f(X_n|\theta)$?

Bayesian estimation and inference

- Bayes Theorem:

$$\begin{aligned} \text{Fact 1: } P(A \cap B) &= P(B \cap A) \\ \text{Fact 2: } P(A \cap B) &= P(A|B)P(B) \\ P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

- In practice, letting X_n be a random sample $x_1, ..., x_n$:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X)}$$

  - What is $f(X_n|\theta)$? The likelihood function
  - What is $f(\theta)$?

Bayesian estimation and inference

- Bayes Theorem:

$$\begin{aligned} \text{Fact 1: } P(A \cap B) &= P(B \cap A) \\ \text{Fact 2: } P(A \cap B) &= P(A|B)P(B) \\ P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

- In practice, letting X_n be a random sample $x_1, ..., x_n$:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X)}$$

  - What is $f(X_n|\theta)$? The likelihood function
  - What is $f(\theta)$? A distribution of the parameter that doesn't depend on the data (i.e., a "prior" distribution, prior to the data being observed)
  - Should our estimate of θ depend on $f(X)$?

Bayesian estimation and inference

- Bayes Theorem:

$$\begin{aligned} \text{Fact 1: } P(A \cap B) &= P(B \cap A) \\ \text{Fact 2: } P(A \cap B) &= P(A|B)P(B) \\ P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

- In practice, letting X_n be a random sample $x_1, ..., x_n$:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X)}$$

  - What is $f(X_n|\theta)$? The likelihood function
  - What is $f(\theta)$? A distribution of the parameter that doesn't depend on the data (i.e., a "prior" distribution, prior to the data being observed)
  - Should our estimate of θ depend on $f(X)$? No. $f(X)$ doesn't depend on θ.

Bayesian estimation and inference

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X)}$$

- Note that this is the “probability” (actually the density) of θ
- Integrating across all possible values of θ, it must integrate to one
- Thus, $f(X)$ is just some constant that must facilitate integrating to one:

$$\int_{-\infty}^{\infty} f(\theta|X_n)d\theta = 1 = \frac{1}{f(X)} \int_{-\infty}^{\infty} f(X_n|\theta)f(\theta)d\theta$$

- More intuitively, when comparing any two potential parameter values (i.e., calculating the probability of one relative to the other), $f(X)$ cancels out and is irrelevant:

$$\frac{f(\theta_1|X_n)}{f(\theta_2|X_n)} = \frac{f(X_n|\theta_1)f(\theta_1)}{f(X_n|\theta_2)f(\theta_2)}$$

- Thus, $f(X)$ is irrelevant and the only true difference between the **Bayesian posterior** – $f(\theta|X_n)$ – and the likelihood function is the prior $f(\theta)$

Bayesian example

- Key formulas:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X_n)} \tag{1}$$

$$\frac{f(\theta_1|X_n)}{f(\theta_2|X_n)} = \frac{f(X_n|\theta_1)f(\theta_1)}{f(X_n|\theta_2)f(\theta_2)}$$

Bayesian example

- Key formulas:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X_n)} \tag{1}$$

$$\frac{f(\theta_1|X_n)}{f(\theta_2|X_n)} = \frac{f(X_n|\theta_1)f(\theta_1)}{f(X_n|\theta_2)f(\theta_2)} \tag{2}$$

- Goals for the example: Illustrate...
  - ... how to “downweight” your prior so it doesn’t have “too much” influence on results
  - ... Bayesian statistics without a computer
  - ... “conjugate priors” for those who want to use Bayesian statistics
  - ... how, in practice, it really doesn’t matter what “the constant” is
  - ... build some foundations for Problem set 2, question 1c

Bayesian example

- Key formulas:

$$f(\theta|X_n) = \frac{f(X_n|\theta)f(\theta)}{f(X_n)} \tag{1}$$

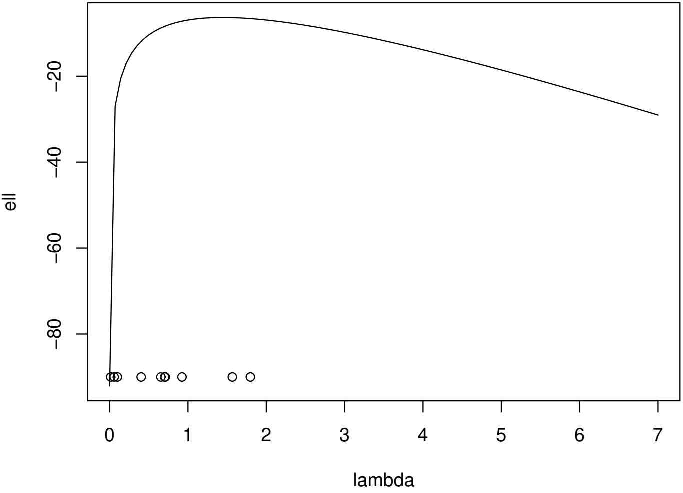
$$\frac{f(\theta_1|X_n)}{f(\theta_2|X_n)} = \frac{f(X_n|\theta_1)f(\theta_1)}{f(X_n|\theta_2)f(\theta_2)} \tag{2}$$

- Goals for the example: Illustrate...
  - ... how to “downweight” your prior so it doesn’t have “too much” influence on results
  - ... Bayesian statistics without a computer
  - ... “conjugate priors” for those who want to use Bayesian statistics
  - ... how, in practice, it really doesn’t matter what “the constant” is
  - ... build some foundations for Problem set 2, question 1c
- Example:
  - You observe an iid random sample $x_1, ..., x_n$ drawn from a normal distribution with unknown mean μ and known variance σ^2
    - In simple problems, we often assume some parameters like variance are known
    - True problem is two-parameter estimation problem, where we’d use Gibbs sampling
  - Your prior is that $\mu \sim N(\mu_0, \tau_0^2)$
  - Your goal is to estimate the posterior distribution of μ: $f(\mu_n|x_n, \sigma^2)$

Bayesian statistics

- Bayesian vs. Frequentist conversations often turn annoyingly philosophical
- What is Bayesian statistics really about?
  1. Having a prior distribution
     - Without a prior, the posterior *is* the likelihood
     - Many frequentists hate that a prior influences the results (in a non-transparent way)
  2. Full posterior distributions
     - MLE is about the mode of the distribution

A log likelihood function



Bayesian statistics

- Bayesian vs. Frequentist conversations often turn annoyingly philosophical
- What is Bayesian statistics really about?
  1. Having a prior distribution
     - Without a prior, the posterior *is* the likelihood
     - Many frequentists hate that a prior influences the results in a non-transparent way
  2. Full posterior distributions
     - MLE is about the mode of the distribution
     - Hypothesis testing is about determining whether the true mode is somewhat different than the one you estimated
     - But why would we only care about the mode?
     - Is it impossible we'd care about probabilities of parameter values other than the mode? e.g., $Pr\big(\theta \in (-\varepsilon, \varepsilon)\big)/Pr\big(\theta \in (1-\varepsilon, 1+\varepsilon)\big)$ even when $\hat{\theta}_{MLE}$ is 1/2?
- When would you want to do something Bayesian?
  1. Decision theory (e.g., it matters whether θ is close to zero or one, and simply knowing the MLE is inadequate)
  2. The likelihood is intractable: Bayesian statisticians have developed amazing simulation methods for complex or poorly behaved models. Two major classes of situations:
     - One parameter models with miserable likelihood functions: Metropolis algorithm
     - Multi-parameter models where your prior for each parameter "matches" with the likelihood function (i.e., conjugate priors): Gibbs sampling
     - Allegedly you can combine these with the Hamiltonian Monte Carlo algorithm
     - See me if you're interested in this stuff

Summary so far

- **Maximum likelihood**
    - Your data comes from some distribution
    - You analytically determine the parameter value most likely to have generated such data (among the potential values of the parameters in your distribution/model)
    - Very desirable properties: Consistent and asymptotically efficient
- **Bayesian statistics**
    - Philosophically different, in principle
    - Also focused on the likelihood function
    - Combines this likelihood with a prior
    - Various practical advantages for solving complex models
    - Useful when you care about more than just the single most likely parameter value
- What do you notice?

Summary so far

- **Maximum likelihood**
  - Your data comes from some distribution
  - You analytically determine the parameter value most likely to have generated such data (among the potential values of the parameters in your distribution/model)
  - Very desirable properties: Consistent and asymptotically efficient
- **Bayesian statistics**
  - Philosophically different, in principle
  - Also focused on the likelihood function
  - Combines this likelihood with a prior
  - Various practical advantages for solving complex models
  - Useful when you care about more than just the single most likely parameter value
- What do you notice?
- This is all assuming we know the distribution from which the data is drawn

Method of moments estimation

- Our first foray into non-parametric econometrics (more next lecture)
- It turns out lots of economics is formulated as equalities (in expectation)
  - Structural examples:
    - Price = Marginal cost
    - The marginal rate of substitution is equal to the price ratio: $(\partial U/\partial x)/(\partial U/\partial y) = p_x/p_y$
    - If an employer sees two groups equally (i.e., no discrimination), then the productivity of the marginal hire (i.e., that for which the employer is indifferent between hiring and not hiring) from each group should be equal[2]
  - Reduced form examples:
    - Treatment is uncorrelated with characteristic x: $E\left[(T_i - \bar{T})(x_i - \bar{x})\right] = 0$
    - X is exogenous: $E(X'\varepsilon) = 0$
    - We have an instrument for treatment that is uncorrelated with potential outcomes (defined next lecture)
- Are these conditions and assumptions alone enough to *identify* (formalized later) our parameters of interest? Sometimes.
  - MLE doesn't use these assumptions even though we're willing to make them
  - MLE *does* require stronger parameteric assumptions that these don't directly invoke

[2] This is the Becker Outcomes Test for discrimination. For a recent review and formal discussion, see Lee, Changhwa, Mallesh Pai, and Rakesh Vohra. "Outcome Tests for Policies." Working Paper, 2021 (https://changhwalee-econ.github.io/outcometest.pdf).

Method of moments estimation

- Suppose your model can be expressed as r equalities
- We always write these equalities as:

$$E\big(g_i(\theta, X)\big) = 0$$

Each equality:

- holds in expectation
- is a function of the k parameters or a subset of them
- is **linearly independent** of the others
  - $g_0(\theta, X)$ is linearly dependent on $g_1(\theta, X), \ldots, g_r(\theta, X)$ (i.e., not linearly independent of them) if there exists some $a_1, \ldots, a_r$ such that $a_1 g_1(\theta, X) + \ldots + a_r g_r(\theta, X) = g_0(\theta, X)$
  - Example: $\mu + \gamma = 0$ is not linearly independent of $\gamma/\beta = 1$ and $\mu = -\beta$
  - $g_0(\theta, X)$ only provides new information (beyond the other assumptions) if it is linearly independent
  - *Linear* dependence is important because we focus on equalities that hold *in expectation* and expectations are linear operators
  - Thus, $E(x + y) = c$ and $E(2x + 2y) = 2c$ are linearly dependent and not distinct assumptions, but $E[x + y] = c$ and $E\big[(x + y)^2\big] = c^2$ are linearly independent and *are* distinct assumptions

How to MOM

- Suppose you have a series of r assumptions (equalities) about k parameters:

$$E\big(g_1(\theta, X)\big) = 0$$
$$\ldots$$
$$E\big(g_r(\theta, X)\big) = 0$$

- If $r < k$, your model is under-identified (more unknowns than equations: familiar)
- If $r = k$, your model is just-identified: There is a unique solution such that each equality holds exactly
    - Method of Moments estimation
- If $r > k$, your model is over-identified: Its assumptions cannot (in general) be satisfied exactly, and how close we can come to satisfying them tells us whether or not its crazy
    - Generalized Method of Moments
    - Note: Most economists prefer over-identified models because they *let us* test whether they are crazy (a just-identified model can always be satisfied even when it's crazy)

- Note 1: We always mean linearly independent conditions
- Note 2: To an econometrician, a model is identified if, for any potential realization of data, there is only one solution in terms of parameters to satisfy the objective function. Related to what applied economists mean.

Method of moments estimation: $r = k$

- We believe $g_j(\theta, X)$ is zero in expectation
- Let's just write $g_j(\theta, X)$ and choose $\hat{\theta}$ such that, with the data we have, the sample mean is zero
- Example:
  - Linear model: $Y = X\beta + \varepsilon$
  - We assume X is exogenous: $E(X'\varepsilon)$
  - Note that this can be rewritten as: $E\big(X'(Y - X\beta)\big) = 0$
  - Choose β such that $\frac{1}{n}\sum_n x_i'(y_i - x_i\beta) = 0$
    (where x_i is the $1 \times k$ vector of regressors for observation i)
  - Note this is why OLS estimated residuals are always mechanically uncorrelated (zero correlation) with all of the regressors (more discussion of this next time)
- For most linear models, the $\hat{\theta}_{MoM}$ can be analytically derived
- Otherwise it can be solved numerically
  - If you can provide the gradient, efficient optimization algorithms
  - Otherwise, inefficient grid search always works
- We will discuss properties of MME's shortly

Generalized Method of Moments (GMM): $r > k$

- Possible that not all equalities can simultaneously be satisfied
- Two implications:
  - We have to pick some objective function to tell us whether some of these equalities are more important than others
    - Even if equally important, if the units are different then we can't necessarily directly compare deviations from zero between these different equations
  - It is informative to know how close we can come
- The GMM objective function:

$$\min_{\theta} \; J(\theta) \equiv \bar{g}_n(\theta)' W \bar{g}_n(\theta)$$

  - n is the sample size
  - $\bar{g}_n(\theta)$ is an $r \times 1$ vector where each element is given by $\frac{1}{n}\sum_{i=1}^{n} g_j(\theta, x_i)$ where i denotes observations and $j \in \{1, ..., r\}$ denotes the r restrictions
  - W is an $r \times r$ positive semi-definite matrix

A silly (but helpful?) GMM example

- Suppose you have the following three equation model:
  1. $y = \eta z + \varepsilon_y, \quad \varepsilon_y \sim N(0, \sigma_y^2)$
  2. $X = \beta z + \varepsilon_x, \quad \varepsilon_x \sim N(0, \sigma_x^2)$
  3. $Cov(\varepsilon_y, \varepsilon_x) = E(\varepsilon_y \varepsilon_x) = \sigma_{xy}$
  - Note: The third assumption has real content and might improve estimates of η, β
- Let's make the silly assumption $\sigma_x, \sigma_y, \sigma_{xy}$ are known
- Two unknowns (k) and three equations (r):

$$g(\theta) = \begin{pmatrix} y - \eta z \\ x - \beta z \\ (y - \eta z)(x - \beta z) - \sigma_{xy} \end{pmatrix}$$

- GMM minimizes: $\bar{g}_n(\theta)' W \bar{g}_n(\theta)$
- Suppose we take $W = I_{3\times 3}$
- Then GMM minimizes:

$$(\bar{y} - \eta \bar{z})^2 + (\bar{x} - \beta \bar{z})^2 + \left[\frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \eta z_i)(x_i - \beta z_i)\right] - \sigma_{xy}\right]^2$$

- Intuitive: Gives each assumption equal weight

A silly (but helpful?) GMM example

- Suppose we take $W = I_{3\times 3}$
- Then GMM minimizes:

$$(\bar{y} - \eta\bar{z})^2 + (\bar{x} - \beta\bar{z})^2 + \left[\frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \eta z_i)(x_i - \beta z_i)\right] - \sigma_{xy}\right]^2$$

- But suppose that $\sigma_x^2 > \sigma_y^2$? Doesn't it make sense that we might worry less about a 1 unit deviation of $\bar{x}$ from $\beta\bar{z}$ than a one unit deviation of $\bar{y}$ from $\eta\bar{z}$?
- There's an intuitive sense that we want to normalize by the variance
- So maybe choose a W that accounts for this:

$$W = \begin{pmatrix} 1/\sigma_y^2 & 0 & 0 \\ 0 & 1/\sigma_x^2 & 0 \\ 0 & 0 & 1/\sigma_{xy} \end{pmatrix} = \begin{pmatrix} Var\big(g_1(\theta, X)\big) & 0 & 0 \\ 0 & Var\big(g_2(\theta, X)\big) & 0 \\ 0 & 0 & Var\big(g_3(\theta, X)\big) \end{pmatrix}^{-1}$$

- Then GMM will minimize:

$$\frac{(\bar{y} - \eta\bar{z})^2}{\sigma_y^2} + \frac{(\bar{x} - \beta\bar{z})^2}{\sigma_x^2} + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \eta z_i)(x_i - \beta z_i)}{\sigma_{xy}} - 1\right]^2$$

A silly (but helpful?) GMM example

- If we choose

$$W = \begin{pmatrix} 1/\sigma_y^2 & 0 & 0 \\ 0 & 1/\sigma_x^2 & 0 \\ 0 & 0 & 1/\sigma_{xy} \end{pmatrix} = \begin{pmatrix} Var\big(g_1(\theta, X)\big) & 0 & 0 \\ 0 & Var\big(g_2(\theta, X)\big) & 0 \\ 0 & 0 & Var\big(g_3(\theta, X)\big) \end{pmatrix}^{-1}$$

  then GMM will minimize:

$$\frac{(\bar{y} - \eta\bar{z})^2}{\sigma_y^2} + \frac{(\bar{x} - \beta\bar{z})^2}{\sigma_x^2} + \left[\frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \eta z_i)(x_i - \beta z_i)}{\sigma_{xy}} - 1\right]^2$$

- But what if we also knew that $\tilde{x} = \beta z + \varepsilon_{\tilde{x}}$ where $E(\varepsilon_x \varepsilon_{\tilde{x}})$ was very high
- There's a sense in which we want to use an additional assumption because more assumptions help us to identify the parameters
- But there's a sense in which, if $E(\varepsilon_x \varepsilon_{\tilde{x}})$ is very high, then this isn't really providing any new information, just giving double the weight to our existing assumption that $x = \beta z + \varepsilon_x$
- What if we could include it, but downweight it based on its similarity to the assumptions that we already have
- One option: Don't only downweight by the variance, but add covariance terms to the off-diagonal elements of W^{-1}

Properties of W

- For any positive semi-definite matrix W (including the identity matrix), $\hat{\theta}_{GMM}$ is a consistent estimator for θ
- Assuming a central limit theorem can be applied (allows for some dependence but not too much), then $\hat{\theta}_{GMM}$ is asymptotically normal
- The asymptotic variance is minimized by choosing $W = \Omega^{-1} \equiv E\left[g(\theta, X)g(\theta, X)'\right]^{-1}$: The inverse of the variance-covariance matrix of the vector of moment conditions, just like we intuited in our silly example!

Properties of W

- For any positive semi-definite matrix W (including the identity matrix), $\hat{\theta}_{GMM}$ is a consistent estimator for θ
- Assuming a central limit theorem can be applied (allows for some dependence but not too much), then $\hat{\theta}_{GMM}$ is asymptotically normal
- The asymptotic variance is minimized by choosing $W = \Omega^{-1} \equiv E\left[g(\theta, X)g(\theta, X)'\right]^{-1}$: The inverse of the variance-covariance matrix of the vector of moment conditions, just like we intuited in our silly example!
- Let D be the **gradient**, a $r \times k$ matrix given by:

$$D = \frac{\partial g(\theta, Y)}{\partial \theta} = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_1(\theta)}{\partial \theta_k} \\ \dots & \dots & \dots \\ \frac{\partial g_r(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_r(\theta)}{\partial \theta_k} \end{pmatrix}$$

- Let $\hat{\theta}_{GMM}\left(\Omega^{-1}\right)$ be the GMM estimator one obtains from choosing Ω^{-1} to be the weights matrix, and assume assumptions are met such that D is consistent and a CLT applies (technical assumptions we'll ignore). Then $\hat{\theta}_{GMM}\left(\Omega^{-1}\right)$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\theta}_{GMM}\left(\Omega^{-1}\right) - \theta\right) \xrightarrow{d} N\left(0, (D'\Omega^{-1}D)^{-1}\right)$$

A gradient of learning

- Who is this gradient fellow?

$$D = \frac{\partial g(\theta, Y)}{\partial \theta} = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_1(\theta)}{\partial \theta_k} \\ \dots & \dots & \dots \\ \frac{\partial g_r(\theta)}{\partial \theta_1} & \dots & \frac{\partial g_r(\theta)}{\partial \theta_k} \end{pmatrix}$$

- Asymptotic normality:

$$\sqrt{n}\big(\hat{\theta}_{GMM}(\Omega^{-1}) - \theta\big) \xrightarrow{d} N\big(0, (D'\Omega^{-1}D)^{-1}\big)$$

- Imagine we ignore rank conditions/over-identification and pretend that there's only one moment condition: $r = 1$ so that Ω^{-1} is a scalar
- Then

$$E(D'\Omega^{-1}D) = \Omega^{-1}E(D'D) = \Omega^{-1}E\left[\frac{\partial g(\theta, Y)}{\partial \theta}\frac{\partial g(\theta, Y)}{\partial \theta'}\right]$$

which looks very much like our old friend the Fisher Information Matrix! (The "squared" (except in matrix form) derivative of the likelihood function with respect to the parameters)

A gradient of learning

$$\sqrt{n}\big(\hat{\theta}_{GMM}(\Omega^{-1}) - \theta\big) \xrightarrow{d} N\big(0, (D'\Omega^{-1}D)^{-1}\big)$$

- The gradient $D = \frac{\partial g(\theta)}{\partial \theta}$ serves the same purpose as the Fisher Information matrix
- It tells us how responsive our identifying information (our constraints: the likelihood function in MLE, the moment conditions in GMM) is to changes in the underlying parameters
- If it's very responsive – meaning that a modest change in parameter values generates very different predictions for the data – then our model will stands to learn a lot from the data, and $H(\theta)$ and $D'D$ are going to be large
- This implies that the asymptotic variance of the estimates is going to be relatively small: $H(\theta)^{-1}$ and $(D'\Omega^{-1}D)^{-1}$ will be small
- We will get a fairly reliable estimate of θ because our model is informative!
- **Key insight**: A model is a series of predictions. The more your predictions change in response to the data, the more we stand to learn from collecting and analyzing data.
    - This corresponds to intuitive notions of epistemology

MLE vs. GMM

- In general, MLE is more asymptotically efficient than GMM
- GMM will reach the Cramer-Rao lower bound when you have an infinite number of moment conditions. Why?

MLE vs. GMM

- In general, MLE is more asymptotically efficient than GMM
- GMM will reach the Cramer-Rao lower bound when you have an infinite number of moment conditions. Why?
- MLE *does* have infinite moment conditions, because the full distribution of data is being used, since the likelihood has predictions for the full distribution of data (i.e., $F(x)$ is defined $\forall x$)
- We already said MLE is more parametric than GMM: It assumes specific distributional families. This is what this means.
- Assuming a family gives an infinite number of parameter-dependent predictions for what you'll see in the data, and that means the data is extremely informative about your parameters
- Because GMM imposes fewer restrictions, it stands to learn less from seeing the data, because it uses that data in a more limited way (e.g., only uses the sample mean of the moment conditions)
- The most fundamental tradeoff of econometrics: Stronger assumptions are more restrictive, but deliver more powerful and precisely estimated estimators

Two outstanding issues for GMM ($r > k$)

- Choose $\hat{\theta}_{GMM}$ to minimize:

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(\theta, y_i)'\Omega^{-1} g(\theta, y_i)$$

  where $\Omega = E\Big[g(\theta, X) g(\theta, X)'\Big]$
- Problem: In general, Ω depends on θ
  - Two-step feasible GMM:
  - Recall that $\hat{\theta}_{GMM}$ is consistent whenever W is p.s.d.
  - Use $W = I$ to get a consistent estimate of θ called $\hat{\theta}_0 = \hat{\theta}_{GMM}(I)$
  - This implies that $\hat{\Omega}(\hat{\theta}_0)$ is a consistent estimate of Ω
  - Use this $\hat{\Omega}(\hat{\theta}_0)$ to calculate $\hat{\theta}_{GMM}$, which is a consistent estimate of θ that is asymptotically more efficient than $\hat{\theta}_0$
  - Could also iterate back and forth between estimates of $\hat{\theta}_{GMM}$ and $\hat{\Omega}$ until $\hat{\theta}_{GMM}$ converges (iterated GMM), which probably makes more sense when your sample size is small

Two outstanding issues for GMM ($r > k$)

- Choose $\hat{\theta}_{GMM}$ to minimize:

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(\theta, y_i)'\Omega^{-1}g(\theta, y_i)$$

  where $\Omega = E\big[g(\theta, X)g(\theta, X)'\big]$
- Problem: In general, Ω depends on θ
  - Two-step feasible GMM
- Second, if not all equalities can simultaneously be satisfied, it is informative to know how close we can come
  - It turns out that if all the moment conditions are correct in expectation (i.e., our model is correctly specified) then:

$$\sqrt{n}\big(\bar{g}(\theta, Y)'\Omega^{-1}\bar{g}(\theta, Y)\big) \xrightarrow{d} \chi^2(r-k)$$

  - The chi-squared distribution is a known distribution, so we can calculate p-values
    - This works as long as we use a consistent estimate of Ω, such as $\hat{\Omega}(\hat{\theta}_{GMM})$

Summary so far

- Three families of estimators
- **MLE**: Assume a likelihood and derive an assumption-consistent estimator
- **Bayesian**: Add a prior distribution and consider the possibility the true parameter takes values other than that at which the likelihood function is maximized
- **GMM**: Don't assume functional forms, just derive semi-parametric estimators that are consistent with economic theory instead of statistical theory

- Compared to GMM, MLE estimators are lower variance (assuming they're correctly specified) because they use more of the data
    - Key question of all of econometrics: What variation from the data am I using and how much does my estimator (my understanding of the world) respond to the data?
- But it makes stronger assumptions than GMM (GMM will converge to MLE if you make infinite moment condition assumptions)
- But what happens if the MLE model is misspecified and based on wrong assumptions? See problem set 2.