Econometrics Math Camp

Day 3: Frequentist and Bayesian Inference

Michael Droste

August 2021

Introduction

- Welcome back!

- Materials (slides, notes, and problems) are available on GitHub: https://github.com/mdroste/metrics-mathcamp-2021/

- It's our last day! Today, the focus is on estimators and statistical inference.

Today's Outline

- Frequentist Inference
  - Statistical Models and Estimators
  - Estimators: Method of Moments and Maximum Likelihood
  - Confidence Sets
  - Hypothesis Testing
- Bayesian Inference
  - Bayes' rule
  - Priors and Posteriors

Statistical Models

- A statistical model is a set of probability distributions indexed by a parameter vector θ.
- A statistical model is said to be parametric if the dimension of the parameter space is finite. Otherwise, the statistical model is said to be nonparametric.
- The idea of statistical inference is: suppose we observe data drawn from the probability distributions. What can we learn about θ?

Frequentist Inference

- Frequentist (or classical) inference starts from the idea that the parameters θ are fixed (but unknown).

- We observe data, say $X_1, ..., X_N$, drawn from $F_\theta(x)$.

- We want to learn something about the true population θ from the data.

- Most often, frequentist econometrics involves one or more of the following:
    1. Point estimation: A best guess of θ
    2. Confidence sets: A set of values that probably covers θ
    3. Hypothesis testing: The probability that a statement involving θ is false

Estimators

- An estimator is a function that maps our data to a “guess” of the parameter vector θ. We will often denote an estimator as $\hat{\theta}$.

- Our estimator is a function of the data. We can be more explicit about this, if we wish, by writing an estimator as $\hat{\theta}(x_1, ..., x_N)$.

- Note that an estimator is a function. An estimate is the result of applying an estimator to a particular dataset.

Properties of Estimators

- In many particular use cases, more than one estimator is available to the econometrician. How do you evaluate the “performance” of an estimator?

- Most handy properties of estimators:
  1. Consistency: $\hat{\theta}(x_1, ..., x_n) \xrightarrow{p} \theta$ as $n \to \infty$
  2. Unbiasedness: $E[\hat{\theta} - \theta] = 0$
  3. Efficiency: $Var(\hat{\theta}(x_1, ..., x_n))$ is minimized[1]

- $\sqrt{n}$-consistency and asymptotic normality: $\exists$ constant V such that $\sqrt{n}(\hat{\theta} - \theta) \to N(0, V)$.

- An important class of estimators satisfy $\sqrt{n}$-consistency and asymptotic normality, which provides a unifying lens for the second half of the econometrics sequence (2140).

[1] Defined loosely. Will be covered more carefully in 2120, in particular the notion of a Cramer-Rao lower bound and minimum variance unbiased estimator.

Classical Method of Moments

- One (historically) common frequentist estimator is the classical method of moments. Introduced by Chebyshev as a part of his proof of the central limit theorem.

- Idea in a nutshell: write expressions relating sample moments to population moments, where population moments are functions of θ. Solve for $\hat{\theta}$.

- More precisely:
  - Observe data $x = (x_1, \ldots, x_n)$ with density $f(x; \theta)$, where $\dim(\theta) = k$.
  - Write first k population moments of x in terms of θ.
  - Write down equations relating sample moments to population moments (in terms of θ).
  - Solve system to obtain estimates, $\hat{\theta}$.

Example: Classical Method of Moments

- Suppose we observe n iid draws from a Poisson distribution with unknown parameter λ. Denote the data $x = (x_1, ..., x_n)$. We wish to use the method of moments to estimate λ.
- The Poisson probability mass function is $f(k; \lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Writing the first pop moment in terms of λ: $E[X] = \lambda$.
- Our estimate is simply the sample mean, $\hat{\lambda} = 1/n \sum x_i$.
- Is this estimator consistent? Why?

Maximum Likelihood Estimation

- Another extremely important class of estimators: maximum likelihood estimation.

- Idea in a nutshell: choose as estimates the parameter values that "maximize the likelihood" of realizing the data. Rather than matching on moments, we match the distribution.

- Define the likelihood as the joint density or mass function evaluated at the realized data, $f(x_1, \ldots, x_n; \theta)$, and write it $L(\theta)$.

- The maximum likelihood estimate $\hat{\theta}^*_{\text{MLE}}$ is the value of θ that maximizes the likelihood, evaluated at the realized data: $\hat{\theta}^*_{\text{MLE}} = \text{argmax}_\theta\, L(\theta)$.

Maximum Likelihood Estimation

- There are a few tricks to make solving maximum likelihood estimators easier.
- With iid data, we can express the likelihood as the product of the likelihood of observing each observation:

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- If the likelihood is strictly positive, we can take logs. The log of the likelihood function is called the log likelihood and usually denoted $\ell(\theta)$. In the i.i.d. case it simplifies as follows:

$$\ell(\theta) = \log L(\theta) = \log \left(\prod_{i=1}^{n} f(x_i; \theta) \right) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

- We can solve this by taking first-order conditions with respect to θ and/or using numerical approximation.

Example: Maximum Likelihood Estimation

- Let's reconsider the Poisson example we used for the method of moments.

- The likelihood function for each observation x_i is the probability mass fn evaluated at x_i,

$$f(x_i; \lambda) = P(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

- Because of i.i.d. data, the log likelihood is

$$\ell(\lambda) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) = (\log \lambda)\left(\sum_{i=1}^{n} x_i\right) - \lambda n + \text{constant}$$

where the 'constant' depends on data but not λ, and so is irrelevant for optimization.

Example: Maximum Likelihood Estimation

- Now solve by taking the first order condition:

$$\frac{\partial \ell}{\partial \lambda} = \frac{1}{\lambda}\left(\sum_{i=1}^{n} x_i\right) - n = 0$$

- The solution to this equation is $\hat{\lambda}^*_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, again the sample mean.

Confidence Sets

- Let the data be denoted $X = (X_1, ..., X_n)$. A $1 - \alpha$ confidence set, $\mathcal{C}(\cdot) \subset \Theta$ is a set-valued function of data with the property that $P(\theta \in \mathcal{C}(X)) = 1 - \alpha$.

- In one dimension, this is usually a confidence interval.

- The interpretation of this object is subtle. Each independent experiment delivers a different dataset, and each dataset generates a confidence set. When you pick up *Science* and read about the independent identical experiments that were conducted, expect the proportion of experiments whose confidence set includes θ to be $1 - \alpha$.

- Even taking a model as given, there are many ways to construct a confidence set. We prefer small confidence sets (by Lebesgue measure, for example) and confidence sets centered at the point estimate.

Hypothesis Tests

- Hypothesis tests allow us to determine the probability that a statement involving θ is false.
- The machinery of hypothesis testing is built on a null hypothesis $H_0 : \theta \in \Theta_0$ and an alternative hypothesis $H_A : \theta \in \Theta_A$, where Θ_0 and Θ_A are disjoint subsets of Θ.
- Hypothesis tests are defined by a decision rule d mapping the realized data into $\{0, 1\}$, where 0 means "accept H_0", and 1 means "reject H_0 in favor of H_A".
- Many common hypothesis tests define a test statistic $T(x)$ to map the realized data x into $\mathbb{R}$, and a fixed rejection region R, and use the following decision rule:

$$d(x) = \begin{cases} 0, & T(x) \notin R \\ 1, & T(x) \in R \end{cases}$$

Hypothesis Tests

- We evaluate hypothesis tests by the two ways that they can go awry.

- Type I error (false rejection): if we reject the null hypothesis when it is in fact true ($\theta \in \Theta_0$). The probability of making a Type 1 error is known as significance, usually denoted α.

- Type II error (false acceptance): if we accept the null hypothesis when the alternative is in fact true ($\theta \in \Theta_A$). The probability of avoiding a Type II error is known as power and is usually denoted $1 - \beta$.

- Frequentists usually begin with a significance level, i.e. $\alpha = 0.05$, and then choose a rejection region $R(\alpha)$ such that the test significance is α. Given the data, the p-value is the smallest significance level at which the null would be rejected:

$$p(x) = inf\{\alpha : T(x) \in R(\alpha)\}$$

Duality of Confidence Sets and Hypothesis Tests

- We can start with a hypothesis test and construct a confidence set by asking what simple null hypothesis would be accepted. More formally:

- **Theorem.** Suppose that for each $\theta_0 \in \Theta$, there exists a test of the simple null hypothesis $H_0 : \theta = \theta_0$ with significance α. Denote its decision rule by d_{θ_0}. Then $\mathcal{C}(x) = \{\theta_0 : d_{\theta_0}(x) = 0\}$ is a $1 - \alpha$ confidence set for θ.

- We can also go the other way, by starting with a confidence set and then constructing a hypothesis test.

- **Theorem.** Let $\mathcal{C}(x)$ be a $1 - \alpha$ confidence set for θ. Then the following test is a hypothesis test of the null hypothesis $H_0 : \theta = \theta_0$ with significance α: accept H_0 if $\theta_0 \in \mathcal{C}(x)$, reject H_0 otherwise.

Bayesian Inference

- Recall that in frequentist inference, the parameter θ is unknown but constant.
- The key distinguishing feature of Bayesian inference is that the population parameter θ is treated as a random variable with its own distribution.
- This difference leads us to a different set of procedures for conducting statistical inference.

Bayesian Priors

- The parameter θ is treated as a random variable with its own distribution, $\pi(\theta)$.

- Think of the prior distribution as encoding *prior information* about the parameter θ available to the researcher prior to observing the data.

- This may come from prior experiments, observational studies, or economic theory.

Bayes Rule: Priors to Posteriors

- Bayes rule provides a logically consistent way to combine prior information with observed data:

  $$\pi(\theta|x) = \frac{f_\theta(x)\pi(\theta)}{f(x)}$$

  where $f(x) = \int_\Theta f_\theta(x)\pi(\theta)d\theta$ is the marginal density of X, $f_\theta(x)$ is the likelihood function.

- We call $\pi(\theta|x)$ the posterior density of θ; it is the object of Bayesian inference.

- The Bayesian uses $\pi(\theta|x)$ to conduct all statistical inference concerning θ.

-

Conjugate Priors

- Bayesian inference is often computationally expensive.

- The posterior distribution is potentially ugly and difficult to sample from (depending on the dimension of the parameter space). High-powered numerical methods, like Markov Chain Monte Carlo (MCMC) algorithms, may be required to evaluate a broad class of problems (more on this in Ec2140).

- Alternatively, one simplifying trick for Bayesian inference is to make use of conjugate priors. A prior is said to be conjugate for a given likelihood function if the posterior function induced by the prior is in the same class of distributions as the prior.

- In practice, choosing conjugate priors can make Bayesian problems much simpler to solve. Analytical solutions also tend to reveal exactly how the prior influences the posterior.

Posterior Objects

- So you've computed a posterior distribution - now what?

- Most often, Bayesian point estimators are simple measures of the posterior's central tendency - the mean, median, or mode. These are simple to compute as long as the parameter space isn't large and/or you have an analytical solution for the posterior (i.e. conjugate prior).

- Likewise, the Bayesian analog of frequentist confidence sets, credible sets, are really just ways to summarize dispesion in the posterior. This is a bit more complicated.

Credible Intervals

- Credible sets (or regions) are the rough Bayesian analog of frequentist confidence sets; they capture dispersion in the posterior distribution rather than the sampling distribution.

- Unlike confidence sets, Bayesian credible sets depend on a prior.

- The most common representation is the highest posterior density interval (HPDI); a 95% credible region using the HPDI is the narrowest region of the parameter space containing 95% of the mass of the posterior.

- Alternative characterizations exist: credible intervals can be constructed to be “symmetric” about the posterior mean, or chosen such that the probability of being below the interval is equal to the probability of being above it.

Wrapping Up

- Good luck in your first year! I hope you enjoyed this portion of math camp, and I hope you will find the material we covered useful this year.

- Feel free to reach out if you have any questions or want to chat about anything - I would love to chat with you at any time.