Introduction to Statistics
(Lectures 11-12: Confidence Sets)

1 Confidence Sets

Overview: Lectures 11-12 will focus on the construction of *confidence sets*. The actions are subsets of the parameter space. Each action is interpreted as a region that (presumably) contains the parameter selected by nature. The loss function we work with is a convex combination of "volume" and "coverage", in a sense we will make precise.

We first show that Bayes rules for this problem are "highest posterior density" sets: the collection of parameters with a sufficiently high posterior probability. We then argue that in large samples, highest posterior density sets can be approximated by "inverting" a Wald test. When the parameter of the model is scalar, inverting the Wald test is tantamount to reporting the interval formed by the Maximum Likelihood estimator $\pm$ c times its estimated "standard error".

We then introduce and analyze two popular algorithms for summarizing the uncertainty in parameter estimation: (parametric) bootstrap confidence intervals and Bayesian "credible" intervals. We use these algorithms to construct confidence/credible regions for one of the possibly many coefficients of the Linear Regression model.

1.1 Description of the problem

Let X be a random variable and let $\{f(x|\theta)\}_{\theta\in\Theta}$ be a statistical model. The problem of constructing a confidence set for θ starts as follows. In the first stage nature picks an element of the parameter space, which in this section we model explicitly as a subset of $\mathbb{R}^k$. The econometrician cannot observe the parameter selected by nature but observes data. Based on the data, the econometrician has to suggest a "region" of the parameter space that is likely to contain the true parameter selected by nature.

In principle, the action space for the econometrician based on the description above is the power set of Θ. A minor technical restriction is that the econometrician can only report regions that are "measurable". This happens because the "volume" of whatever set is reported will be part of the loss function. Let $\mathcal{C}$ denote the class of regions that the econometrician can report and let c denote an element of $\mathcal{C}$.

A *decision rule/algorithm/strategy* for this problem is a mapping

$$C : X \to \mathcal{C}$$

These decision rules will be called "confidence" regions.

1.2 Loss function

The payoff/loss for the econometrician depends on the action taken ($c \in \mathcal{C}$) and the true parameter (θ). The following loss is sometimes used to evaluate confidence sets:

$$\mathcal{L}_\delta(c,\theta) = \delta \mathbf{1}\{\theta \notin c\} + (1-\delta)\int_c d\theta,$$

where $\delta \in (0,1)$. The first component in the loss function above penalizes the econometrician whenever θ does not belong to c. The econometrician could avoid the penalty by reporting a "large" region (for example, reporting Θ). The second component makes large regions unattractive by taking into account their volume.

The expected loss (risk) of a confidence region C is given by

$$\mathbb{E}_{f(x|\theta)}[\mathcal{L}_\delta(C(x),\theta)] = \delta(1 - P_{f(x|\theta)}(\theta \in C(x))) + (1-\delta)\mathbb{E}_{f(x|\theta)}[\text{Vol}(C(x))]. \tag{1.1}$$

The term

$$P_{f(x|\theta)}(\theta \in C(x))$$

is typically referred to as the *coverage probability* of $C(\cdot)$ at θ. This captures how often the confidence region contains the parameter θ when the data is indeed generated by it. The worst coverage probability over the parameter space:

$$\inf_{\theta\in\Theta} P_{f(x|\theta)}(\theta \in C(x))$$

is called the *confidence level* of the confidence set $C(x)$.

1.3 Minimizing Posterior-Loss

We now present a rule that minimizes posterior-loss. Any such rule, by construction, cannot be uniformly improved: there is no other region that has better coverage everywhere and also smaller expected volume. Let $\pi(\cdot)$ denote a prior over Θ and let $\pi(\theta|x)$ denote the posterior.

The posterior loss of an action c is given by:

$$\delta(1 - P_{\pi(\theta|x)}(\theta \in c)) + (1-\delta)\int_c d\theta.$$

The term

$$P_{\pi(\theta|x)}(\theta \in c) = \int_c \pi(\theta|x) d\theta$$

is the posterior probability that θ falls in a set c. Such probability is usually referred to as the *credibility* of c. Algebra shows that the posterior loss can be written as

$$\delta - \int_c \left(\delta \pi(\theta|x) - (1-\delta)\right) d\theta.$$

Thus, for each x, the posterior loss is minimized by choosing

$$C_\delta^{\text{Bayes}}(x) = \{\theta \in \Theta \mid \pi(\theta|x) \geq (1-\delta)/\delta\}. \tag{1.2}$$

The region above is called a "highest posterior density" region. As it is name suggests, the region collects the points in the parameter space for which the posterior density is above certain threshold.

In general, it is hard to give closed form expressions for the highest posterior density region. However, in large samples it is possible to provide (heuristic) approximations for these regions; exploiting the fact that—regardless of the prior—the posterior distribution is approximately distributed as

$$\mathcal{N}_k\left(\widehat{\theta}_{\text{ML}} \quad , \quad \widehat{I}^{-1}\right),$$

where $\widehat{I}$ is the observed information. This means that highest posterior density regions will be approximately of the form

$$\left\{\theta \in \Theta \,\middle|\, \frac{1}{(2\text{det}\widehat{I}))^{k/2}} \exp\left(-\frac{1}{2}(\theta - \widehat{\theta}_{\text{ML}})'\widehat{I}(\theta - \widehat{\theta}_{\text{ML}})\right) \geq (1-\delta)/\delta\right\},$$

which can be written as

$$\left\{\theta \in \Theta \,\middle|\, (\widehat{\theta}_{\text{ML}} - \theta)'\widehat{I}(\widehat{\theta}_{\text{ML}} - \theta) \leq \widehat{c}_\delta\right\},$$

where $\widehat{c}_\delta$ is some function that depends on δ and $\widehat{I}$ but not on θ. Recognize the quadratic form as the Wald statistic for the problem

$$\mathbf{H}_0 : \theta = \theta_0, \text{ vs. } \mathbf{H}_1 : \theta \neq \theta_0.$$

Thus, in large samples, highest-posterior density regions are approximately given by those points in the parameter space in which the Wald statistic is below certain threshold that might depend on the data.

When θ is scalar, the approximation takes the form

$$\Big[\widehat{\theta}_{\text{ML}} - \sqrt{\frac{\widehat{c}_\delta}{\widehat{I}}}\ ,\ \widehat{\theta}_{\text{ML}} + \sqrt{\frac{\widehat{c}_\delta}{\widehat{I}}}\Big]$$

Example: We derive the highest posterior density region for a linear regression model with known variance. We have shown that if the prior density is that of a $\mathcal{N}_k(0, \sigma^2(\mathbb{I}_n/\lambda))$, the posterior of β is given by

$$\beta|Y, X \sim \mathcal{N}_k(m, V),$$

where

$$\begin{aligned} m &\equiv (X'X + \lambda\mathbb{I}_n)^{-1}X'Y, \\ V &\equiv \sigma^2(X'X + \lambda\mathbb{I}_n)^{-1}. \end{aligned}$$

The highest-posterior density region is thus

$$\begin{aligned} C_\delta(x) &= \left\{\beta \in \mathbb{R}^k \;\Big|\; \frac{1}{(2\text{det}V))^{k/2}} \exp\left(-\frac{1}{2}(\beta - m)'V^{-1}(\beta - m)\right) \geq (1-\delta)/\delta\right\}, \\ &= \left\{\beta \in \mathbb{R}^k \;\Big|\; (\beta - m)'V^{-1}(\beta - m) \leq \widehat{c}_{\delta, X, \sigma^2}\right\}. \end{aligned}$$

This is an ellipse centered at the posterior mean. In large samples,

$$m \approx \widehat{\beta}_{\text{OLS}}, \quad V \approx \sigma^2(X'X)^{-1}.$$

Hence, the highest-posterior density region is approximately given by

$$\left\{\beta \in \mathbb{R}^k \;\Big|\; \frac{1}{\sigma^2}(\widehat{\beta} - \beta)'(X'X)(\widehat{\beta} - \beta) \leq \widehat{c}_{\delta, X, \sigma^2}\right\}.$$

This is set of all parameter values for which the Wald test for the null $\beta = \beta_0$ is low enough.

1.4 Confidence Regions and Hypothesis Testing

There is a connection between confidence regions and tests of hypothesis that goes beyond the approximation argument above. Any family of tests of size α can be "inverted" to construct a confidence region with confidence level of at least $1 - \alpha$. Likewise, any confidence region with confidence level $1 - \alpha$ can be transformed into a statistical test with rate of Type 1 error of at most α.

Inverting a family of tests: Let $\{\phi_\theta\}_{\theta\in\Theta}$ denote a collection of non-randomized tests. Suppose each element in the collection is a test for the problem

$$\mathbf{H}_0 : \theta = \theta_0, \quad \mathbf{H}_1 : \theta \neq \theta_0.$$

Suppose further that each test has size of at most α in the sense that

$$\mathbb{P}_{f(x|\theta)}[\phi_\theta(x) = 1] \leq \alpha.$$

The confidence region obtained by "test inversion" is given by

$$C(x) \equiv \{\theta \in \Theta \mid \phi_\theta(x) = 1\}.$$

The coverage of such confidence region at θ is given by

$$\begin{aligned} P_{f(x|\theta)}[\theta \in C(x)] &= 1 - P_{f(x|\theta)}[\theta \notin C(x)] \\ &= 1 - P_{f(x|\theta)}[\phi_\theta(x)] \\ &\geq 1 - \alpha. \end{aligned}$$

Therefore,

$$\inf_{\theta\in\Theta} P_{f(x|\theta)}[\theta \in C(x)] \geq 1 - \alpha,$$

which implies that the confidence region formed by test inversion has confidence level, provided each of the tests ϕ_θ have size of at α.

Using a confidence region to test a hypothesis: A confidence region with $C(x)$ confidence level of at least $1 - \alpha$ can be readily used to test a hypothesis of the form

$$\mathbf{H}_0 : \theta = \theta_0 \quad \text{vs.} \quad \mathbf{H}_1 : \theta \neq \theta_0.$$

Consider the test ϕ that rejects the null hypothesis above whenever $\theta \notin C(x)$. The rate of Type I error of the test is

$$\begin{aligned} P_{f(x|\theta)}[\phi(x) = 1] &= P_{f(x|\theta)}[\theta \notin C(x)] \\ &= 1 - P_{f(x|\theta)}[\theta \in C(x)] \\ &\leq 1 - \inf_{\theta\in\Theta} P_{f(x|\theta)}[\theta \in C(x)] \\ &\leq 1 - (1 - \alpha) = \alpha. \end{aligned}$$

1.5 Parametric Bootstrap and Bayesian Credible sets

1.5.1 The Parametric Bootstrap

Suppose that we have a statistical model $\{f(x|\theta)\}_{\theta\in\Theta}$ for data X. The estimator $\widehat{\theta}_{\mathrm{ML}}$ is a random variable, as it depends on the data. If θ were known, we could approximate the distribution of $\widehat{\theta}_{\mathrm{ML}}$ by Monte-Carlo methods: generate I independent draws from $x \sim f(x;\theta)$, and evaluate $\widehat{\theta}_{\mathrm{ML}}$ over the I new data sets that we have generated.

The true parameter that generated the data is unknown; hence it is not possible to generate draws from the model $f(x;\theta)$. However, since there is an estimator for θ, one could perform the Monte-Carlo approximation exercise described above by using draws from the model:

$$f(x|\widehat{\theta}_{\mathrm{ML}}).$$

If we do this, we will effectively end with I estimators (one for each new data set). Suppose that the sample size is large enough to guarantee that $\widehat{\theta}_{\mathrm{ML}}$ is close to the true θ (whatever the true value might be) with high probability. If the parametric model $f(x|\theta)$ varies smoothly with θ, the c.d.f. based on the I estimators can be shown to be a reasonable approximate for the distribution of $\widehat{\theta}_{\mathrm{ML}}$. If the parameter of interest is some function $g(\theta)$, the procedure above also provides an approximation for $g(\widehat{\theta}_{\mathrm{ML}})$, provided g is smooth.

The exercise described above is known as the parametric bootstrap and it is a general statistical technique used to estimate the distribution of $\widehat{\theta}_{\mathrm{ML}}$ or any function smooth function $g(\widehat{\theta}_{\mathrm{ML}})$.

Example: Consider the linear regression model with unknown variance

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 \mathbb{I}_n).$$

The parameters of the model are $\beta = (\beta_1, \ldots, \beta_k)'$ and σ^2. Suppose we want to construct a confidence region for the β_1. The ML estimators for β and σ^2 are

$$\widehat{\beta}_{\mathrm{ML}} = (X'X)^{-1}X'Y, \quad \widehat{\sigma}^2_{\mathrm{ML}} = (Y - X\widehat{\beta}_{\mathrm{ML}})'(Y - X\widehat{\beta}_{\mathrm{ML}})/n.$$

The parametric bootstrap algorithm starts by generating I new data sets, independently according to the distribution

$$Y(i) \sim \mathcal{N}_n(X\widehat{\beta}_{\mathrm{ML}}\ , \ \widehat{\sigma}^2_{\mathrm{ML}}\mathbb{I}_n).$$

For each of these data sets, we can compute

$$\widehat{\beta}_{\text{ML}}(i) = (X'X)^{-1}X'Y(i), \quad \widehat{\sigma}^2_{\text{ML}} = (Y(i) - X\widehat{\beta}_{\text{ML}}(i))'(Y(i) - X\widehat{\beta}_{\text{ML}}(i))/n,$$

and collect the first component of $\widehat{\beta}_{\text{ML}}(i)$:

$$\widehat{\beta}_{\text{ML}}(i)_1 = e_1'\widehat{\beta}_{\text{ML}}(i),$$

where e_1 is the first column of the identity matrix of dimension k. Let $\widehat{q}_\eta$ denote the η quantile of $\{\widehat{\beta}_{\text{ML}}(i)_1\}_{i=1}^I$. The parametric bootstrap confidence interval for β_1 (of "nominal" level $1-\alpha$) for β_1 is thus given by

$$C_{\text{Bootstrap}}(Y, X) \equiv [\widehat{q}_{\alpha/2}\ ,\ \widehat{q}_{(1-\alpha)/2}].$$

1.6 Credible Sets based on the quantiles of the posterior

The quantiles of the posterior provide an off-the-shelf algorithm to construct "credible" regions for any real-valued function $g(\theta)$. Start with the parametric model $\{f(x|\theta)\}$ and let $\pi(\theta)$ be a prior. Suppose that we can generate I draws from the posterior distribution $\pi(\theta|x)$. Denote these draws $\{\theta(i)\}_{i=1}^I$. For each of these draws we can construct

$$\Big\{g(\theta(i))\Big\}_{i=1}^{I}.$$

Let $\widehat{p}_\eta$ denote the η quantile of $\{g(\theta(i))\}_{i=1}^I$. The interval with credibility of approximately $1-\alpha$ is equal to

$$C_{\text{Bayes}} \equiv [\widehat{b}_{\alpha/2}\ ,\ \widehat{b}_{(1-\alpha)/2}].$$

Example: Let's go back to the linear regression model with unknown variance. The parameter of interest is still β_1. The prior is $\mathcal{N}_k(0, \sigma^2(\mathbb{I}_n/\lambda))$. The posterior of β (conditional on σ^2) is given by

$$\beta|Y, X, \sigma^2 \sim \mathcal{N}_k(m, V),$$

where

$$\begin{aligned} m &\equiv (X'X + \lambda\mathbb{I}_k)^{-1}X'Y, \\ V &\equiv \sigma^2(X'X + \lambda\mathbb{I}_k)^{-1}. \end{aligned}$$

The posterior distribution for distribution for σ^2 is an Inverse-Gamma(a_n,b_n) where

$$a_n = a_0 + n/2, \quad b_n = b_0 + \frac{1}{2}((Y - Xm)'(Y - Xm) + \lambda m'm).$$

We can generate I draws for σ^2 from the Inverse-Gamma posterior. For each draw $\sigma^2(i)$ we can also take a draw from $\beta|Y, X, \sigma^2$, and collect the first element $\beta(i)_1$. The $\alpha/2$ and $(1-\alpha)/2$ quantiles, provide an approximation to the credible set for β_1.