Maximum Likelihood, Set Estimation

December 24, 2013

Something we should already have mentioned

A $t_n(\mu, \Sigma)$ distribution converges, as $n \to \infty$, to a $N(\mu, \Sigma)$.

Consider the univariate case, where the $t_n(0, 1)$ pdf is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)}\left(1+\frac{(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2}.$$

Using the calculus fact that $(1 + a/n)^n \to e^a$ as $n \to \infty$, it is easy to show that the part of the pdf that depends on x converges to $\exp(-(x-\mu)^2/2)$.

Note also that the t distribution has moments only up to order $\nu - 1$. So it does not have a moment generating function.

Stein's result

- In the standard normal linear model, with a loss function of the form $(\beta - \hat{\beta})' W (\beta - \hat{\beta})$ with W p.s.d., the OLS estimator of $\underset{k \times 1}{\beta}$ is admissible for $k \leq 2$, but not for $k > 2$.

- He proved this by constructing an estimator that dominates OLS.

- However, his estimator is also not admissible.

- Bayesian posterior means with proper priors are of course admissible.

- However, only a narrow class of them dominates OLS, and the class will vary with W.

- So long as an estimator is admissible, that it also dominate OLS is not necessarily desirable.

- These results reflect standard good practice in applied work. When there are many regressors, everyone understands that it is possible to make the predictions from OLS regression estimates turn out badly by including regressors whose estimates have high standard errors. So researchers exclude variables based on prior beliefs.

- But it might be better sometimes to formulate priors explicitly probabilistically, instead of excluding variables informally.

Maximum Likelihood Estimation

- The MLE of θ is the value of θ that maximizes $p(Y \mid \theta)$.
- It may not exist.
- It is rarely justifiable as a Bayesian estimator.
- While it is often thought of as a non-Bayesian estimator, it is not generally unbiased and does not generally have any other good properties except being a function of sufficient statistics.
- Under general conditions that we will study later, it has "approximately" good properties when the sample size is large enough.
- It is usually the starting point for the task of describing the shape of the likelihood. But there are some cases where it is by itself not much use: many local maxima, all of similar height; cases where the peak of the LH is narrow and far from the main mass of probability.

Set Estimation

- This is a procedure that is hard to rationalize from a Bayesian perspective, so we'll come back to that at the end.
- A $100(1-\alpha)\%$ confidence set for the parameter θ in the parameter space Θ is a mapping from observations Y into subsets $S(Y) \subset \Theta$ with the property that for every $\theta \in \Theta$, $P[\theta \in S(Y) \,|\, \theta] = (1-\alpha)$.
- $1-\alpha$ is the **coverage probability** of the interval.
- Such a mapping may not exist, so the definition is commonly relaxed to say that $S(Y)$ is $100(1-\alpha)\%$ interval if

$$\min_{\theta \in \Theta} P[\theta \in S(Y) \,|\, \theta] = 1 - \alpha \,.$$

What is "confidence"?

- It is in practice nearly always treated as if it represented posterior probability. In both popular press and applied economic literature you will see a result that a 95% interval $S(Y)$ for θ has realized value (a, b) described as a result that "it is 95% sure that θ is between a and b" or "the probability that θ is between a and b is 95%".

- So is it connected to posterior probability? Yes, to some extent.

- In the SNLM, confidence sets generated in the usual way (which we will see shortly) have, under a flat prior on β and $\log \sigma$, posterior probability equal to their coverage probabilities.

- In general, a $100(1-\alpha)\%$ confidence set must have posterior probability of at least $1-\gamma$ with unconditional probability at least $1-\alpha/\gamma$. So, e.g., an interval with coverage probability .99 must have posterior probability at least .9 for a set of Y's with pre-sample probability (accounting for uncertainty about θ via the prior) at least .9.

- For 95% confidence sets this result is pretty weak: 95% intervals must have posterior probability at least .9 with unconditional probability at least .5.

-

$$E[P[\theta \in S(Y) \,|\, Y]] = E[P[\theta \in S(Y)]] = E[P[\theta \in S(Y) \,|\, \theta]] = 1 - \alpha \,.$$

Example: Bounded interval parameter space

- We are estimating μ which we know must lie in $[0, 1]$. We have available an estimator $\hat{\mu}$ with the property that $\{\hat{\mu} \,|\, Y\} \sim N(\mu, .1^2)$.
- A 95% confidence interval for μ is therefore $\hat{\mu} \pm .196$.
- Notice that the fact that we know $\mu \in [0, 1]$ did not enter the calculation of the confidence interval. In fact to keep it a subset of the parameter space, we must make the interval $\{\hat{\mu} \pm .196\} \cap [0, 1]$.
- With non-zero probability, the confidence set is empty.
- With non-zero probabiity the confidence set is a very short interval, with very small posterior probability, which conventional mistaken interpretations would treat as indicating great precision of the inference.

Example: Red-green color blind at the traffic light

A witness to a traffic accident is red-green color blind, but can perfectly distinguish yellow. The traffic light is unusual, arranged horizontally. The witness, we have determined, does not like to admit colorblindness, and when asked the color of red and green objects simply announces one or the other color at random, with equal probabilities.

His deposition in this accident states that he observed the traffic light, and it was yellow.

Do we say "with 100%" confidence the light was yellow", or "with 50% confidence the light was yellow"?

Both statements could be valid, but we would have had to commit before seeing the witness's answer to how we would behave if he reported red or green. If we would say "with 50% confidence the light was red" when the report was red, then we have to quote the same confidence level when the light is yellow. But if when the report is red we would say instead "with 100% confidence the light was either red or green", then we are using a 100% confidence set and we should say the light is yellow with 100% confidence.

Of course this is ridiculous. The posterior probability of yellow given the report of yellow is 1.0, regardless of the prior, so every sensible person would simply say the light was surely yellow, and the fact that the witness was red-green color blind is irrelevant, given that the light was not in fact red or green.

"Significant" and "insignificant" results

- What we say here applies about equally to confidence sets and to minimum-size posterior probability sets.
- There is a big difference between a result that posterior probability is concentrated in a small (from the point of view of the substance of the problem) region around $\beta = 0$ and the result that the sample data is so uninformative that the posterior probability is spread widely, with a 95% HPD region therefore including $\beta = 0$.
- The former says we are quite sure that β is substantively small. The latter says β could be very big, indeed from looking at the data alone seems more likely to be big in absolute value than small in absolute value.
- Yet is is not uncommon to see one regression study, which found an "insignificant" effect of a variable X, cited as contradicting another study which found a "significant" effect, without any attention to what the probability intervals were and the degree to which they overlap.