MAXIMUM LIKELIHOOD, SET ESTIMATION, MODEL CRITICISM

1. 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.

2. 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.

Date: November 1, 2004.

3. 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.

4. 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 $\theta \in \Theta$, $P[\theta \in S(Y) \mid \theta] = (1-\alpha)$.
- $1-\alpha$ is the **coverage probability** of the set. Such a random set can have different coverage probabilities for different values of θ, but then it is not an exact confidence set.
- An exact confidence set may not exist, so the definition is commonly relaxed to say that $S(Y)$ is $100(1-\alpha)\%$ set if

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

  that is, if its coverage probability is at least $1-\alpha$ for every θ.

5. 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.
- $$E[P[\theta \in S(Y) \mid Y]] = E[P[\theta \in S(Y)]] = E[P[\theta \in S(Y) \mid \theta]] = 1-\alpha\,.$$

  That is, the prior probability, before the data is seen, that an exact $(1-\alpha)\%$ set will contain the true value of the parameter is $1-\alpha$, regardless of the prior distribution on the parameter, and this is the prior expected value of the posterior probability of the set. So the posterior probability of the set, if it is not always equal to the confidence level, must be higher in some samples, lower in others.

6. 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 $\{\hat{\mu} \mid Y\} \sim N(\mu, .1^2)$.
- 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 probability 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.

7. 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.

This is a special case of a general problem with using pre-sample probability statements as if they were posterior probability statements — pre-sample probabilities depend on samples that did not in fact occur.

8. Example: data with different variances

- Suppose Θ is the two-point space $\{0,1\}$ and when $\theta = 0$ our data Y are $N(0,.5)$ while when $\theta = 1$ $Y \sim N(1,1)$.
- An apparently natural way to form a confidence set would be to have it include both 0 and 1 when $Y \in (-.644,.82)$, 0 only when $Y < -.644$, and 1 only when $Y > .82$. This would indeed be an exact 95% confidence set, as you should be able to check.
- But when Y is below -1, the likelihood ratio in favor of $\theta = 1$ is large. Indeed the strongest likelihood ratios in favor of $\theta = 1$ occur for large negative (or positive) Y's.
- So making such observations deliver a confidence set containing only $\theta = 0$ is unreasonable.
- There are ways to construct more reasonable confidence sets in this example, and of course posterior probability intervals would not show this kind of anomaly.

9. "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 it 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.

10. Old business

- Lancaster's definition of a natural conjugate prior: It implies that any prior pdf that has the form $\pi_0(\beta)\ell(Y^*,\theta)$, where $\ell(Y^*,\theta)$ is what we have called a conjugate prior, is a "Lancaster-conjugate" prior, regardless of what π_0 is.
- Gelman, Carlin, Stern and Rubin label Lancaster's "natural conjugate" priors just plain "conjugate" priors, and use "natural conjugate" to refer to priors that are in the same family as the likelihood function. Probably this is the best usage. So our definition of "conjugate" in lecture should instead be labeled "natural conjugate", and Lancaster in correspondence has agreed that the text should make this distinction.

11. A general method for constructing confidence regions

- For each $\theta \in \Theta$, choose a test statistic $T(Y,\theta)$, where Y is the observable data.
- For each θ choose a **rejection region** $R(\theta) \subset \Theta$ such that $P[T(Y,\theta) \in R(\theta) \mid \theta] = \alpha$.
- Define $S(Y) = \{\theta \mid T(Y,\theta) \notin R(\theta)\}$.
- Then $S(Y)$ is a $100(1-\alpha)\%$ confidence region for θ.

12. Remarks on the general method

- Regions constructed this way are exact confidence regions.
- When Y is continuously distributed, it is always possible to find $T(Y,\theta)$'s and $R(\theta)$'s to implement this idea.
- It is generally not possible to use this idea to produce confidence regions for individual elements of the θ vector or linear combinations of them.
- There are obviously many ways to choose the T and R functions, and thus many confidence regions, for a given α value.

13. Pivotal quantities

- Sometimes we can find a collection of functions $T^*(Y,\theta)$ with the property that the distribution of $\{T^*(Y,\theta) \mid \theta\}$ does not depend on θ. In this case we call T^* a **pivotal quantity** or **pivot** for θ.
- Constructing confidence regions based on pivots is particularly easy: One defines a single region R^* such that $P[T^*(Y,\theta) \in R] = \alpha$, and this rejection

region defines the test used to construct the confidence region for all values of θ. Most confidence regions actually used in practice are based on pivots.

- Certain kinds of pivots also make confidence regions based on them correspond to posterior probability regions with probabilities matching the confidence levels under certain flat priors.
- A leading special case is the SNLM, in which

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \text{ and } \frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{\hat{u}'\hat{u}}$$

  are a pair of jointly pivotal quantities and confidence regions based on the distribution of these statistics conditional on β, σ turn out to coincide with posterior probability regions with probabilities corresponding to $1-\alpha$ if the $(1/\sigma)d\sigma\, d\beta$ flat prior is used.
- Other cases: scale parameters in general, like α in the Gamma or σ in the t distribution, allow this kind of construction, as do location parameters in general, like μ in the t distribution.
- Most confidence regions in actual use are not only based on pivots, they are based on location-scale pivots that allow this flat-prior Bayesian interpretation.

14. Tests

- Each $T(Y,\theta)$, $R(\theta)$ pair in our construction of a confidence region makes up what is known as a **statistical test** of the **null hypothesis** that θ is the true value of the parameter. The parameter α is known as the **significance level** or **size** of the test.
- So an exact confidence region can always be interpreted as a collection of statistical tests with significance level α.
- More generally, we can consider tests of hypotheses that do not consist of a single point in Θ. For such compound hypotheses, DeGroot and Schervish distinguish level, or significance level, and size. The null hypothesis is some set $\Omega_0 \in \Theta$ and the test still takes the form of a statistic $T(Y)$ and rejection region R. Only now it is possible that $P[T(Y) \in R \mid \theta]$ differs across θ's in Ω_0. The standard definitions now say that the test has significance level α if $P[T(Y) \in R \mid \theta] \leq \alpha$ for all $\theta \in \Omega_0$, and that it has size α if it has level γ for all $\gamma \geq \alpha$.

15. Power

- The **power function** of a test is $P[T(Y) \in R \mid \theta]$ considered as a function of θ over $\Theta \ominus \Omega_0$. (It could be extended to range over Ω_0 also.)

- We would like a test to have a small size and have large values of the power function for θ outside Ω_0.
- We would like a test to be **unbiased**, meaning that the infimum of $P[T(Y) \in R \mid \theta]$ over $\Theta \ominus \Omega_0$ is no smaller than the supremum over Ω_0 of the same thing. (Here $\ominus$ is the set difference operator: $A \ominus B = A \cap B^c$.

16. Linear combinations of parameters

- In the SNLM with $d\sigma/\sigma$ prior, consider a $m \times k$ matrix R used to form m linear combinations of β.
$$\{R\beta \mid Y, X\} \sim t_{T-k}(R\hat{\beta}, R\Sigma_\beta R')\,, \tag{$*$}$$
where $\Sigma_\beta = (\hat{u}'\hat{u}/(T-k))(X'X)^{-1}$.
- This implies (it is not hard to show, but we are skipping the algebra) that
$$\frac{\{(\beta-\hat{\beta})'R'(R\Sigma_\beta R')^{-1}R(\beta-\hat{\beta}) \mid Y, X\}}{m} \sim F(m, T-k)\,.$$
- There is a completely analogous non-Bayesian result for the same pivot: same as ($*$) but with the conditioning on β, σ instead of on Y, X.

17. Confidence sets for linear combinations

- Obviously we can use this pivot to develop elliptical confidence sets for arbitrary linear combinations of coefficients, or individual coefficients (where they are equivalent to intervals based on the t_{T-k} distribution).
- Many programs, including R, report automatically what is called the F-statistic for the equation or regression. It is the statistic we are discussing with R a selection matrix — all zeros and ones, such that $R\beta$ is the $(k-1) \times 1$ vector obtained by deleting the constant term from the regression.
- What is the constant term? Most of the time, we include as the first or last column of X a vector of ones. If we write the equation for one observation as
$$Y_t = \beta_0 + \sum_{j=1}^{k} \tilde{X}_{jt}\beta_j + \varepsilon_t\,,$$
This is equivalent to $Y = X\beta + \varepsilon$, with $X = \begin{bmatrix} \underset{T\times 1}{\mathbf{1}} & \tilde{X} \end{bmatrix}$, with β consisting of the $k+1$ β_j's.
- If this F-statistic is not large, it may imply that $R\beta = 0$ is inside a standard confidence set or **credible set** (some authors' term for a set with a given posterior probability). For the particular R we are considering, this means that a parameter vector with every element zero except for the constant term is

inside the confidence/credible set, and thus that in some sense the whole regression is “insignificant”.