Statistical Theory
Quarter 4

James Pringle
taught by Constantine the Greek

26 March 2013 - 17 May 2013

1 Tuesday, 26 March 2013

Test review

The second problem had to do with reducing the data into a function. X to $g(X)$. When you do this do you have a reduction in the information about the parameter? You can write

$$\text{pr}(X|\theta) = p(g(X)|\theta)p(X|g(X),\theta) \tag{1}$$

Take the score from here. It is the derivative of the log. So the score is

$$S_{FM} = d\log(p(g(X)|\theta)/d\theta + d\log p(X|g(X),\theta)/d\theta \tag{2}$$

The variance of the first thing is

$$\text{var}g + \text{var}x|g \tag{3}$$

and the covariance is

$$E[S_g S_{X|g}] \tag{4}$$

This is a sequential accrue-al of information. When you add pieces of data like this, then the information is additive. The scores are uncorrelated. First condition on g in (4), you have

$$E[S_g S_{X|g}] = E[E[S_g S_{X|g}]|g] = E[S_{X|g}E[S_g]] = 0 \tag{5}$$

since the expectation of a score is 0.

Lecture

In the remaining part of the class, we are looking at large samples. Whether there is a general estimator that at least approximately always achieves "this information bound" for unbiased estimators. Two weeks ago, we left off at a point. We resume there.

This only happens if the estimator is proportional to the score $d \log(p(X|\theta)/d\theta$. We will discuss the properties of the maximum likelihood estimator (consistency–in large samples it is closer and closer to the true value). It gets closer and closer to the variance of the score.

Review from chap 7. Definitions of convergence in probability / almost everywhere. Convergence in distribution.

In many situations, you have a limiting random variable be normal (pg 5). You may want a 95% confidence interval. You know it is approximately

$$\hat{\theta} \pm 1.96 \frac{\sqrt{v_0}}{\sqrt{n}} \tag{6}$$

but you don't know v_0. So you have a $\hat{v} \rightarrow_p v_0$. So using the convergence properties and Slutsky's theorem you can make good approximations.

Two important results: MLE is consistent, with variance approximately the Cramer - Rao bound. The proofs are simple, but key.

Consistency

Based on the MLE being a minimizer of the Kullback Leibler distance. That is the reason for the consistency. What is this KL distance? Reminder: for two distributions $p_1(X)$ and $p_2(X)$, how do you tell if the distros are the same? You can use the metric $\sup |p_1(x) - p_2(x)|$. The KL distance is

$$k_{21} = E_{p_1(x)} \left[\log \left(\frac{p_1(x)}{p_2(x)} \right) \right] = \int \log \left(\frac{p_1(x)}{p_2(x)} \right) p_1(x) dx \tag{7}$$

It is another such summary. It is nonnegative and equal to 0 iff $p_1 = p_2$. Also

$$-k_{21} = E_{p_1(x)} \left[\log \left(\frac{p_1(x)}{p_2(x)} \right) \right] \leq \log E_{p_1(x)} \frac{p_2(x)}{p_1(x)} = 0 \tag{8}$$

because $E_{p_1(x)} \frac{p_2(x)}{p_1(x)} = \int \frac{p_2(x)}{p_1(x)} p_1(x) dx = 1$

Intution for consistency of MLE: In a parametric model, with parameter θ, (there are some conditions for consistency). We have $\text{pr}(X|\theta), \theta \in \Theta$ and $X_i, i = 1, \cdots, n$, iid. The MLE $\hat{\theta}$ maximizes $\prod_i^n \text{pr}(X_i|\theta)$. Or equivalently, maximized $1/n \sum \log(\text{pr}(X_i|\theta))$ or minimizes $1/n \sum -\log(\text{pr}(X_i|\theta))$. With respect to what? The minus sign "looks better" if it is $1/n \sum \log(\text{pr}(X|\theta_0))$ where θ_0 is the truth. But we also have that it minimizes

$$\frac{1}{n} \sum [\log \text{pr}(X_i|\theta_0) - \log \text{pr}(X_i|\theta)] = \frac{1}{n} \sum [\log \frac{\text{pr}(X_i|\theta_0)}{\text{pr}(X_i|\theta)}] \tag{9}$$

$$\rightarrow E_{\theta_0} \log \frac{\text{pr}(X_i|\theta_0)}{\text{pr}(X_i|\theta)} \tag{10}$$

$$= K_{\theta, \theta_0} \tag{11}$$

This limit must exist. One thing that requires a lot of effort to justify is that the limit of the minimizer should be the minimizer of the limit. The minimizer of the limit is θ_0. We want the limit of the minimizer (the probability limit of $\hat{\theta}$. And the θ_0 is the minimizer of the limiting process.

Now let's say there is a finite number of values for θ and so that the model has a finite number of distributions. See slide 6 for the assumptions.

Can replace conditions A3, A4 (3 and 4, Frangakis doesn't think those conditions are right) with Θ being compact (closed and bounded); $\log p(X|\theta)$ continuous in θ, X. There should exist $d(x) : |logp(X|\theta)| \leq d(x)$ so that $E_{\theta_0} d(x) < \infty$.

Variance approximately the Cramer - Rao bound

QUESTIONS

1. Projection of score
2. Prove Slutsky's theorem
3. Reveiw the proof on page 7.

2 Thursday, 28 May 2013

Review

There was a question from the last class about the condition (A4, consistency of MLE). We looked at $\log \text{pr}(X|\theta)$. It must be continuous in θ for every X, and there must exist a function $d(x) : |\log \text{pr}(X|\theta)| \leq d(x)$ for all θ and $E_{\theta_0}(d(x)) < \infty$.

Lecture

Look at theorems 10.1.6 and 10.1.12 in Casella and Berger. One is the Neymann Scott paradox. These are examples of inconsistent MLEs.

Suppose you have a set of patients, denoted by i. They have cancer of the liver. Take a biopsy of the cancer and measure the RNA expression of a particular genes. Take two measurements of that (technical replicates). This can be done by splitting the biopsy into two. So we have X_{1i} and X_{2i} for each person (and assumed to be independent given μ_i and σ^2. Also assume the X_{1i} and X_{2i} are independent across i. The measurements are assumed to be distributed as $N(\mu_i, \sigma^2)$.

Let's say that $\mu_i \in \mathbb{R}$ and $\sigma^2 > 0$. Suppose we want to find the MLE of these parameters. Are the MLE for μ_i and σ^2 consistent?

If you do not condition on μ_i then you could say that X_{2i}, X_{1i}, μ_i are iid. The MLE is

$$\prod_i \frac{1}{2\pi\sigma^2} \exp\{-\frac{(x_{1i}-\mu_i)^2+(X_{2i}-\mu_i)^2}{2\sigma^2}\} \tag{12}$$

Maximizing, you get that $\hat{\mu_i} = \bar{X_i}$ and

$$\sigma_n^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ji}-\bar{X_i})^2 \tag{13}$$

Asymptotically, $1/n-p$ goes to $1/n$. Obviously, $\hat{\mu_i}$ is not consistent for μ_i as μ grows. Let

$$\frac{S_i^2}{\sigma^2} = \frac{\sum_{j=1}^{2}(X_{ji}-\bar{X_i})^2}{\sigma^2} \sim \chi^2_{1d.f.} \tag{14}$$

This leads us to $\hat{\sigma}_n^2/\sigma^2 = 1/2n\sum_{i=1}^{n}\frac{S_i^2}{\sigma^2}$. This converges to 1/2. The reason is that we lose degrees of freedom from (14). Note: we can still consistently estimate the σ^2. Just multiply by 2.

Example 2

You have a $N(\mu,\sigma^2)$ and a $N(\mu,1)$. Suppose X is half the sum of these two. Easy to find the mean. How to find the unknown variance? Can take a sample.

Likelihood of $X_1,\cdots,X_n|\mu,\sigma^2=\theta$.

$$L(X|\theta) = \prod_{i=1}^{n}\{\frac{1}{2\sqrt{2\pi}\sigma}\exp(-1/2\sigma^2(x_i-\mu)^2) + \frac{1}{2\sqrt{2\pi}}\exp(-1/2(x_i-\mu)^2)\} \tag{15}$$

Suppose you have an indicator Z_i which is 1 if X_i comes from $N(\mu,\sigma^2)$ and 0 if it comes from the other one. Then that is why you add up the densities times 1/2.

Now try to find the supremum of (15) over μ and σ^2. If you tell a computer to try to maximize this, then it will crash. Choose $\mu = X_1$ for maximizing algorithm. Then (15) becomes

$$L(X|\theta) = \frac{1}{2\sqrt{2\pi}}(1/\sigma+1)\prod_{i=2}^{n}\{\frac{1}{\sqrt{2\pi}\sigma}\exp(-1/2\sigma^2(x_i-\mu)^2)+\frac{1}{\sqrt{2\pi}}\exp(-1/2(x_i-\mu)^2)\} \tag{16}$$

The mixture problem is extremely important in physics. They measure a manifestation of a process that you are looking at. Most of these problems are mixture problems.

Now, note 4. The fact that the MLE failed does not mean that you cannot measure μ and σ^2 consistently. There is a local stationary point.

You can find the variance of $\bar{X}$ and find the sample second moment.

QUESTIONS

1. Check out LyX.
2. Check out XeTeX

3 Tuesday, 2 April 2013

Review

Talked about the direction of the course. The natural flow of things.

Lecture

Delta method for univariate (5.5.24) and vector (5.5.28).

QUESTIONS

1. What the heck is the score?

4 Tuesday, 16 April 2013

Lecture

Given $E(Y|X_i\beta) = X'\beta$. You solve $\sum(Y - X'\beta)X = 0$. For estimating equations, $\sum(Y_i - X_i'\beta)g(X_i) = 0$.

Semiparametric models: Another set of parameters is (β, η) needed to characterize the distribution. We are talking about iid data D_i from a distribution.β is small dimension, η is remaining factors.

EXAMPLE 1: For (X_i, Y_i) iid, then $E(Y_i|X_i\beta) = g(X_i, \beta)$. in this example, the η is the remaining part of (X_i, Y_i).

EXAMPLE 2: The Cox model of proportional hazards. States that the hazard $\lambda(t|Z, \beta) = \lambda(t)\exp(Z'\beta)$. So η is a label about what remains unknown. Here, $\lambda(t)$ is what is unknown.

EXAMPLE 3: When looking at data X_i, you want to see the mean, $E(X_i) = \beta$.

2. For mathematical convenience. The focus is often on estimators $\hat{\beta}$ of β. That are (i) asymptotically linear, i.e.,

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_i \phi(D_i, \beta, n) + o_p(1) \tag{17}$$

(ii) regular if

$$\sqrt{m}\left(\begin{bmatrix} \beta_n \\ \eta \end{bmatrix} - \begin{bmatrix} \beta \\ \eta \end{bmatrix} \right) \to c \tag{18}$$

then $\sqrt{m}(\hat{\beta} - \beta)$ and $\sqrt{m}(\hat{\beta} - \beta_n)$ have the same distribution.

QUESTIONS

1. Cox regression

5 Thursday, 18 April 2013

Lecture

We are basically on chapter 9 of the text in Casella and Berger. The newest thing on this part is getting confidence regions simultaneuosly for multiple parameters.

5.1 Introduction

Suppose we observe a random sample with variance known, $X_1, X_2, \cdots, X_n$ iid $N(\mu, \sigma_0^2)$. We know that $\bar{X}$ is MVUE and consistent for μ. May want to keep track of the variance of the estimators.

We know $\bar{X}$ does no contain much information about plausible values of μ, i.e. $\text{pr}(\bar{X} \mid \mu, \sigma^2)$ is much better.

One way to consider "plausible" values is to sketch the likelihood $\text{pr}(\mathbf{X} \mid \mu, \sigma^2)$. If you plot this likelihood, it picks the sample average. Get $\propto \exp(-\frac{n(\mu-\bar{X})^2}{2\sigma_0^2})$. Usually the likelihood function looks like a normal density. You get intervals (abline(h = FOO)) where the likelihood is equal for the endpoints. For example $l_{max}/l_a = 8$. Or in other words you are looking for all μ such that

$$\frac{\text{pr}(\mathbf{X} \mid \mu, \sigma^2)}{\text{pr}(\mathbf{X} \mid \hat{\mu}, \sigma^2)} \geq c \tag{19}$$

Note if we define an interval like this... the right limit (RL) of the usual 95 % confidence interval for μ has a likelihood ratio

$$\frac{\text{pr}(\mathbf{X} \mid \hat{\mu}, \sigma^2)}{\text{pr}(\mathbf{X} \mid \mu = \text{RL}, \sigma^2)} = 6.8 \tag{20}$$

Another way is to consider a (random over $\mathbf{X}$ interval $[l(\mathbf{X}), u(\mathbf{X})]$ whose probability of covering the desired value is greater than or equal to a constant no matter the value.

We know that

$$\left[\bar{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}\right] \tag{21}$$

has 95% probability of covering μ, no matter what μ is. Sometimes people like a frequentist interval or a Bayesian interval or a likelihood interval.

5.2 Definition

Suppose that X is a random variable that represents an observation from a study (could be conducted repeatedly). It has a likelihood $\text{pr}(X|\theta); \theta \in \Theta$. The random region $I(X) = (l(X), u(X))$ is a $100(1-\alpha)\%$ confidence interval for $\tau(\theta)$ if

1. $l(X) \leq u(X)$

2. $\text{pr}(I(X) \text{ contains } \tau(\theta|\theta)) \geq 1 - \alpha$ for all $\theta \in \Theta$.

Neyman was a Polish statistician. He had a distinguished career at Berkeley. In 1923 he was studying the design of studies. He was working with finite populations.

For example take a pool of 1000, a finite population, take a SRS of 100 people. When you get a SRS without replacement, there is a negative correlation between getting one person then the next. So the X's are not independent in the $\bar{X}$.

Definition: The coverage probability, $\pi(\theta)$, or $I(X)$ for $\tau(\theta)$ is $\text{pr}(\tau(\theta) \in I(X)|\theta)$. **Note:** The text defines the confidence coefficient of $I(X)$. It is the $\inf \pi(\theta)$ (so a 95% CI has confidence coefficient $\geq 95\%$).

Long confidence intervals reflects a lot of uncertainty. Let's now look at a couple of criteria. Some of them nobody pays any attention to.

5.3 Some criteria for evaluating C.I.s, section 9.3 in the text

Criterion 1

Length of the CI or the expected length. The length isn't always fixed. It may be dependent on the data, so we may want to think about the expected value. The idea is to choose CI with small lengths.

Example of the $t_{(n-1)d.f.}$ where we look at the sample mean and relate it by $\chi^2_{(n-1)d.f.}$. The quantity is a known pivotal number. Therefore, we choose any lower and upper values such that the mass in between those bounds is 0.95 or $1-\alpha$. One can try to determine what u and l that give the minimal interval length. By Theorem 9.3.2 in CB, for a number of distributions with a unimodal density, the solution is where the likelihood is the same for each.

Suppose (i) $\int_l^u f(x)dx = 1 - \alpha$. (ii) $f(l) = f(u) > 0$ and (iii) $l \leq x \leq u$ where x^* is the mode of f. Then we have (l, u) is the shortest that satistifies (i).

Two more criteria are based on the following concept of false coverage.

definition: For a confidence interval I(X), the probability of false coverage of θ' given θ is $\text{pr}(\theta' \in I(X)|\theta)$ (depends on open / closed interval).

We have concepts of accuracy and bias. We are invited to read the two additional criteria and the chapter.

Questions

1.

For next time

Pivotal quantities.