Statistics 612: Regular Parametric Models and Likelihood Based Inference

Moulinath Banerjee

December 6, 2006

We continue our discussion of likelihood based inference for parametric models; in particular, we will talk more about information bounds in the context of parametric models, and the role they play in likelihood based inference. We first introduce the multiparameter version of the celebrated Cramer-Rao inequality.

I will not describe the underlying assumptions in details. These are the usual sorts of assumptions one makes for parametric models, in order to be able to establish sensible results. See Page 11 of Chapter 3 of Wellner's notes for a detailed description of the conditions involved. For a multidimensional parametric model $\{p(x,\theta) : \theta \in \Theta \subset \mathbb{R}^k\}$, the information matrix $I(\theta)$ is given by:

$$I(\theta) = E_\theta(\dot{l}(X,\theta), \dot{l}(X,\theta)^T) = -E_\theta\, \ddot{l}(X,\theta)\,,$$

where

$$\dot{l}(X,\theta) = \frac{\partial}{\partial\,\theta}\, l(X,\theta)$$

being a $k \times 1$ column vector (recall that $l(x,\theta) = \log\, p(x,\theta)$), and

$$\ddot{l}(x,\theta) = \frac{\partial^2}{\partial\,\theta\,\partial\,\theta^T}\, l(X,\theta)\,,$$

is a $k \times k$ matrix. Consider a smooth real-valued function $q(\theta)$ that is estimated by some statistic $T(X)$, and let $\dot{q}(\theta)$ denote the derivative of q (written as a $k \times 1$ vector). Let $b(\theta) = E_\theta(T(X)) - q(\theta)$ be the bias of the estimator T, and let $\dot{b}(\theta)$ denote the derivative of the bias. We then have:

$$\mathrm{Var}_\theta(T(X)) \geq (\dot{q}(\theta) + \dot{b}(\theta))^T\, I^{-1}(\theta)\, (\dot{q}(\theta) + \dot{b}(\theta))\,.$$

In particular, if $T(X)$ is unbiased for $q(\theta)$, then

$$\mathrm{Var}_\theta(T(X)) \geq \dot{q}(\theta)^T\, I^{-1}(\theta)\, \dot{q}(\theta)\,.$$

For a proof of this result, see Page 12 of Chapter 3 of Wellner's notes – the proof runs along lines similar to the one–dimensional case. We will not be worried about the construction of exact

unbiased estimators for $q(\theta)$ that attain the information bound; in the vast majority of situations this is not feasible. Rather, we focus on the connection of the MLE $\hat{\theta}_n$ to the information bound arising from the multiparameter inequality above. Consider the asymptotically linear representation of the MLE given by:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta)^{-1} \dot{l}(X_i, \theta) + o_p(1) .$$

Invoke the Delta method to obtain:

$$\sqrt{n}(q(\hat{\theta}_n) - q(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{q}(\theta)^T I(\theta)^{-1} \dot{l}(X_i, \theta) + o_p(1) .$$

It is easily seen that the asymptotic variance of $\sqrt{n}(q(\hat{\theta}_n) - q(\theta))$ is exactly $\dot{q}(\theta)^T I^{-1}(\theta)\, q(\theta)$, the information bound arising from the multiparameter Cramer Rao inequality. The function $\dot{q}(\theta)^T I(\theta)^{-1} \dot{l}(x, \theta)$ (that provides a linearization of the MLE) is called the *efficient influence function* for estimating $q(\theta)$. Motivated by the above considerations, we define efficient influence functions and information bounds for vector-valued functions of θ.

Let ν be a Euclidean parameter defined on a regular parametric mode; $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We can identify ν with the parametric function $q : \Theta \to \mathbb{R}^m$ defined by:

$$q(\theta) = \nu(P_\theta), \text{ for } P_\theta \in \mathcal{P} .$$

Fix $P = P_\theta$ and suppose that q has a derivative $\dot{q}_{k \times m}$ at θ. Define the information bound for ν as:

$$I^{-1}(P \mid \nu, \mathcal{P}) = \dot{q}(\theta)^T I^{-1}(\theta)\, \dot{q}(\theta) . \tag{0.1}$$

Also, define the efficient influence function for ν as:

$$\tilde{l}(\cdot, P \mid \nu, \mathcal{P}) = \dot{q}(\theta)^T I^{-1}(\theta)\, \dot{l}_\theta . \tag{0.2}$$

The definitions of the two quantities above seem to depend on the parametrization θ, but in fact are independent of it, as the notation suggests. This is the content of the following proposition.

Proposition: The information bound $I^{-1}(P \mid \nu, \mathcal{P})$ and the efficient influence function $\dot{l}(\cdot, P \mid \nu, \mathcal{P})$ are invariant under smooth changes of parametrization.

A derivation of this is given on Page 18, Chapter 3 of Wellner's notes; or, you can also look at the document posted on the 612 webpage that discusses this phenomenon separately.

So far, we have been interested in estimating the entire parameter vector θ. We now write $\theta = (\nu, \eta)$, where ν is an m–dimensional sub-parameter of interest and η is a *nuisance parameter*, and estimation of ν needs to be carried out in the presence of the nuisance parameter. We

introduce partitions of the score functions and the information matrix in correspondence with the partitioning (ν, η) of the parameter vector θ. We write:

$$\dot{l}(x,\theta) = \left[\begin{array}{c} \frac{\partial}{\partial \nu} l(x,\nu,\eta) \\ \frac{\partial}{\partial \eta} l(x,\nu,\eta) \end{array} \right] \equiv \left[\begin{array}{c} \dot{l}_\nu(x,\nu,\eta) \\ \dot{l}_\eta(x,\nu,\eta) \end{array} \right],$$

and

$$\frac{\partial}{\partial \eta} \dot{l}_\nu(x,\nu,\eta) \equiv \frac{\partial^2}{\partial \nu \, \partial \eta} l(x,\nu,\eta) = \ddot{l}_{\nu\,\eta}(x,\nu,\eta)_{m \times k-m} \,,$$

and

$$\frac{\partial}{\partial \nu} \dot{l}_\eta(x,\nu,\eta) \equiv \frac{\partial^2}{\partial \eta \, \partial \nu} l(x,\nu,\eta) = \ddot{l}_{\eta\,\nu}(x,\nu,\eta)_{k-m \times m} \,.$$

We define $\ddot{l}_{\nu\,\nu}$ and $\ddot{l}_{\eta\,\eta}$ similarly. Now,

$$I(\theta_0) = \left[\begin{array}{cc} I_{\nu_0\,\nu_0} & I_{\nu_0\,\eta_0} \\ I_{\eta_0\,\nu_0} & I_{\eta_0\,\eta_0} \end{array} \right] = \left[\begin{array}{cc} I_{11} & I_{12} \\ I_{21} & I_{22} \end{array} \right] = - \left[\begin{array}{cc} E_{\theta_0} \, \ddot{l}_{\nu_0\,\nu_0} & E_{\theta_0} \, \ddot{l}_{\nu_0\,\eta_0} \\ E_{\theta_0} \, \ddot{l}_{\eta_0\,\nu_0} & E_{\theta_0} \, \ddot{l}_{\eta_0\,\eta_0} \end{array} \right].$$

We also write $\dot{l}(x,\theta_0) = (\dot{l}_1^T, \dot{l}_2^T)^T$, with

$$\dot{l}_1 = \frac{\partial}{\partial \nu} l(x,\nu_0,\eta_0) \;\text{ and }\; \dot{l}_2 = \frac{\partial}{\partial \eta} l(x,\nu_0,\eta_0) \,.$$

Then, note that $I_{11} = E_{\theta_0}[\dot{l}_1 \, \dot{l}_1^T]$, $I_{12} = E_{\theta_0}[\dot{l}_1 \, \dot{l}_2^T]$ and $I_{22} = E_{\eta_0}[\dot{l}_2, \dot{l}_2^T]$, with $I_{12} = I_{21}^T$. Now define:

$$I_{11.2} = I_{11} - I_{12} \, I_{22}^{-1} \, I_{21} \;\text{ and }\; I_{22.1} = I_{22} - I_{21} \, I_{11}^{-1} \, I_{12} \,.$$

These are p.d. matrices (and are known as Schur complements in the linear algebra literature). Write:

$$I(\theta_0)^{-1} = \left[\begin{array}{cc} I^{11} & I^{12} \\ I^{21} & I^{22} \end{array} \right].$$

From one of the homework problems, it will follow that:

$$I(\theta_0)^{-1} = \left(\begin{array}{cc} I_{11} & I_{12} \\ I_{21} & I_{22} \end{array} \right)^{-1} = \left(\begin{array}{cc} I_{11.2}^{-1} & -I_{11.2}^{-1} \, I_{12} \, I_{22}^{-1} \\ -I_{22.1}^{-1} \, I_{21} \, I_{11}^{-1} & I_{22.1}^{-1} \end{array} \right).$$

Let $q(\theta) = \nu$. Then $\dot{q}(\theta)^T = [I_{m \times m}, 0_{m \times k-m}]$. Now,

$$\sqrt{n} \, (\hat{\nu}_n - \nu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{q}(\theta_0)^T \, I(\theta_0)^{-1} \, \dot{l}(X_i,\theta_0) + o_p(1) \,.$$

We have:

$$\begin{aligned} \dot{q}(\theta_0)^T \, I(\theta_0)^{-1} \, \dot{l}(x,\theta_0) &= [I_{m \times m} : 0_{m \times k-m}] \left[\begin{array}{cc} I^{11} & I^{12} \\ I^{21} & I^{22} \end{array} \right] \left[\begin{array}{c} \dot{l}_1 \\ \dot{l}_2 \end{array} \right] \\ &= [I^{11} : I^{12}] \left[\begin{array}{c} \dot{l}_1 \\ \dot{l}_2 \end{array} \right] \\ &= I^{11} \, \dot{l}_1 + I^{12} \, \dot{l}_2 \\ &= I_{11.2}^{-1} \, [\dot{l}_1 - I_{12} \, I_{22}^{-1} \, \dot{l}_2] \,. \end{aligned}$$

Hence, we can write:

$$\sqrt{n}\,(\hat{\nu}_n - \nu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{11.2}^{-1}\,[\dot{l}_1(X_i) - I_{12}\,I_{22}^{-1}\,\dot{l}_2(X_i)] + o_p(1)\,.$$

We denote by $\tilde{l}_1$ the efficient influence function $I_{11.2}^{-1}\,[\dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2] \equiv I_{11.2}^{-1}\,\lambda_1^\star$ for estimation of the parameter ν; here $l_1^\star = \dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2$ is called the efficient score function for estimating the parameter ν in the presence of η at parameter value $\theta_0 = (\nu_0, \eta_0)$. Note that

$$\sqrt{n}(\hat{\nu}_n - \nu_0) \to_d N(0, I_{11.2}^{-1})\,.$$

Also check that $\mathrm{Cov}\,\tilde{l}_1 = I_{11.2}^{-1}$. Now, $I_{11.2}^{-1}$ is the information bound for estimating ν when the value of the nuisance parameter η is unknown.

Consider the problem of estimating ν_0 when information is available on η, i.e. we know that $\eta = \eta_0$. In this case, we have a regular parametric model of dimension m given by $\{p(x, \nu, \eta_0) : \nu \text{ varying}\}$, and the ordinary score function in this model at parameter value ν_0 is simply given by $\dot{l}_1$ itself. If $\hat{\nu}_0$ denotes the M.L.E. of ν in this (lower–dimensional) model, we have:

$$\sqrt{n}\,(\hat{\nu}_n^0 - \nu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{11}^{-1}\,\dot{l}_1(X_i) + o_p(1)\,;$$

the information bound for estimating ν at ν_0 is simply I_{11}^{-1} (which is also the asymptotic variance of the normalized MLE in this model). How does this compare to $I_{11.2}^{-1}$, the information bound when η is unknown? Now, note that $I_{11}^{-1} = I^{11} - I^{12}\,(I^{22})^{-1}\,I^{21}$ and $I_{11.2}^{-1} = I^{11}$; hence $I_{11.2}^{-1} - I_{11}^{-1} = I^{12}\,(I^{22})^{-1}\,I^{21}$, which is a p.d. matrix. This shows, that as one might intuitively expect, the information bound for estimating ν is smaller when η is known (as compared to when it is not). The loss of information between the two situations is the difference of the efficient information; this is $I_{12}\,I_{22}^{-1}\,I_{21}$.

The phenomenon above can be given a geometric interpretation if we consider the functions $(\dot{l}_1, \dot{l}_2)$ as a vector of functions living in $L_2^0(P_{\theta_0})$, the Hilbert space of mean 0 square integrable functions with respect to the probability measure P_{θ_0}. Recall that the inner product between random variables u and $\tilde{u}$ in $L_2(P_{\theta_0})$ is $\langle u, \tilde{u}\rangle = E_{\theta_0}(u\,\tilde{u})$. For simplicity of presentation, consider the situation where ν is 1–dimensional. Then $\dot{l}_1$ is an element of $L_2^0(P_{\theta_0})$. Consider the closed linear subspace of $L_2^0(P_{\theta_0})$ that is formed by linear combinations of the components of $\dot{l}_2$ (each component of course lives in $L_2^0(P_{\theta_0})$). If $\mathcal{S}$ denotes this subspace, then we know that $\dot{l}_1$ admits a unique decomposition as $\pi_{\mathcal{S}}\,\dot{l}_1 + \pi_{\mathcal{S}^\perp}\,\dot{l}_1$. We claim that $l_1^\star = \pi_{\mathcal{S}^\perp}\,\dot{l}_1$. To show this, observe that:

$$\dot{l}_1 = \dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2 + I_{12}\,I_{22}^{-1}\,\dot{l}_2\,;$$

the latter is obviously in $\mathcal{S}$, so all we need to show is that the former lives in $\mathcal{S}^\perp$. To this end, it suffices to show that $\langle \dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2, \alpha^T\,\dot{l}_2\rangle = 0$ for all vectors α. We have:

$$\langle \dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2, \alpha^T\,\dot{l}_2\rangle \;\; = \;\; E\,[(\dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2)\,(\alpha^T\,\dot{l}_2)]$$

$$
\begin{aligned}
&= \alpha^T E\,[\dot{l}_2\,\dot{l}_1] - I_{12}\,I_{22}^{-1}\,E(\dot{l}_2\,\dot{l}_2^T)\,\alpha \\
&= \alpha^T I_{21} - I_{12}\,I_{22}^{-1}\,I_{22}\,\alpha \\
&= 0\,.
\end{aligned}
$$

Hence, with $\|\ \ \|_0$ denoting the norm in $L_2^0(P_{\theta_0})$, we have:

$$\|\dot{l}_1\|_0^2 = \|\dot{l}_1^\star\|_0^2 + \|\dot{l}_1 - \dot{l}_1^\star\|_0^2\,.$$

The left side is precisely I_{11}, and the right side decomposes as $I_{11.2} + I_{12}\,I_{22}^{-1}\,I_{21}$. Thus, the information loss is precisely the squared length of the (orthogonal) projection of the usual score function in the lower–dimensional submodel (with η known) into the closed linear subspace spanned by the components of the score function for η. If the score function for ν is uncorrelated with the score function for η, so that $I_{12} = 0$, then there is no information loss, and knowledge of η makes no difference to the efficiency of estimation. If ν is m–dimensional, with $m > 1$ then a similar interpretation can be given. We can write $\dot{l}_1 = (\dot{l}_{1,1}, \dot{l}_{1,2}, \ldots, \dot{l}_{1,k})$ with a similar decomposition for $\dot{l}_1^\star$, and in this case $\dot{l}_{1,i}^\star = \pi_{\mathcal{S}^\perp}\,\dot{l}_{1,i}$. This is a consequence of the (easily checked) fact that

$$\mathrm{Cov}\,(\dot{l}_1 - I_{12}\,I_{22}^{-1}\,\dot{l}_2, \dot{l}_2) = 0_{m\times k-m}\,.$$

Exercise: Once again consider one–dimensional ν. Compute the projection of $\dot{l}_1$ into $\mathcal{S}$ by minimizing $E_{\theta_0}\,(\dot{l}_1 - \alpha^T\,\dot{l}_2)^2$ and check that this matches what we have established above.

Tests for composite null hypotheses: We can concentrate on tests of (composite) null hypotheses of the form $\eta = \eta_0$, where ν is left unspecified. We first consider the likelihood ratio statistic for testing $\eta = \eta_0$.

Likelihood ratio statistic: With $l_n(\theta) \equiv \sum_{i=1}^n l(X_i, \theta)$ denoting the log-likelihood function based on n observations, we have:

$$2\,\log\,\lambda_n = 2\,\left(l_n(\hat{\theta}_n) - l_n(\hat{\theta}_n^0)\right)\,,$$

where $\hat{\theta}_n = (\hat{\nu}_n, \hat{\eta}_n)$ and $\hat{\theta}_n^0 = (\hat{\nu}_n^0, \eta_0)$. Now, we can write:

$$2\,\log\,\lambda_n = 2\,[l_n(\hat{\nu}_n, \hat{\eta}_n) - l_n(\nu_0, \eta_0)] - 2\,\left[l_n(\hat{\nu}_n^0, \eta_0) - l_n(\nu_0, \eta_0)\right] \equiv I_n - II_n\,.$$

Now, by the asymptotics of the likelihood ratio statistic for testing $\theta = \theta_0$ in the full k dimensional model, we have:

$$I_n = \sqrt{n}\,(\hat{\theta}_n - \theta_0)^T\,I(\theta_0)\,\sqrt{n}\,(\hat{\theta}_n - \theta_0) + o_p(1)\,,$$

and by the asymptotics of the likelihood ratio statistic for testing $\nu = \nu_0$ in the (reduced) m dimensional model (with η known to be fixed at η_0),

$$II_n = \sqrt{n}\,(\hat{\nu}_n - \nu_0)^T\,I_{\nu_0\,\nu_0}\,\sqrt{n}\,(\hat{\nu}_n - \nu_0) + o_p(1)\,.$$

Recall that $\dot{l} = (\dot{l}_1^T, \dot{l}_2^T)^T$. Let $Z = (Z_1^T, Z_2^T)$ be a normal random vector with dispersion matrix $I(\theta_0)$. Now, using the facts that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{bmatrix} \dot{l}_1 \\ \dot{l}_2 \end{bmatrix} \rightarrow_d \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N(0, I(\theta_0))$$

and that

$$\begin{bmatrix} \sqrt{n}\,(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\nu}_n^0 - \nu_0) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}}\, I(\theta_0)^{-1} \sum_{i=1}^{n} \dot{l}(X_i) \\ \frac{1}{\sqrt{n}}\, I_{11}^{-1} \sum_{i=1}^{n} \dot{l}_1(X_i) \end{bmatrix} + o_p(1)$$

we conclude that

$$\begin{bmatrix} \sqrt{n}\,(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\nu}_n^0 - \nu_0) \end{bmatrix} \rightarrow_d \begin{bmatrix} I(\theta_0)^{-1} Z \\ I_{11}^{-1} Z_1 \end{bmatrix}.$$

It follows by continuous mapping that $I_n - II_n$ converges in distribution to

$$Z^T I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} Z - Z_1^T I_{11}^{-1} I_{11} I_{11}^{-1} Z_1$$

and this, by one of the homework problems, follows a χ^2_{k-m} distribution.

Score and Wald statistics: The Wald statistic for testing $\eta = \eta_0$ is given by:

$$W_n = n\,(\hat{\eta}_n - \eta_0)^T \hat{I_{22.1}}\,(\hat{\eta}_n - \eta_0)\,,$$

and has a limiting χ^2_{k-m} distribution, since $\sqrt{n}\,(\hat{\eta}_n - \eta_0) \rightarrow_d N(0, I_{22.1}^{-1})$.

To set up the score statistic, define:

$$Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(X_i, \theta)\,.$$

To test for the full parameter $\theta = \theta_0$, recall that the score statistic is given by $Z_n(\theta_0)^T I^{-1}(\theta_0)\, Z_n(\theta_0)$ (and converges to a χ^2_k distribution). To test $\eta = \eta_0$ define the score statistic S_n as

$$S_n = Z_n(\hat{\theta}_n^0)^T I^{-1}(\hat{\theta}_n^0)\, Z_n(\hat{\theta}_n^0)\,.$$

Now, note that

$$Z_n(\hat{\theta}_n^0) = \begin{bmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_\nu(X_i, \hat{\nu}_n^0, \eta_0) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) \end{bmatrix}.$$

Denoting the bottom component of the vector on the extreme right of the above display by $Z_{n,2}(\hat{\theta}_n^0)$, the score statistic becomes:

$$S_n = Z_{n,2}(\hat{\theta}_n^0)^T I^{-1}(\hat{\theta}_n^0)_{22}\, Z_{n,2}(\hat{\theta}_n^0)\,.$$

Next, $I^{-1}(\hat{\theta}_n^0)_{22} \to_p I^{-1}(\theta_0)_{22} = I_{22.1}^{-1}$. Also,

$$\begin{aligned}
Z_{n,2}(\hat{\theta}_n^0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_\eta(X_i, \nu_0, \eta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) - \dot{l}_\eta(X_i, \nu_0, \eta_0)) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_2(X_i) + \left[\frac{1}{n} \sum_{i=1}^{n} (\dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) - \dot{l}_\eta(X_i, \nu_0, \eta_0)) \right] \sqrt{n} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_2(X_i) + \frac{1}{n} \sum_{i=1}^{n} \ddot{l}_{\eta\,\nu}(X_i, \nu_0, \eta_0) \sqrt{n} \, (\hat{\nu}_n^0 - \nu_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_2(X_i) + [-I_{21} + o_p(1)] \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{11}^{-1} \, \dot{l}_1(X_i) + o_p(1) \right] + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\dot{l}_2(X_i) - I_{21} \, I_{11}^{-1} \, \dot{l}_1(X_i) \right) + o_p(1) \, .
\end{aligned}$$

One of the steps above needs some clarification. The result:

$$\left[\frac{1}{n} \sum_{i=1}^{n} (\dot{l}_\eta(X_i, \hat{\nu}_n^0, \eta_0) - \dot{l}_\eta(X_i, \nu_0, \eta_0)) \right] \sqrt{n} = \frac{1}{n} \sum_{i=1}^{n} \ddot{l}_{\eta\,\nu}(X_i, \nu_0, \eta_0) \sqrt{n} \, (\hat{\nu}_n - \nu_0) + o_p(1) \, ,$$

though intuitively meaningful does need a little argument. You cannot use a Taylor expansion with a standard remainder term for vector valued functions. Consider the difference on the left side componentwise (there are $k-m$ components). Let $\eta = (\eta_1, \eta_2, \ldots, \eta_{k-m})$. Now, the j'th component of the difference is:

$$\frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \hat{\nu}_n^0, \eta_0) - \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \nu_0, \eta_0) \, .$$

We can write:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \hat{\nu}_n^0, \eta_0) &= \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \nu_0, \eta_0) + \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial}{\partial \nu} \, \dot{l}_{\eta_j}(X_i, \nu_0, \eta_0) \right]_{1 \times m} (\hat{\nu}_n^0 - \nu_0) \\
&\quad + \frac{1}{2\,n} \, (\hat{\nu}_n - \nu_0)^T \, \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \, \nu^2} \, \dot{l}_{\eta_j}(X_i, \tilde{\nu}_{n,j}, \eta_0) \, (\hat{\nu}_n - \nu_0) \, ,
\end{aligned}$$

where $\tilde{\nu}_{n,j}$ lies on the straight line joining $\hat{\nu}_n^0$ and ν_0. Hence:

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \hat{\nu}_n^0, \eta_0) - \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\eta_j}(X_i, \nu_0, \eta_0) \right] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial}{\partial \nu} \, \dot{l}_{\eta_j}(X_i, \nu_0, \eta_0) \right]_{1 \times m} \sqrt{n} (\hat{\nu}_n^0 - \nu_0)$$

$$+\frac{1}{2\,n}\,(\hat{\nu}_n-\nu_0)^T\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\,\nu^2}\,\dot{l}_{\eta_j}(X_i,\tilde{\nu}_{n,j},\eta_0)\;\sqrt{n}\,(\hat{\nu}_n-\nu_0)\,.$$

The second term on the right side of the above display is $o_p(1)$ by virtue of the fact that $\sqrt{n}(\hat{\nu}_n^0-\nu_0)$ is $O_p(n^{-1/2})$ and furthermore

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\,\nu^2}\,\dot{l}_{\eta_j}(X_i,\tilde{\nu}_{n,j},\eta_0)=O_p(1)\,,$$

using the fact that $\tilde{\nu}_{n,j}$ converges to ν_0 (in probability) and the assumption that for all i,j,k,

$$\left|\frac{\partial^3}{\partial\,\theta_i,\partial\,\theta_j\,\partial\,\theta_k}\,l(x,\theta)\right|\le M_{ijk}(x)\,,$$

uniformly over θ in a neighborhood of θ_0, where $E_{\theta_0}(M_{ijk})<\infty$ (see, for example, Assumption (A.3) on Page 5 of Chapter 4 of Wellner's notes). It follows that:

$$\sqrt{n}\,\left[\frac{1}{n}\sum_{i=1}^{n}\dot{l}_\eta(X_i,\hat{\nu}_n^0,\eta_0)-\frac{1}{n}\sum_{i=1}^{n}\dot{l}_\eta(X_i,\nu_0,\eta_0)\right]=\frac{1}{n}\sum_{i=1}^{n}\left[\ddot{l}_{\eta\,\nu}(X_i,\nu_0,\eta_0)\right]_{k-m\times m}\sqrt{n}(\hat{\nu}_n^0-\nu_0)+o_p(1)\,.$$

1 Problems

(1) (a) Consider a regular parametric model $\{f(x,\theta):\theta\in\Theta\subset\mathcal{R}^k\}$. Fix $\theta_0\in\Theta$ and let $I(\theta_0)$ denote the (positive definite) information matrix at the point θ_0. Let $\theta=(\nu,\eta)$ be a partitioning of θ and let $\theta_0=(\nu_0,\eta_0)$. Here ν is an m–dimensional parameter. Write the information matrix as:

$$I(\theta_0)=\begin{pmatrix} I_{11} & I_{12}\\ I_{21} & I_{22}\end{pmatrix}.$$

As defined in class, let

$$I_{11.2}=I_{11}-I_{12}I_{22}^{-1}I_{21}$$

and let $I_{22.1}$ be defined similarly with 2 and 1 swapped. Write,

$$I(\theta_0)^{-1}=\begin{pmatrix} A & B\\ C & D\end{pmatrix}$$

and use the fact that $I(\theta_0)\,I(\theta_0)^{-1}=I=I(\theta_0)^{-1}\,I(\theta_0)$ to show that,

$$I(\theta_0)^{-1}=\begin{pmatrix} I_{11.2}^{-1} & -I_{11.2}^{-1}\,I_{12}\,I_{22}^{-1}\\ I_{22.1}^{-1}\,I_{21}\,I_{11.2}^{-1} & I_{22.1}^{-1}\end{pmatrix}.$$

(b) Use (a) to deduce that if $Z_{k\times1}$ follows a multivariate normal distribution with dispersion matrix $I(\theta_0)$, then

$$Z^T\,I(\theta_0)^{-1}\,Z^T-Z_2^T\,I_{22}^{-1}\,Z_2\sim\chi^2_m\,.$$

Here Z_1 denotes the first m components of Z and Z_2 the remaining $k-m$. What is the distribution of

$$Z^T\, I(\theta_0)^{-1}\, Z^T - Z_1^T\, I_{11}^{-1}\, Z_1\,?$$

(c) Consider the Rao/Score statistic for testing $H_0 : \eta = \eta_0$. This is

$$R_n \equiv Z_n(\hat{\theta}_n^0)^T\, I^{-1}(\hat{\theta}_n^0)\, Z_n(\hat{\theta}_n^0)$$

where, $\hat{\theta}_n^0 = (\hat{\nu}_n^0, \eta_0)$ is the MLE of θ under the null hypothesis and

$$Z_n(\theta) = \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\nu(X_i, \nu, \eta) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\eta(X_i, \nu, \eta) \end{pmatrix}.$$

Show that under H_0, R_n has an asymptotic χ^2_{k-m} distribution.

(d) Consider the null hypothesis $H_0 : \nu = \nu_0$. Let $\hat{\theta} = (\hat{\nu}, \hat{\eta})$ and let $\hat{\eta}_0$ be the MLE of η obtained under H_0. Show that, under H_0,

$$\sqrt{n}(\hat{\eta} - \hat{\eta}_0) = -I_{22}^{-1}\, I_{21}\, \sqrt{n}\,(\hat{\nu} - \nu_0) + o_p(1)\,. \qquad (\star)$$

Now consider the likelihood ratio statistic, $2\,\log\,\lambda_n$ for testing H_0. Show, that under H_0,

$$2\,\log\,\lambda_n = \sqrt{n}(\hat{\theta} - \theta_0)\, I(\theta_0)\, \sqrt{n}(\hat{\theta} - \theta_0) - \sqrt{n}(\hat{\eta}_0 - \eta_0)\, I_{22}\, \sqrt{n}(\hat{\eta}_0 - \eta_0) + o_p(1)\,.$$

Now, using the representation $(\star)$ or otherwise, show that

$$2\,\log\,\lambda_n = n\,(\hat{\nu} - \nu_0)^T\, I_{11.2}\,(\hat{\nu} - \nu_0) + o_p(1)\,.$$

Hence, deduce the asymptotic distribution of the likelihood ratio statistic.

(e) Show that $I_{12}\, I_{22}^{-1}\, \dot{l}_2$ is the closest element to $\dot{l}_1$ in the span of $\dot{l}_2$ in the sense that

$$\text{argmin}_\alpha\, E\,(\dot{l}_1(X) - \alpha^T\, \dot{l}_2(X))^2 = I_{12}\, I_{22}^{-1}\,.$$

(2) (a) Let $X_1, X_2, \ldots, X_n$ be a sample from the exponential distribution with parameter θ and let $Y_1, Y_2, \ldots, Y_n$ be a sample from an exponential μ distribution. Also, let the first sample be independent of the second. Consider testing the null hypothesis $\mu = 2\,\theta$. Use an appropriate reparametrization to recast the null hypothesis into the form $\psi = \psi_0$ for some fixed ψ_0, compute the likelihood ratio and Wald statistics and determine their asymptotic distributions under the null.

(b) For $i = 1, 2, \ldots, k$ let $X_{i1}, X_{i2}, \ldots, X_{in}$ be independent samples from Poisson distributions, $\text{Poi}(\theta_i)$ respectively. Find the likelihood ratio test and its asymptotic distribution for testing $H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$.

(3) Let X follow the Beta(α, β) distribution. Thus,

$$p(\theta, x) = p(\alpha, \beta, x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}\,1\,(x \in (0,1))\,.$$

This is a regular parametric model. Let,

$$\psi(\alpha) \equiv \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\,.$$

(a) Compute the scores for α and β based on one X following Beta(α, β). Hence, compute the information matrix.

(b) What is the information for α if β is known? What if β is unknown? Draw a picture of the scores to illustrate this geometrically.

(c) Now, suppose that we have i.i.d. observations $X_1, X_2, \ldots, X_n$ from a Beta(α_0, β_0) distribution. Let $q(\theta) = \psi(\alpha) - \psi(\beta)$.

(i) Compute the efficient influence function and the information bound for the estimation of q at the point $\theta_0 \equiv (\alpha_0, \beta_0)$.

(ii) Propose a method of moments estimator of the parameter $q(\theta)$; i.e. find some $h(X)$ such that $E_\theta(h(X))$ is $q(\theta)$ and use $n^{-1}\sum_{i=1}^n h(X_i)$ as an estimate of $q(\theta)$. Compute $\mathrm{Var}_\theta(h(X_1))$ and determine the asymptotic distribution of your estimate.

(4) (a) Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Consider testing the null hypothesis $H_0 : \mu = \mu_0, \sigma = \sigma_0$. Show that the likelihood ratio statistic can be written as:

$$2\,\log\,\lambda_n = \frac{n(\overline{X}-\mu_0)^2}{\sigma_0^2} + n\,\log\,\frac{\sigma_0^2}{\hat{\sigma}^2} + n\,\left[\frac{\hat{\sigma}^2}{\sigma_0^2} - 1\right]\,.$$

Hence, deduce that $2\,\log\,\lambda_n$ converges under H_0 to a χ^2_2. Here $\hat{\sigma}^2$ is the usual MLE for σ^2.

(b) Consider the same data and model as in (a), but now the goal is to test $H_0 : \mu = 0$ versus $\mu > 0$. Derive the likelihood ratio test for this problem. What is the asymptotic distribution of the likelihood ratio statistic under H_0?

(5) Consider a model for the joint distribution of two random variables Y and Z in which Z has a Bernoulli distribution with success probability $\eta \in [0,1]$, and the conditional distribution for Y given $Z = z$ is exponential with failure rate $\lambda\, e^{\gamma\, z}$. Then verify that (Y, Z) has joint density:

$$f_\theta(y,z) = \lambda\, e^{\gamma\, z}\,\exp(-\lambda\, e^{\gamma\, z}\, y)\,\eta^z\,(1-\eta)^{1-z},\ z \in \{0,1\}, y > 0\,,$$

where $\theta = (\lambda, \gamma, \eta)$. This is a parametric version of the Cox proportional hazards model and γ, the regression parameter is of primary interest. Let $\{Y_i, Z_i\}_{i=1}^n$ be i.i.d. observations from

this model.

(a) Discuss the computation of the M.L.E's of (λ, γ, η) in this model. Explicitly determine the limit distributions of the M.L.E's.

(b) Is the limit distribution of $(\hat{\lambda}, \hat{\gamma})$, these being MLE's, affected by knowledge of the Bernoulli parameter η? Explain your answer.

(c) Define $\nu(\theta) = P_\theta(Y \geq y_0 \mid Z = 1)$, where y_0 is a fixed positive number. Consider the estimation of ν using the estimator

$$\hat{\nu}_1 = \frac{n^{-1} \sum_{i=1}^n 1\{Y_i \geq y_0, Z_i = 1\}}{n^{-1} \sum_{i=1}^n 1\{Z_i = 1\}} .$$

Why is this a reasonable estimator of ν? Compute its asymptotic distribution. How does $\hat{\nu}_1$ compare to $\hat{\nu}_2 = \nu(\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ?