FINAL REVIEW PROBLEMS 3

1. The purpose of this problem is to: (i) gain experience working with conditional probability by deriving Bayes's Theorem; (ii) develop a striking similarity between an approximation to Bayesian inference and the limit distribution for the maximum-likelihood estimator in Note 13; (iii) sharpen our interpretation of a frequentist confidence interval by comparing it to a conditional (posterior) probability interval; (iv) examine the relevance of the Note 13 theory for decision making by making a comparison with Bayesian (really Savage's) decision theory.

Let Y denote a random vector with distribution conditional on $Z = z$ equal to $\mathcal{P}_z$. (Below we shall specialize to the case of random sampling, with $Y = (Y_1, \ldots, Y_n)$, $Z = (Z_1, \ldots, Z_n)$, and (Y_i, Z_i) i.i.d. for $i = 1, \ldots, n$. For now, it is convenient to let (Y, Z) denote the complete observation, without insisting that it is the result of a random sample.) Suppose that we have specified a family of conditional distributions for Y conditional on $Z = z$ such that the (population) conditional distribution $\mathcal{P}_z$ is in this family. The family of conditional distributions is given by a family of conditional density functions, indexed by a parameter θ which takes on values in a parameter space Θ, which is a subset of $\mathcal{R}^K$. The densities $f(y \,|\, z, \theta)$ are with respect to a single measure m. The assumption that the family contains the population distribution means that there is some value $\theta^* \in \Theta$ such that

$$\Pr\{Y \in A \,|\, Z = z, \mathcal{P}_z\} = \mathcal{P}_z(A) = \int_A f(y \,|\, z, \theta^*)\, dm(y). \tag{1}$$

(In the continuous case, m is Lebesgue measure and we can replace $dm(y)$ by dy; in the discrete case, m is counting measure, and we can replace the integral by a sum. With random sampling, we would have

$$f(y \,|\, z, \theta) = \prod_{i=1}^{n} f(y_i \,|\, z_i, \theta),$$

and it might be wise to change the notation to $f^{(n)}(y \,|\, z, \theta)$, which is the likelihood function for the full sample, to avoid confusing it with the likelihood function for a single observation.)

Suppose that we introduce a distribution on the parameter space Θ. This is a personal, subjective distribution that represents uncertainty about the value of θ^*. This distribution is conditional on $Z = z$, but we shall consider the case where it does not depend on z:

$$\Pr\{\theta^* \in B \,|\, Z = z\} = \int_B \pi(\theta)\, d\theta.$$

This (prior) distribution has density π with respect to Lebesgue measure on $\mathcal{R}^K$.

Multiplying the conditional density $f(y \mid z, \theta)$ by the marginal density $\pi(\theta \mid z) = \pi(\theta)$ gives the joint density $f(y, \theta \mid z)$ (with all densities conditional on z):

$$\Pr\{Y \in A, \theta^* \in B \mid Z = z\} = \int_A \int_B f(y \mid z, \theta)\pi(\theta)\, dm(y)\, d\theta.$$

The joint density $f(y, \theta \mid z)$ can be factored the other way, into a product of the conditional density $\pi(\theta \mid z, y)$ and the marginal density $f(y \mid z)$ (with all densities still conditional on z):

$$\Pr\{Y \in A, \theta^* \in B \mid Z = z\} = \int_A \int_B \pi(\theta \mid z, y) f(y \mid z)\, dm(y)\, d\theta.$$

So we have

$$f(y, \theta \mid z) = f(y \mid z, \theta)\pi(\theta) = \pi(\theta \mid z, y) f(y \mid z),$$

which gives Bayes's Theorem:

$$\pi(\theta \mid z, y) = f(y \mid z, \theta)\pi(\theta)/f(y \mid z).$$

A convenient shorthand is

$$\pi(\theta \mid z, y) \propto f(y \mid z, \theta)\pi(\theta),$$

where it is understood that we can recover the normalizing constant because the density integrates to 1:

$$\int_\Theta \pi(\theta \mid z, y)\, d\theta = c \int f(y \mid z, \theta)\pi(\theta)\, d\theta = 1$$
$$\Rightarrow \quad c = 1 \Big/ \int f(y \mid z, \theta)\pi(\theta)\, d\theta = 1/f(y \mid z).$$

(c is a "constant" when we condition on the sample value (z, y).)

Let (z, y) denote the sample realized value of (Z, Y). Suppose that the log-likelihood function, regarded as a function of θ with (z, y) fixed at the sample value, is well approximated by the quadratic part of a Taylor series expansion around the maximum-likelihood estimate $\hat{\theta}$:

$$L(\theta) = \log f(y \mid z, \theta) \approx L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})' H(\hat{\theta})(\theta - \hat{\theta}),$$

where $H(\theta)$ is the Hessian matrix for the log-likelihood function:

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}$$

(and the linear term $[\partial L(\hat{\theta})/\partial\theta'](\theta - \hat{\theta})$ drops out because of the first-order condition for the maximum-likelihood estimate $\hat{\theta}$). Suppose that the likelihood function $f(y \mid z, \theta)$ is quite concentrated around $\hat{\theta}$. Then we have

$$\pi(\theta) \approx \pi(\hat{\theta})$$

over the range of values for which the likelihood is not essentially equal to 0. So outside of a set B the likelihood is essentially 0, and the set B is small enough so that $\pi(\theta)$ is approximately constant for $\theta \in B$. Then

$$f(y \mid z, \theta)\pi(\theta) \approx f(y \mid z, \theta)\pi(\hat{\theta})$$

for all $\theta \in \Theta$. A quadratic log likelihood combines with a prior that is dominated by the data to give the following approximation:

$$\pi(\theta \mid z, y) \propto \exp[-\frac{1}{2}(\theta - \hat{\theta})'(-H(\hat{\theta}))(\theta - \hat{\theta})].$$

(The proportionality "constant" includes $\exp[L(\hat{\theta})]$ and $\pi(\hat{\theta})$.) This is the density of a multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix equal to $-[H(\hat{\theta})]^{-1}$:

$$\theta^* \mid z, y \overset{a}{\sim} \mathcal{N}(\hat{\theta}, -[H(\hat{\theta})]^{-1}).$$

(A precise statement of this result is in the Bernstein-von Mises Theorem: *A Course in Large Sample Theory*, Thomas Ferguson, Chapman & Hall, 1996, Chapter 21; *Asymptotic Statistics*, A.W. van der Vaart, Cambridge University Press, 1998, Chapter 10.)

If we are interested in a linear combination $l'\theta^*$, then this result can be used to provide an approximate conditional (posterior) .95 interval:

$$\Pr\{l'\theta^* \in [l'\hat{\theta} - 1.96 \cdot \text{SE}, l'\hat{\theta} + 1.96 \cdot \text{SE}] \mid z, y\} \approx .95, \tag{2}$$

where

$$\text{SE} = (-l'[H(\hat{\theta})]^{-1}l)^{1/2}.$$

(a) Suppose that we have random sampling, so that (Y_i, Z_i) i.i.d. for $i = 1, \ldots, n$. Then we can apply the Note 13 results to obtain a limit distribution for $\sqrt{n}(\hat{\theta} - \theta^*)$, form an asymptotic pivot, and obtain an approximate .95 confidence interval for $l'\theta^*$. Is there a way to do this so that the interval we obtain is numerically equal to the approximate Bayesian .95 interval in (2)?

(b) Provide a brief discussion of how the interpretation of the frequentist .95 confidence interval differs from that of the Bayesian .95 posterior interval.

(c) Totrep needs to make a decision. (Trade-Off Talking Rational Economic Person, from *Notes on the Theory of Choice* by David Kreps, Westview Press, 1988.) There is a profit function, $g(a, \theta^*)$, which depends upon Totrep's action a and on the value of θ^* in (1). Totrep does not know θ^* but does have access to the data (Y_i, Z_i) for $i = 1, \ldots, n$ and would like to maximize expected profit. How might Totrep use the frequentist theory for maximum likelihood developed in Note 13? How might Totrep use the Bayesian approximation developed above?

2. Consider the panel probit model in Note 13:

$$\tilde{Y}_{it}^* = X_{it}'\alpha + \tilde{V}_i + \tilde{\epsilon}_{it},$$

with $\tilde{V}_i$, $\tilde{\epsilon}_{i1}$, ..., $\tilde{\epsilon}_{iT}$ mutually independent, $\tilde{V}_{it} \sim \mathcal{N}(0, \tilde{\sigma}_v^2)$, and $\tilde{\epsilon}_{it} \sim \mathcal{N}(0,1)$. (We have divided the original latent variable Y_{it}^* by σ_ϵ.) We observe

$$Y_{it} = \begin{cases} 1, & \text{if } \tilde{Y}_{it}^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

(a) Show that

$$\Pr(Y_{it} = 1 \,|\, X_{it}) = \Phi(X_{it}'\delta)$$

and provide a formula for δ.

(b) Show that there is a certain predictive effect that can be evaluated just using δ, without needing α and $\tilde{\sigma}_v$.

(c) Continue to assume random sampling on i. Try to generalize the latent variable model as much as you can, but keeping the implication that for some value of δ:

$$\Pr(Y_{it} = 1 \,|\, X_{it}) = \Phi(X_{it}'\delta) \qquad (t = 1, \ldots, T). \tag{$*$}$$

(The formula connecting δ to α and $\tilde{\sigma}_v$ will no longer hold.)

(d) Assume only that there is random sampling on i and that there is a value for δ such that $(*)$ holds. Consider applying a maximum-likelihood probit program to the cross-section observations for period t:

$$\hat{\delta}_t = \arg\max_a \sum_{i=1}^N \Big(Y_{it} \log[\Phi(X_{it}'a)] + (1 - Y_{it}) \log[1 - \Phi(X_{it}'a)] \Big).$$

Do this separately for $t = 1, \ldots, T$. Explain why each of these $\hat{\delta}_t$ is a consistent (as $N \to \infty$) estimator for δ.

(e) Stack the estimates in (d) into $\hat{\gamma}$ with $\hat{\gamma} \xrightarrow{p} \gamma$:

$$\hat{\gamma} = \begin{pmatrix} \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_T \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \vdots \\ \delta \end{pmatrix}.$$

Show that there is a moment function ψ that satisfies the key condition and

$$\sum_{i=1}^N \psi((Y_i, X_i), \hat{\gamma}) = 0.$$

Conclude that

$$\sqrt{N}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

and provide a formula for Λ.

(f) Use the minimum-distance framework in Note 12 to combine the T separate estimates of δ and provide a confidence interval for $l'\delta$ (a linear combination of the coefficients). Provide enough detail so that a research assistant could program the procedure in Matlab. The research assistant is happy to take derivatives, but you need to be clear on what the function is and which derivatives are needed.

3. Suppose that Z_1 and Z_2 are binary random variables: Z_1 takes on only the values 0 and 1, Z_2 takes on only the values 0 and 1, and

$$0 < \text{Prob}\{Z_1 = l, Z_2 = m\} < 1 \qquad (l, m = 0, 1).$$

Consider the (population) linear predictor of Y given $1, Z_1, Z_2, Z_1 \cdot Z_2$:

$$E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 \cdot Z_2.$$

(a) Does

$$E(Y \mid Z_1, Z_2) = E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2)?$$

Explain.

(b) Suppose that data (Y_i, Z_{i1}, Z_{i2}) are available from a random sample of $i = 1, \ldots, n$ individuals. The following four sample means have been tabulated:

$$\bar{Y}_{00}, \quad \bar{Y}_{01}, \quad \bar{Y}_{10}, \quad \bar{Y}_{11},$$

where

$$\bar{Y}_{lm} = \frac{\sum_{i=1}^n Y_i 1(Z_{i1} = l, Z_{i2} = m)}{\sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m)} \qquad (l, m = 0, 1).$$

($1(B)$ is the indicator function that equals 1 if the event B occurs and equals 0 otherwise.) Use these means to provide an estimate of β_3. Is this a consistent estimator of β_3 as $n \to \infty$? Explain.

(c) The following four sample variances have been tabulated:

$$S_{00}, S_{01}, S_{10}, S_{11},$$

where

$$S_{lm} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_{lm})^2 1(Z_{i1} = l, Z_{i2} = m)}{\sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m)},$$

and the sample sizes

$$n_{00}, n_{01}, n_{10}, n_{11},$$

where

$$n_{lm} = \sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m) \qquad (l, m = 0, 1).$$

Let $\hat{\beta}_3$ denote your estimator for β_3 in (b). Use $\hat{\beta}_3$ along with these sample variances and sample sizes to provide a (approximate) .95 confidence interval for β_3. Does the probability that the interval covers β_3 converge to .95 as $n \to \infty$? Explain. Do not make additional assumptions such as homoskedasticity.