Due date: Thursday, Nov. 29

You may disregard measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, “all functions” vs. “all measurable functions,” etc. (unless the problem is explicitly asking about such issues).

In addition, you should assume that all necessary regularity conditions are in effect for the asymptotic tests discussed here (smoothness, consistency of MLE, etc.).

1. Logistic regression with random $X$

Consider a univariate logistic regression model where we observe $n$ i.i.d. pairs $(X_i, Y_i) \in \mathbb{R} \times \{0, 1\}$. The covariate is random with a known distribution, $X_i \overset{i.i.d.}{\sim} U[-1, 1]$, and

$$P_{\alpha, \beta}(Y_i = 1 \mid X_i = x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}.$$

(a) Show that the maximum likelihood estimator for $(\alpha, \beta)$ solves

$$\sum Y_i = \sum \pi_i(\hat{\alpha}_n, \hat{\beta}_n)$$

$$\sum Y_i X_i = \sum \pi_i(\hat{\alpha}_n, \hat{\beta}_n) X_i,$$

where $\pi_i(\alpha, \beta) = e^{\alpha + \beta X_i}/(1 + e^{\alpha + \beta X_i})$.

(b) For $\alpha = 0$, $\beta = 4$, calculate the Fisher information for a single pair $(X_i, Y_i)$; give it as an integral and also calculate it numerically (you do not need to analytically evaluate the integral). Note your answer should not depend on $X_i$, which is a random variable in this problem.

(c) Find the asymptotic distribution of the MLE $(\hat{\alpha}, \hat{\beta})$, giving the asymptotic variance numerically. You may just assume that our theorems from class apply without showing it (they do).

(d) For each of a few different $n$ values:

(i) Generate a large number (e.g. 1000) of data sets of size $n$, and for each one compute the MLE $(\hat{\alpha}, \hat{\beta})$ (you can use statistical software to compute the MLE, e.g. the glm function in R).

(ii) Plot histograms of $\hat{\alpha}$ and $\hat{\beta}$ (if you use R, I recommend setting freq=FALSE to get a density histogram instead of a frequency histogram).

(iii) Overlay the Gaussian curve based on the approximate distribution from part (c) (you can use the dnorm function in R). About how big does $n$ need to be for the normal approximation to be pretty good?

(e) Repeat parts (c) and (d) for $\alpha = -6$ and $\beta = 4$.

2. MLE consistency for concave log-likelihoods

Assume $X_1, X_2, \ldots \overset{i.i.d.}{\sim} p_{\theta_0}(x)$ for some dominated family $\mathcal{P} = \{p_\theta : \theta \in \Theta = (a, b)\}$, with $-\infty \leq a < b \leq \infty$. Assume that $\ell_1(\theta; X_1) = \log p_\theta(X_1)$ is almost surely concave in $\theta$, and that $E_{\theta_0} |\ell_1(\theta; X_1)| < \infty$ for all $\theta \in \Theta$. Additionally, assume the model is identifiable: that $p_\theta \neq p_{\theta_0}$ for all $\theta \neq \theta_0$. 

Prove that the MLE is consistent: if \( \hat{\theta}_n \in \arg\max \ell_n(\theta; X_1, \ldots, X_n) \) then \( \hat{\theta}_n \overset{p}{\rightarrow} \theta_0 \) (we can arbitrarily define \( \hat{\theta}_n = \pm \infty \) if no maximizer exists).

(Hint: This is not simply a modification of the proof from class. Start by showing that, if

\[
\ell_n(\theta) > \max \{ \ell_n(\theta - \varepsilon), \ell_n(\theta + \varepsilon) \},
\]

then \( \hat{\theta}_n \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon) \).

3. Probabilistic big-O notation

Let \( X_1, X_2, \ldots \) denote a sequence of real-valued random variables \((|X_n| < \infty \text{ almost surely for each } n)\). We say the sequence is bounded in probability (or sometimes tight) if for every \( \varepsilon > 0 \) there exists a constant \( M_\varepsilon > 0 \) for which

\[
P(|X_n / g(n)| > M_\varepsilon) < \varepsilon, \quad \forall n.
\]

Informally, there is “no mass escaping to infinity” as \( n \) grows. Like deterministic big-O notation, these symbols are extremely useful for making rigorous asymptotic arguments in a clean and intuitive way.

For a real-valued function \( g(n) \), we say \( X_n = o_p(g(n)) \) if \( X_n / g(n) \overset{p}{\rightarrow} 0 \) as \( n \to \infty \), and \( X_n = O_p(g(n)) \) if the sequence \((X_n / g(n))_{n \geq 1}\) is bounded in probability. More generally, if \( X_n \in \mathcal{X} \), a generic sample space with norm \( \| \cdot \| \), we can say \( X_n = O_p(g(n)) \) or \( o_p(g(n)) \) if \( \|X_n\| = O_p(g(n)) \) or \( o_p(g(n)) \), respectively.

Prove the following facts for \( X_n \in \mathbb{R} \):

(a) If \( X_n \Rightarrow X \) for any random variable \( X \), then \( X_n = O_p(1) \).
(b) If \( X_n = o_p(g(n)) \) then \( X_n = O_p(g(n)) \).
(c) If \( X_n = O_p(f(n)) \) and \( Y_n = o_p(g(n)) \), then \( X_nY_n = o_p(f(n)g(n)) \).
(d) If \( X_n = O_p(f(n)) \) and \( Y_n = O_p(g(n)) \), then \( X_nY_n = O_p(f(n)g(n)) \).

4. One-step estimator

We say a sequence of estimators \( \hat{\theta}_n \) is \( g(n) \)-consistent for \( \theta \) if \( g(n)(\hat{\theta}_n - \theta) = O_p(1) \) (the MLE is typically \( \sqrt{n} \)-consistent).

We have shown that, when the MLE exists, it is usually asymptotically efficient (achieves the Cramer-Rao lower bound). But there are some problems where the MLE is intractable to compute, or does not exist, or exists but is not consistent.

In such a problem, suppose we can compute some other estimator \( \tilde{\theta}_n \), which is consistent but possibly inefficient. We could try to improve \( \tilde{\theta}_n \) by taking one Newton step on the log-likelihood to get a new estimator, called a one-step estimator:

\[
\hat{\theta}_n = \tilde{\theta}_n - \left( \nabla^2 \ell(\tilde{\theta}_n; X) \right)^{-1} \nabla \ell(\tilde{\theta}_n; X).
\]

Perhaps surprisingly, \( \hat{\theta}_n \) is asymptotically efficient (asymptotically just as good as the MLE would usually be).

Assume \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_{\theta_0}(x) \) from a dominated family with \( \Theta \subseteq \mathbb{R} \), and

- The true value \( \theta_0 \) is in the interior of \( \Theta \)
- \( p_{\theta_0}(x) > 0 \) for all \( x, \theta \)
- \( \ell(\theta; X) \) has two continuous derivatives with respect to \( \theta \)
- \( \mathbb{E}_{\theta_0} \ell'_1(\theta_0; X_1) = 0 \) and \( \text{Var}_{\theta_0} \ell'_1(\theta_0; X_1) = -\mathbb{E}_{\theta_0} \ell''_1(\theta_0; X_1) = J(\theta_0) \in (0, \infty) \),

2
• There exists \( \varepsilon > 0 \) for which

\[
E_{\theta_0} \left[ \sup_{\theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]} |\ell'_{\chi}(\theta; X_1)| \right] < \infty
\]

Prove that if \( \tilde{\theta}_n \) is \( \sqrt{n} \)-consistent, then \( \hat{\theta}_n \) is asymptotically normal and asymptotically efficient:

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, J_1(\theta_0)^{-1})
\]

(Not graded, just for fun: \( \tilde{\theta}_n \) need not actually be \( \sqrt{n} \)-consistent. At what rate do we really need \( \tilde{\theta}_n \) to be converging? Feel free to make extra regularity assumptions as necessary.)

5. Superefficiency Revisited

Recall the superefficient estimator for \( \theta \) given \( X_1, \ldots, X_n \) i.i.d. \( \sim N(\theta, 1) \), from the last problem set:

\[
\delta_n = \begin{cases} 
\bar{X}_n & |\bar{X}_n| > n^{-1/4} \\
0 & |\bar{X}_n| \leq n^{-1/4}
\end{cases}, \quad \text{where} \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]

Show that, pointwise in \( \theta \), as \( n \to \infty \),

\[
n \text{MSE}(\delta_n; \theta) \to 1\{\theta \neq 0\},
\]

but that the convergence is not uniform in \( \theta \); in fact,

\[
\sup_{\theta \in \mathbb{R}} n \text{MSE}(\delta_n; \theta) \to \infty
\]

(Hint: consider the sequence \( \theta_n = n^{-1/4} \))