Problem Set 4 (Lectures 7-8)

Problem 1 (Cramer-Rao Lower Bound for a scalar parameter, 50 points). In class we showed that the OLS estimator achieves the smallest variance among all unbiased estimators. The proof in the notes looks like a bunch of algebra, but it is actually based on a more general result we did not cover: the Cramer-Rao Lower Bound.

Let $\widehat{\theta}(x)$ be the estimator of a real-valued parameter θ in the statistical model with pd.f. $f(x,\theta)$. Unfortunately, there is no theorem that says that ML estimators will be unbiased or that the ML estimators will achieve the lowest possible variance among unbiased estimators (if by chance they happen to be unbiased).

Instead of establishing the optimality of ML estimators, we will show that in parametric models we can provide a lower bound on the variance of any given estimator. The lower bound will be given as a function of the bias. The bound will be important, as in large samples, there are theorems that guarantee that ML estimators (which are asymptotically unbiased) will approach this bound.

Here is what I would like you to show:

Proposition 1 (Cramér-Rao Bound)**.** *Suppose that the estimator $\widehat{\theta}$ and the statistical model satisfy:*

$$\int_{\mathbb{R}} \Big[\widehat{\theta}(x)\frac{\partial}{\partial\theta}f(x,\theta)\Big]dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} \Big[\widehat{\theta}(x)f(x,\theta)\Big]dx \tag{0.1}$$

and

$$\int_{\mathbb{R}} \Big[\frac{\partial}{\partial\theta}f(x,\theta)\Big]dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} \Big[f(x,\theta)\Big]dx = 0, \tag{0.2}$$

(both of which require that we can change the order in which we take integrals and derivatives). If these conditions are satistfied:

$$Var_{P_\theta}\Big[\widehat{\theta}(x)\Big] \geq \Big[\frac{\partial}{\partial\theta}\mathbb{E}_{P_\theta}[\widehat{\theta}(x)]\Big]^2 \Big/ Var_{P_\theta}\Big[S_\theta(x)\Big],$$

where

$$S_\theta(x) \equiv \frac{\partial}{d\theta}\ln f(x,\theta)$$

is called the score *of the statistical model $\{f(x,\theta)\}_{\theta\in\Theta}$ and $Var_{P_\theta}\Big[S_\theta(x)\Big]$ is called the Fisher information of the statistical model at θ.*

I will help you a bit with the proof. Just fill in the blanks (if you can give a different proof of this result—which probably you can do, based on the lecture notes—go for it!)

Proof. (**5 points each**). The covariance between any estimator $\widehat{\theta}$ and the score (which is a random variable) is

$$\begin{aligned}
\mathbb{E}_\theta\Big[\widehat{\theta}(x)S_\theta(x)\Big] &= \int_{\mathbb{R}} \widehat{\theta}(x)\frac{\partial}{d\theta}\ln f(x,\theta) f(x,\theta)dx \\
&= \int_{\mathbb{R}} \widehat{\theta}(x)\frac{\partial}{d\theta} f(x,\theta)dx \\
&= \boxed{} \\
&\quad \text{(where we have used Equation 0.1)} \\
&= \frac{\partial}{d\theta}\boxed{}.
\end{aligned}$$

where $\mathbb{E}_\theta[\widehat{\theta}(x)]$ is the bias of the estimator $\widehat{\theta}(x)$ at θ. Assumption 0.2 implies

$$\mathbb{E}_\theta[S_\theta(x)] = \boxed{},$$

which implies

$$\mathbb{E}_\theta\Big[\widehat{\theta}(x)S_\theta(x)\Big] = \mathbb{E}_\theta\Big[\Big(\widehat{\theta}(x) - \mathbb{E}_\theta[\widehat{\theta}(x)]\Big)S_\theta(x)\Big]$$

Hence, by the Cauchy-Scharwz inequality:[1]

$$\begin{aligned}
\mathbb{E}_\theta\Big[\widehat{\theta}(x)S_\theta(x)\Big]^2 &\leq \boxed{}\boxed{} \\
&= \mathrm{Var}_\theta\Big[\widehat{\theta}(x)\Big]\boxed{}
\end{aligned}$$

Therefore,

$$\mathrm{Var}_\theta\Big[\widehat{\theta}(x)\Big] \geq \frac{\partial}{d\theta}\boxed{}.$$

□

Corollary (**15 Points**) Let $\widehat{\theta}$ be any *unbiased* estimator for the mean parameter θ in the model

[1] For any two random variables X and Y:

$$\mathbb{E}_{\mathbb{P}}[XY] \leq \mathbb{E}_{\mathbb{P}}[X^2]^{1/2}\mathbb{E}_{\mathbb{P}}[Y^2]^{1/2}$$

See pg. 24 Durrett (2010).

for the data $(x_1, \ldots, x_n)$, where $x_i \sim \mathcal{N}(\theta, \sigma^2)$, i.i.d. and σ^2 is known. Show that the estimator $\widehat{\theta}(x_1, \ldots, x_n) = (1/n)\sum_{i=1}^n x_i$ is the ML estimator and it achieves the smallest mean squared error relative to all unbiased estimators.

Problem 2 (Score and Fisher Information Matrix in the Linear Regression model, 15 points) Derive the score and the Fisher Information matrix in the Normal Linear Regression model. How does the variance of the OLS estimator relates to the Fisher Information matrix? Would you expect this relation to hold in other parametric models?

Problem 3 (Variance of the Ridge Estimator, 15 points) In class we showed that the variance of the Ridge estimator is

$$\mathbb{V}_\beta(\widehat{\beta}_{\text{Ridge}}) = \sigma^2(X'X + \lambda \mathbb{I}_k)^{-1} X'X (X'X + \lambda \mathbb{I}_k)^{-1}$$

Prove or disprove the following statement: when $k < n$, the trace of this variance has to be smaller than that of the ML/OLS estimator.

Problem 4 (Expression for the Ridge Estimator, 30 points) In class we derived the expression for the posterior distribution of $\beta|Y$ using Bayes' Theorem and the Gaussian p.d.f.s. An alternative derivation would be to write down the joint distribution of $(\beta', Y')'$ and use the formula for the conditional mean and variance. For example, see the entry on conditional distributions in wiki (click here). Convince yourself that the formulae are the same. You might need to use the Woodbury Identity Formula a couple of times to establish the connection.

References

DURRETT, R. (2010): *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 4th ed.