Problem Set 4 (Lectures 7-8)

Problem 1 (Cramer-Rao Lower Bound for a scalar parameter, 40 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 always 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 (meaning, statistical models where θ is finite-dimensional) we can provide a lower bound on the variance of any given estimator. The lower bound we present depends on the bias. The bound will be important, as in large samples, there are theorems that guarantee that ML estimators (which are asymptotically unbiased) will eventually approach the lower bound on the variance.

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

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

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,

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

□

[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).

Corollary (**5 Points**) Let $\widehat{\theta}$ be any *unbiased* estimator for the mean parameter θ in the model 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 sample mean estimator $\widehat{\theta}(x_1, \ldots, x_n) = (1/n)\sum_{i=1}^n x_i$ maximizes the likelihood and it achieves the smallest mean squared error relative to all unbiased estimators (HINT: Use the Cramer-Rao Lower bound) .

Problem 2 (Maximum Likelihood in the Linear Regression model, 25 points) We have shown that the OLS estimator $\widehat{\beta}_{\text{OLS}}$ maximizes the likelihood for the Normal Linear Regression model: $Y \sim \mathcal{N}_n(X\beta, \sigma^2\mathbb{I}_n)$ when σ^2 is known. In this exercise I would like you to treat σ^2 as an unknown parameter, and derive the ML estimators of β and σ^2. We will do this in two steps.

1. (10 points) Treating σ^2 as known, derive the score $S(x, \beta)$ and the Fisher Information matrix $\mathcal{I}(\beta)$ in the Normal Linear Regression model. Show that, at a given parameter β,

$$\widehat{\beta}_{\text{ML}} = \mathcal{I}(\beta)^{-1}S(x, \beta) + \beta \sim \mathcal{N}_k(\beta, \mathcal{I}(\beta)^{-1}).$$

   I am making this connection because one of the large sample approximations that you will show for ML estimators in the near future is that

$$\widehat{\theta}_{\text{ML}} = \mathcal{I}_n(\theta)^{-1}S_n(x_1, \ldots, x_n, \theta) + \beta \approx \mathcal{N}_k(\theta, \mathcal{I}_n(\theta)^{-1}).$$

   This means that ML estimators will eventually achieve the Cramer-Rao Lower bound.

2. (15 points) Treating σ^2 as unknown, I would like you to derive the ML estimator for both parameters β and σ^2. I would suggest you to start by deriving the score for this model (you already have one part) and solve the F.O.C. You will note that the ML estimator for β is still the same as before. OPTIONAL: Derive the Fisher information matrix for this model. What is the distribution of $\widehat{\sigma}^2_{\text{ML}}$?

Problem 3 (Bayesian Estimation in the Linear Regression model, 25 points) We have derived the ML estimators for (β, σ^2) in a Linear Regression model where both of these parameters are unknown. I will now ask you to derive the posterior mean estimators of (β, σ^2), which means we will require a prior π over (β, σ^2). Consider the following one for some $\lambda > 0$:

$$\beta|\sigma^2 \sim \mathcal{N}_k(\beta_0, \sigma^2(\lambda\mathbb{I}_k)^{-1}).$$

and

$$\sigma^2 \sim \text{Inverse-Gamma}(a_0, b_0),$$

which means that the p.d.f of σ^2 is up to a constant equal to

$$(\sigma^2)^{-a_0-1}\exp\left(-b_0/\sigma^2\right).$$

Please derive the posterior mean of β and the posterior mean of σ^2.

HINT: This exercise can be really painful (tons of algebra) if you do not approach from the right angle. I would start by computing the distribution of

$$\beta|Y,\sigma^2.$$

Note that this is a conditional posterior distribution (as I am conditioning on the data, but also on σ^2). If you condition on σ^2, you should be able to derive the posterior in exactly the same way we did in class. The posterior mean of σ^2 is slightly more difficult to derive, the idea is to derive the distribution of $Y|\sigma^2$ (yes, eliminating β out). One way to do this is to write

$$Y = X\beta + \sqrt{\sigma^2}\epsilon_1, \quad \epsilon_1 \sim \mathcal{N}_n(0, \mathbb{I}_n),$$

and

$$\beta = \beta_0 + \sqrt{\sigma^2}\epsilon_2, \quad \epsilon_2 \sim \mathcal{N}_k(0, (\lambda\mathbb{I}_k)^{-1}),$$

where ϵ_1 and ϵ_2 are independent. Using the representation above you can derive the distribution of $Y|\sigma^2$ and then proceed to derive the posterior distribution of $\sigma^2|Y$. Good luck!

Problem 4 (OLS vs. Ridge Variance, 10 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 $k < n$ and X has full-column rank, the trace of this variance has to be smaller than $\sigma^2\text{tr}((X'X)^{-1})$. HINT: I am not looking for an algebraic proof of this result. I want you to exploit the properties of the Ridge estimator as a Bayesian estimator.

Problem 5 (Optional: Expression for the Ridge Estimator) 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.