DS-GA 3001.009 Applied Statistics: Homework #6 Solutions

Due on Thursday, November 16, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. Revisit the example of bivariate Gaussian location model we covered in class:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \dots, \begin{bmatrix} x_n \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta_0 \\ \eta_0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),$$

where $\rho \in [-1, 1]$ is known.

(a) Recall that the estimating equation based on the score for θ_0 is

$$\frac{1}{n} \sum_{i=1}^{n} \left[x_i - \widehat{\theta} - \rho(y_i - \widehat{\eta}) \right] = 0.$$

If $\widehat{\eta} = \eta_0$ is the true nuisance, from the above equation, determine the probability distribution of $\widehat{\theta} - \theta_0$ which only depends on (n, ρ).

(b) Repeat (a) if $\widehat{\eta} = \eta_0 + \varepsilon$ with a fixed constant ε. Your answer should depend on (n, ρ, ε).

(c) Now consider the efficient score equation

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\theta}) = 0.$$

Write out the probability distribution of $\widehat{\theta} - \theta_0$. How does $\mathbb{E}[(\widehat{\theta} - \theta_0)^2]$ compare with (a) and (b)?

Solution:

(a) The estimating equation gives that

$$\widehat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - \theta_0) - \rho(y_i - \eta_0) \right].$$

Each term in the average is distributed as $\mathcal{N}(0, \sigma^2)$ with

$$\sigma^2 = \mathsf{Var}(x_i) + \rho^2 \mathsf{Var}(y_i) - 2\rho \mathsf{Cov}(x_i, y_i) = 1 - \rho^2,$$

and therefore $\widehat{\theta} - \theta_0 \sim \mathcal{N}(0, (1 - \rho^2)/n)$.

(b) If $\widehat{\eta} = \eta_0 + \varepsilon$, then

$$\widehat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - \theta_0) - \rho(y_i - \eta_0) \right] - \rho\varepsilon.$$

By the result in (a), we have $\widehat{\theta} - \theta_0 \sim \mathcal{N}(-\rho\varepsilon, (1 - \rho^2)/n)$.

(c) The new estimating equation gives $\widehat{\theta} - \theta_0 = n^{-1} \sum_{i=1}^n (x_i - \theta_0) \sim \mathcal{N}(0, 1/n)$. We compute that $\mathbb{E}[(\widehat{\theta} - \theta_0)^2] = 1/n$, whereas the results in (a) and (b) are $(1-\rho^2)/n$ and $(1-\rho^2)/n + \rho^2\varepsilon^2$, respectively. Therefore, the MSE of $\widehat{\theta}$ from the efficient score equation is higher than the counterpart with known nuisance η_0 in (a), while is lower than the result of (b) as long as $\varepsilon^2 > 1/n$.

2. In this problem, we consider a simple error-in-variable model

$$\begin{aligned} y &= \theta_0 z_0 + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0,1), \\ x &= z_0 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

Here the observables are (x, y), the target parameter is θ_0, the nuisance parameter is z_0, and the errors $(\varepsilon_1, \varepsilon_2)$ are independent. The parameter σ is known.

(a) Write out the log-likelihood of (x, y) given (θ_0, z_0).

(b) Compute the score functions $s^{\theta}_{(\theta_0,z_0)}(x, y)$ and $s^{z}_{(\theta_0,z_0)}(x, y)$.

(c) Compute the efficient score function $s^{\text{eff}}_{(\theta_0,z_0)}(x, y)$ for θ_0.

(d) Now suppose that we have n i.i.d. observations $(x_1, y_1), \cdots, (x_n, y_n)$, as well as a nuisance estimate $\widehat{z}$. Find the estimator $\widehat{\theta}$ based on the efficient score function.

Solution:

(a) The log-likelihood is

$$\ell_{\theta_0,z_0}(x, y) = -\frac{(x - z_0)^2}{2\sigma^2} - \frac{(y - \theta_0 z_0)^2}{2} + \text{const}.$$

(b) The score functions are

$$\begin{aligned} s^{\theta}_{(\theta_0,z_0)}(x, y) &= \left.\frac{\partial \ell_{\theta,z}(x, y)}{\partial \theta}\right|_{(\theta,z)=(\theta_0,z_0)} = z_0(y - \theta_0 z_0), \\ s^{z}_{(\theta_0,z_0)}(x, y) &= \left.\frac{\partial \ell_{\theta,z}(x, y)}{\partial z}\right|_{(\theta,z)=(\theta_0,z_0)} = \frac{z_0(x - z_0)}{\sigma^2} + \theta_0(y - \theta_0 z_0). \end{aligned}$$

(c) We can compute that

$$\begin{aligned} \mathbb{E}[s^{\theta}_{(\theta_0,z_0)}(x, y) s^{z}_{(\theta_0,z_0)}(x, y)] &= z_0\theta_0, \\ \mathbb{E}[(s^{z}_{(\theta_0,z_0)}(x, y))^2] &= \frac{z_0^2}{\sigma^2} + \theta_0^2. \end{aligned}$$

Consequently, the efficient score function is

$$\begin{aligned} s^{\text{eff}}_{(\theta_0,z_0)}(x, y) &= s^{\theta}_{(\theta_0,z_0)}(x, y) - \frac{\mathbb{E}[s^{\theta}_{(\theta_0,z_0)}(x, y) s^{z}_{(\theta_0,z_0)}(x, y)]}{\mathbb{E}[s^{z}_{(\theta_0,z_0)}(x, y)^2]} s^{z}_{(\theta_0,z_0)}(x, y) \\ &= \frac{z_0^2}{z_0^2 + \theta_0^2\sigma^2} \left[z_0(y - \theta_0 z_0) - \theta_0(x - z_0) \right]. \end{aligned}$$

(d) Based on the efficient score function, $\widehat{\theta}$ is the solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n} s^{\text{eff}}_{(\widehat{\theta},\widehat{z})}(x_i, y_i) = \frac{\widehat{z}^2}{\widehat{z}^2 + \widehat{\theta}^2\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{z}(y_i - \widehat{\theta}\widehat{z}) - \widehat{\theta}(x_i - \widehat{z})\right].$$

It is then easy to compute that

$$\widehat{\theta} = \frac{\widehat{z}\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(\widehat{z}^2 + x_i - \widehat{z})}.$$

3. Coding: we will implement Stein's semiparametric estimator for the symmetric location model $y_1, \cdots, y_n \sim f(y - \theta_0)$, where in our experiment $f(y) = e^{-|y|}/2$ is the Laplace density. We will experiment on three estimators of θ_0:

   - the sample mean;
   - the MLE with the knowledge of f - you should derive the form of the MLE here and find it to be a very simple statistic of $(y_1, \cdots, y_n)$;
   - Stein's semiparametric estimator without the knowledge of f.

   Based on inline instructions, fill in the missing codes in `https://tinyurl.com/5zjf4bzd`. Be sure to submit a pdf with your codes, outputs, and colab link.

   Solution: see `https://tinyurl.com/mpbbb678`.