Simulation Report

Qinglong Tian

updated on April 1, 2022

Setting 1

Data Generating Mechanism

The joint distribution of $(Y, \mathbf{X}) \in \mathbb{R}^4$ from the **source** population has a multivariate normal distribution, which is given by

$$(Y, \mathbf{X})^{\mathrm{T}} \sim \mathrm{MVN}(\mu_{Y\mathbf{X}}, \Sigma_{Y\mathbf{X}}),$$

where

$$\mu_{Y\mathbf{X}} = (2, 1, 1, 1)^{\mathrm{T}}, \quad \Sigma_{Y\mathbf{X}} = \begin{bmatrix} 1.44 & 0.9 & 0.81 & 0.729 \\ 0.9 & 1.1 & 0.3 & 0.09 \\ 0.81 & 0.3 & 1.1 & 0.3 \\ 0.729 & 0.09 & 0.3 & 1.1 \end{bmatrix}.$$

The condition distribution of Y given $\mathbf{X}$ on the source population is given by

$$p_s(y|\mathbf{x}) \sim \mathrm{Norm}(\alpha_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1, \sigma^2),$$

where

$$\alpha_0 = 0.422, \quad \boldsymbol{\alpha}_1 = (0.663, 0.421, 0.494)^{\mathrm{T}}, \quad \sigma^2 = 0.142.$$

The conditional distribution of $\mathbf{X}$ given Y is given by

$$p(\mathbf{x}|y) \sim \mathrm{MVN}(\mu_{\mathbf{X}|Y}, \Sigma_{\mathbf{X}|Y}),$$

where

$$\mu_{\mathbf{X}|Y} = \begin{bmatrix} -0.25 \\ -0.125 \\ -0.0125 \end{bmatrix} + \begin{bmatrix} 0.625 \\ 0.5625 \\ 0.50625 \end{bmatrix} y, \quad \Sigma_{\mathbf{X}|Y} = \begin{bmatrix} 0.5375 & -0.20625 & -0.365625 \\ -0.20625 & 0.644375 & -0.1100625 \\ -0.365625 & -0.1100625 & 0.73094375 \end{bmatrix}.$$

The marginal distribution of Y on the target distribution is

$$p_t(y) \sim \mathrm{Norm}(\mu = 1.5, \sigma^2 = 2.25).$$

Consequently, function $\rho(y; \boldsymbol{\beta})$ is given by

$$\rho(y; \boldsymbol{\beta}) = \beta_1 y + \beta_2 y^2,$$

where

$$\beta_1 = -0.722, \quad \beta_2 = 0.125.$$

The sample size of the source distribution data is denoted by n while the sample size of the target distribution data is denoted by m. The proportion π is computed by $\pi = n/(n+m)$. The Monte Carlo sample size is $B = 2000$. In this simulation, we estimate $\boldsymbol{\beta}$ using both the true $p_s(y|\mathbf{x})$ and the fitted (but correctly specified) $\widehat{p}_s(y|\mathbf{x})$.

Simulation Results

The simulation results using the true model $p_s(y|\mathbf{x})$ are given in Table 1.

n	m	β_1				β_2			
		Mean	Bias	SE	SD	Mean	Bias	SE	SD
500	500	-0.680	0.0427	0.113	3.425	0.119	-0.00630	0.0286	1.355
500	1000	-0.717	0.00556	0.123	3.679	0.116	-0.00857	0.04436	1.529
1000	500	-0.703	0.0189	0.114	3.444	0.121	-0.00392	0.0266	1.351
1000	1000	-0.687	0.0352	0.0943	3.410	0.119	-0.00617	0.0198	1.340
1500	1500	-0.691	0.0315	0.0824	3.412	0.120	-0.00544	0.0169	1.342

Table 1: Estimation of $\boldsymbol{\beta}$ using the true model $p_s(y|\mathbf{x})$.

The simulation results using the fitted mode $\hat{p}_s(y|\mathbf{x})$ are given in Table 2.

n	m	β_1				β_2			
		Mean	Bias	SE	SD	Mean	Bias	SE	SD
500	500	-0.681	0.04173	0.112	3.420	0.119	-0.00633	0.0284	1.356
500	1000	-0.717	0.00538	0.122	3.684	0.117	-0.00846	0.0445	1.531
1000	500	-0.702	0.0204	0.112	3.445	0.121	-0.00425	0.0262	1.352
1000	1000	-0.689	0.0333	0.0954	3.414	0.119	-0.00609	0.0198	1.342
1500	1500	-0.690	0.0322	0.0820	3.414	0.119	-0.00556	0.0168	1.343

Table 2: Estimation of $\boldsymbol{\beta}$ using the fitted model $\widehat{p}_s(y|\mathbf{x})$.

Discussions

This simulation study shows that the estimation approach gives good results as the bias is generally small and the standard error is reasonable. In both tables, SE is the standard error and SD is the mean of all $B = 2000$ estimated standard deviations. It is obvious that either my derivation/programming is incorrect or the formula does not work. I checked the functions to compute $\mathrm{E}(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})$ and did not find anything. It is of interest to investigate the effect of the ratio of m/n on the Bias/SE in the following simulation.

I used optim() function for simplicity. The running time is not fast but not too terrible as each row takes at most a few hours to run on the laptop. If we want to run large number of n and m, I can write the optimization algorithm instead of using optim().

References