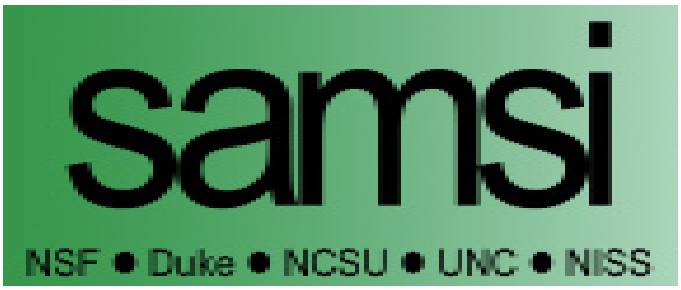

A Note on Inference in a Bivariate Normal Distribution Model

Jaya Bishwal and Edsel A. Peña

Technical Report #2009-3
December 22, 2008

This material was based upon work partially supported by the National Science Foundation under Grant DMS-0635449 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

A Note on Inference in a Bivariate Normal Distribution Model

Jaya Bishwal* Edsel A. Peña†

December 22, 2008

Abstract

Suppose observations are available on two variables Y and X and interest is on a parameter that is present in the marginal distribution of Y but not in the marginal distribution of X, and with X and Y dependent and possibly in the presence of other parameters which are nuisance. Could one gain more efficiency in the point estimation (also, in hypothesis testing and interval estimation) about the parameter of interest by using the full data (both Y and X values) instead of just the Y values? Also, how should one measure the information provided by random observables or their distributions about the parameter of interest?

We illustrate these issues using a simple bivariate normal distribution model. The ideas could have important implications in the context of multiple hypothesis testing or simultaneous estimation arising in the analysis of microarray data, or in the analysis of event time data especially those dealing with recurrent event data. It is also hoped that this note could be used for pedagogical purposes especially in the teaching of mathematical statistics concepts such as completeness, sufficiency, unbiased estimation, dependence, marginalization, efficiency, statistical (Fisher) information, correlation, and regression.

Key Words and Phrases: Completeness; correlation; Cramer-Rao Inequality; Fisher information; dependence; efficiency; marginal inference; regression; residual information; sufficiency; unbiased estimation.

1 Setting and Motivation

Consider a situation where observations are available on two random variables, V and W, which are possibly dependent, and it is of interest to make inference about a parameter vector θ which is present in the marginal distribution of V, but not in the marginal distribution of W. Furthermore, there could be a nuisance parameter vector ξ appearing in the joint distribution of (V, W). In making inference about θ, should one use the V observations only, or does one gain more efficiency by also utilizing the W observations? We examine this situation using a bivariate normal distribution model.

*J. Bishwal is Assistant Professor in the Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223. His e-mail address is J.Bishwal@uncc.edu.

†E. Peña is Professor in the Department of Statistics, University of South Carolina, Columbia, SC 29208. His e-mail address is pena@stat.sc.edu.

Let $(Y_i, X_i), i = 1, 2, \ldots, n$, be independent and identically distributed (IID) random vectors from a bivariate normal distribution with mean vector (μ, ν) and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

so that the common joint density function of the (Y_i, X_i)s is of form

$$f_{(Y,X)}(y, x|\mu, \nu, \sigma_1^2, \sigma_2^2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-Q(x, y|\mu, \nu, \sigma_1^2, \sigma_2^2, \rho)\right\}$$

with

$$Q(x, y|\mu, \nu, \sigma_1^2, \sigma_2^2, \rho) = \frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu}{\sigma_1}\right)^2 - 2\rho\left(\frac{y-\mu}{\sigma_1}\right)\left(\frac{x-\nu}{\sigma_2}\right) + \left(\frac{x-\nu}{\sigma_2}\right)^2\right].$$

The value of the parameter $\nu \in \Re$ is assumed known, and without loss of generality it could be taken to be $\nu = \nu_0 = 0$, while $\mu \in \Re$ is unknown and is the parameter of interest. The parameters in the covariance matrix Σ are $\sigma_1 > 0$, the standard deviation of the Y_is; $\sigma_2 > 0$, the standard deviation of the X_is; and $\rho \in (-1, 1)$, the correlation coefficient between Y_i and X_i.

Suppose it is of interest to estimate μ, or equivalently, test hypothesis or construct confidence intervals about μ, with the other parameters viewed as nuisance. A simplistic approach is to reason out that since μ is the mean of the Y_is and the marginal distributions of the X_is do not at all involve μ, then inference about μ should only be based on the Y_is. Indeed, since marginally $Y_i, i = 1, \ldots, n$, are IID $N(\mu, \sigma_1^2)$, then an estimator of μ is the usual sample mean

$$\delta_1 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{1}$$

which is unbiased for μ and has variance $\mathrm{Var}(\delta_1) = \sigma_1^2/n$. In fact, recall that in the case where ν is not known, the uniformly minimum variance unbiased estimator (UMVUE) of μ is $\bar{Y}$ by invoking the Lehmann-Scheffe Theorem (cf., Casella and Berger (1990)). On the other hand, it begs to reason that if the correlation coefficient ρ is not equal to zero, then the X_is should also contain information about μ, even if the marginal distributions of the X_is do not depend on μ, and consequently that perhaps one could improve on the estimator δ_1 when ν is known, by utilizing both the Y_is and X_is. In this note we shall examine this issue under two cases: (i) the parameters in Σ are all known; and (ii) the parameters in Σ are all unknown.

The motivation behind this note is borne out by considering methods used in multiple testing arising in the analysis of microarray data (cf., Dudoit and van der Laan (2008)) and also in marginal modeling approaches in event time analysis (cf., Therneau and Grambsch (2000)). The situation is also relevant in matched-pair studies, e.g., in twin studies, where the observable variables are correlated and of interest is a parameter in the marginal distribution of the variable observed in the first member but not in the marginal distribution of the variable observed in the second member of each pair. The situation in the analysis of microarray data is that you will have M genes and for each gene you will have sample data. The random observables for these genes need not be independent especially for those genes that are important. One is usually interested in testing M pairs of hypotheses associated with each of the genes, or estimating parameters for each of the genes. Many existing procedures for doing the multiple testing or the simultaneous estimation of parameters possess the characteristic that the procedure for the mth gene only utilizes data for the mth gene and not from the other genes. This leads to the question on whether these procedures are inefficient. In the area of event time analysis, especially those dealing with recurrent events, a typical approach is to specify marginal models for each of the event position occurrences, and perform the analysis based on the associated marginal data. In such settings, there is usually a strong type of dependence among the observables, possibly arising from the data accrual scheme, and consequently the question arises on whether there is loss in efficiency by utilizing marginal modeling and inference procedures. Thus, we hope that by considering in this note a very simple structure using Gaussian distributions, we will be able to highlight some of the consequences of marginalization in performing inference. It is our belief that this is a more profound issue permeating many areas more complicated than the simple setting considered in this note.

2 Σ Parameters Known

Let us first consider the situation when σ_1, σ_2, and ρ are all known, and recall that we have set $\nu = \nu_0 = 0$, else just consider subtracting ν_0 from each of the X_is. Let $\beta = \rho\sigma_1/\sigma_2$ be the regression coefficient. Then it is easy to see by applying the exponential family theorem regarding complete

sufficient statistics (cf., Casella and Berger (1990)) that a complete sufficient statistic for μ is

$$S \equiv S(\mathbf{X}, \mathbf{Y}) = \bar{Y} - \beta \bar{X} \tag{2}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$. Since S is unbiased for μ, then in fact, by Lehmann-Scheffe Theorem (cf., Casella and Berger (1990)), the estimator given by

$$\delta_2 = S = \bar{Y} - \beta \bar{X} \tag{3}$$

is the UMVUE of μ. Its variance is easily obtained to be $\text{Var}(\delta_2) = \sigma_1^2(1 - \rho^2)/n$. Consequently, since both δ_1 and δ_2 are unbiased for μ, the efficiency of δ_2 relative to δ_1 could be measured by the variance ratio. This is given by

$$\text{Eff}(\delta_2 : \delta_1) = \frac{1}{1 - \rho^2}. \tag{4}$$

Observe that δ_2 is always more efficient than δ_1 when $\rho \neq 0$, so that one gains by also utilizing the X_i-values aside from the Y_i-values, with the efficiency increasing as the degree of dependence between Y and X increases.

3 Σ Parameters Unknown

The results in Section 2 are hardly surprising, except for the fact perhaps that the estimator utilizing the full data could considerably be more efficient than the estimator using only the variable that is directly related to the parameter of interest from a marginal distribution perspective. In this section we now consider the more interesting situation where there are nuisance parameters, which are the parameters of the covariance matrix Σ. In this setting, S in (2) is not anymore a statistic, hence δ_2 is not an estimator, though clearly δ_1 is still an unbiased estimator of μ. An obvious idea is to obtain an estimator of the regression coefficient β, say $\hat{\beta}$, and then to plug-in this estimator for β in (3) to generate an estimator for μ. To implement this idea, we consider the usual estimator of β (cf., Casella and Berger (1990); Neter, Kutner, Nachtsheim, and Wasserman (1996)) given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{5}$$

Our estimator of μ in this setting will be

$$\delta_3 = \bar{Y} - \hat{\beta}\bar{X}. \tag{6}$$

By utilizing the Factorization theorem and/or the exponential family theorem, it is straightforward to see that a minimal sufficient statistic for the parameter $(\mu, \sigma_1^2, \sigma_2^2, \rho)$ is given by

$$S^*(\mathbf{Y}, \mathbf{X}) = \left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2, \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i X_i \right). \tag{7}$$

Note that $\hat{\beta}$ is a function of S^*. However, observe as a consequence of Theorem 1 below that S^* is *not* a complete sufficient statistic. Because of the unbiasedness of both δ_1 and δ_3, the function of S^* given by $g(S^*) = \delta_1 - \delta_3 = \hat{\beta}\bar{X}$, has expectation zero whatever the parameter value, but it is not identically zero. Thus, we cannot utilize the Lehmann-Scheffe Theorem to claim that δ_3 is the UMVUE of μ; in fact, from the efficiency result of Theorem 1, it does not actually dominate δ_1, so that it could not be the UMVUE of μ, though it is a function of the minimal sufficient statistic and is unbiased.

We now present results pertaining to the unbiasedness of δ_3 for μ and an efficiency result for δ_3 relative to δ_1.

Theorem 1 *The estimator δ_3 is unbiased for μ, and if $n > 3$, then*

$$\text{Eff}(\delta_3 : \delta_1) = \left(\frac{1}{1 - \rho^2} \right) \left(\frac{n-3}{n-2} \right),$$

so that $\delta_3 = \bar{Y} - \hat{\beta}\bar{X}$ is more efficient than $\delta_1 = \bar{Y}$ iff $|\rho| > 1/\sqrt{n-2}$.

Before proving this result, observe that in contrast to the case with Σ known wherein δ_2 uniformly dominates δ_1, in this setting with Σ unknown, δ_3 does not anymore uniformly dominate δ_1. In fact, $\delta_1 = \bar{Y}$ is a better estimator than $\delta_3 = \bar{Y} - \hat{\beta}\bar{X}$ in a neighborhood of ρ at zero, though this is a shrinking neighborhood as n increases. This also shows that neither δ_1 nor δ_3 can be a UMVUE for μ in this situation with Σ unknown. The loss of efficiency incurred by δ_3 relative to δ_2 is the price one pays by the need to estimate the nuisance parameters. Interestingly, the results in Theorem 1 are invariant with respect to the values of the variances σ_1^2 and σ_2^2. We now prove the theorem.

Proof of Theorem 1: The distributional model for $(Y_i, X_i), i = 1, 2, \ldots, n$, is equivalent to having $X_1, \ldots, X_n$ IID from $N(0, \sigma_2^2)$, letting $Z_1, \ldots, Z_n$ be IID from $N(0, 1)$ and independent of the X_is, and then defining the Y_is according to

$$Y_i = \mu + \beta X_i + \sigma_1 \sqrt{1 - \rho^2} Z_i, \ i = 1, 2, \ldots, n. \tag{8}$$

With $\bar{Z} = \sum_{i=1}^n Z_i/n$, then $\bar{Y} = \mu + \beta \bar{X} + \sigma_1 \sqrt{1 - \rho^2} \bar{Z}$, so that for $i = 1, 2, \ldots, n$,

$$\begin{aligned} Y_i - \bar{Y} &= \beta(X_i - \bar{X}) + \sigma_1 \sqrt{1 - \rho^2}(Z_i - \bar{Z}) \\ (X_i - \bar{X})(Y_i - \bar{Y}) &= \beta(X_i - \bar{X})^2 + \sigma_1 \sqrt{1 - \rho^2}(Z_i - \bar{Z})(X_i - \bar{X}). \end{aligned}$$

For $i = 1, 2, \ldots, n$, let

$$c_i = c_i(X_1, \ldots, X_n) = \frac{X_i - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}. \tag{9}$$

Then, it follows from the above expressions that $\hat{\beta} = \beta + \sigma_1 \sqrt{1 - \rho^2} \sum_{i=1}^n c_i(\mathbf{X})(Z_i - \bar{Z})$, and

$$\delta_3 = \mu + \sigma_1 \sqrt{1 - \rho^2} \left[\bar{Z} - \bar{X} \sum_{i=1}^n c_i(\mathbf{X})(Z_i - \bar{Z}) \right]. \tag{10}$$

By conditioning on $\mathbf{X}$, then using the independence between $\mathbf{X}$ and $\mathbf{Z}$, and the fact that the Z_is have zero means, it follows immediately from (10) that $\mathrm{E}(\delta_3|\mathbf{X}) = \mu$, and the unbiasedness of δ_3 for μ follows.

By the iterated variance rule (cf., Casella and Berger (1990)), and since $\mathrm{Var}[\mathrm{E}(\delta_3|\mathbf{X})] = 0$, then $\mathrm{Var}(\delta_3) = \mathrm{E}[\mathrm{Var}(\delta_3|\mathbf{X})]$. But, from the representation of δ_3 in (10),

$$\begin{aligned} \mathrm{Var}(\delta_3|\mathbf{X}) &= \sigma_1^2(1 - \rho^2) \left\{ \mathrm{Var}(\bar{Z}|\mathbf{X}) + \bar{X}^2 \mathrm{Var} \left[\sum_{i=1}^n c_i(Z_i - \bar{Z})|\mathbf{X} \right] - \right. \\ &\qquad \left. 2\bar{X} \mathrm{Cov} \left[\bar{Z}, \sum_{i=1}^n c_i(Z_i - \bar{Z})|\mathbf{X} \right] \right\}. \end{aligned}$$

Normal distributional theory (cf., Casella and Berger (1990)) now yields that $\mathrm{Var}(\bar{Z}|\mathbf{X}) = 1/n$ and the independence between $\bar{Z}$ and the vector $(Z_i - \bar{Z}, i = 1, 2, \ldots, n)$, so that the covariance term above becomes zero. Furthermore, it is easy to see that $\mathrm{Var}(Z_i - \bar{Z}) = (n-1)/n$ and $\mathrm{Cov}(Z_i - \bar{Z}, Z_j - \bar{Z}) = -1/n, (i \neq j)$. Since

$$\sum_{i=1}^n c_i(\mathbf{X}) = 0 \quad \text{and} \quad \sum_{i=1}^n c_i(\mathbf{X})^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

then

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^{n} c_i(Z_i - \bar{Z})|\mathbf{X}\right] &= \frac{n-1}{n}\sum_{i=1}^{n} c_i^2 - \frac{1}{n}\sum_{i\neq j} c_i c_j \\ &= \sum_{i=1}^{n} c_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} c_i\right)^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.\end{aligned}$$

Consequently,

$$\mathrm{Var}(\delta_3|\mathbf{X}) = \sigma_1^2(1-\rho^2)\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]. \tag{11}$$

Note that the conditional variance expression in (11) is actually directly obtainable by simply recalling the variance of the estimator of the intercept term in the simple linear regression model with Gaussian errors and fixed regressors (cf., Casella and Berger (1990); Neter et al. (1996)). In this simple linear regression model, the error variance is $\sigma_1^2(1-\rho^2)$, which is the conditional variance of Y, given $X = x$, in the bivariate normal model. This simple linear regression structure is contained in the representation of the Y_is given in (8).

By normal distribution theory (cf., Casella and Berger (1990)), $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent. Furthermore,

$$\frac{1}{\sigma_2^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1},$$

where χ^2_k represents a central chi-square distribution with k degrees-of-freedom. Also, it is easy to check that $\mathrm{E}(1/\chi^2_{n-1}) = 1/(n-3)$ provided $n > 3$. Since $\mathrm{E}(\bar{X}^2) = \mathrm{Var}(\bar{X}) = \sigma_2^2/n$, then by taking expectation of the expression in (11) with respect to $\mathbf{X}$, we obtain

$$\mathrm{Var}(\delta_3) = \frac{\sigma_1^2(1-\rho^2)}{n}\left(1 + \frac{1}{n-3}\right) = \mathrm{Var}(\delta_1)(1-\rho^2)\left(\frac{n-2}{n-3}\right). \tag{12}$$

The efficiency results in the theorem follow from the above expression. ‖

4 An Information Viewpoint

The results in the preceding sections prompt the question on the quantification of information content about parameters contained in random observables or their distributions. In the model considered, we could for instance ask about the information about μ that is contained in the random vector (Y, X), or that which is contained in the marginal random variables Y or X. It is

actually more precise to talk about the information about μ contained in the joint distribution of (X, Y) or from the marginal distributions of X or Y, though the terminology using the random variables leads to a simpler notation below. We have already observed that since the marginal distribution of X is independent of μ, then X by itself could not contain information about μ, but in conjunction with Y, that it does contain information about μ. We may therefore ask how much information is added to that already contained in Y by also using X.

Recall that statistical information about a parameter could be quantified by the Fisher information (cf., Casella and Berger (1990)). For a random vector $\mathbf{V}$ with joint density function $f_{\mathbf{V}}(\mathbf{v}|\theta)$ where the parameter $\theta \in \Theta$, under certain regularity conditions, the Fisher information matrix for θ contained in the distribution of $\mathbf{V}$ is defined by

$$\mathcal{I}_{\mathbf{V}}(\theta) = \mathrm{Cov}_{\mathbf{V}|\theta}\left[\nabla_\theta \log f_{\mathbf{V}}(\mathbf{V}|\theta)\right] = -\mathrm{E}_{\mathbf{V}|\theta}\left[\nabla_{\theta^t}\nabla_\theta \log f_{\mathbf{V}}(\mathbf{V}|\theta)\right] \tag{13}$$

where $\nabla_\theta = \partial/\partial\theta$. The subscript $\mathbf{V}$ in $\mathcal{I}_{\mathbf{V}}(\cdot)$ is to indicate that information is with respect to the distribution of $\mathbf{V}$. Using such a notation, we may for instance then have $\mathcal{I}_{\mathbf{V}_2|\mathbf{V}_1}(\theta)$ as the information about θ contained in the conditional distribution of $\mathbf{V}_2$, given $\mathbf{V}_1$. Note that $\mathcal{I}_{\mathbf{V}_2|\mathbf{V}_1}(\theta)$ will depend on θ as well as on $\mathbf{V}_1$. We define the *residual* Fisher information about θ contained in $(\mathbf{V}_1, \mathbf{V}_2)$, but not in $\mathbf{V}_1$, via

$$\mathcal{I}_{(\mathbf{V}_1,\mathbf{V}_2)\setminus\mathbf{V}_1}(\theta) \equiv \mathcal{I}_{(\mathbf{V}_1,\mathbf{V}_2)}(\theta) - \mathcal{I}_{\mathbf{V}_1}(\theta). \tag{14}$$

However, since by the multiplication rule, $f_{(\mathbf{V}_1,\mathbf{V}_2)}(\mathbf{v}_1, \mathbf{v}_2|\theta) = f_{\mathbf{V}_1}(\mathbf{v}_1|\theta) f_{\mathbf{V}_2|\mathbf{V}_1}(\mathbf{v}_2|\mathbf{v}_1;\theta)$, then we have the identity

$$\mathcal{I}_{(\mathbf{V}_1,\mathbf{V}_2)}(\theta) = \mathcal{I}_{\mathbf{V}_1}(\theta) + \mathrm{E}_{\mathbf{V}_1}\{\mathcal{I}_{\mathbf{V}_2|\mathbf{V}_1}(\theta)\}.$$

Thus, we obtain the identity relating residual information and conditional information given by

$$\mathcal{I}_{(\mathbf{V}_1,\mathbf{V}_2)\setminus\mathbf{V}_1}(\theta) = \mathrm{E}_{\mathbf{V}_1}\{\mathcal{I}_{\mathbf{V}_2|\mathbf{V}_1}(\theta)\}. \tag{15}$$

Going back now to the bivariate normal setting considered in this note and just by focusing on one observation ($n = 1$), we easily obtain, by applying the formula in (13), the following

Fisher informations about μ based on the different random observables or more precisely their distributions:

$$\mathcal{I}_{(Y,X)}(\mu) = \frac{1}{\sigma_1^2(1-\rho^2)}; \quad \mathcal{I}_Y(\mu) = \frac{1}{\sigma_1^2}; \quad \text{and} \quad \mathcal{I}_X(\mu) = 0. \tag{16}$$

Assuming that the nuisance parameters are known, by invoking the Cramer-Rao Inequality (cf., Casella and Berger (1990)), we obtain that any unbiased estimator of μ which depends on $(Y_i, X_i), i = 1, 2, \ldots, n$, will have a variance that is at least equal to

$$\text{CRLB}(\mu) = \frac{1}{n\mathcal{I}_{(Y,X)}(\mu)} = \frac{\sigma_1^2(1-\rho^2)}{n}. \tag{17}$$

Since δ_2 achieves this lower bound, this also verifies that δ_2 is the UMVUE of μ. It actually turns out that *even* when the nuisance parameters are not known, the lower bound for the variance of unbiased estimators of μ is still (17), a consequence of the parameter orthogonality of μ and Σ. However, in this situation this lower bound is not achievable, which can be seen by appealing to the necessary and sufficient condition for the Cramer-Rao lower bound to be achieved in exponential families (cf., Casella and Berger (1990)). Note in particular that the variance of δ_3, given in (12), exceeds the lower bound in (17), whereas when $\rho \neq 0$, the variance of δ_1 also exceeds this lower bound. These observations, together with the earlier comment that neither δ_1 nor δ_3 could be the UMVUE of μ, lead to the question on whether there actually exists a UMVUE for μ when Σ is unknown, a setting where the minimal sufficient statistic is not complete. Note that neither the Rao-Blackwell Theorem and Lehmann-Scheffe Theorem nor the Cramer-Rao Variance Inequality Theorem is applicable in deciding UMVUEness. Since it is an interesting exercise for students to argue that no such UMVUE exists for μ when Σ is unknown, we leave the proof of this non-existence to the reader and simply offer the hint: *consider sub-models!*

From (16), we also obtain that the residual Fisher information about μ contained in (Y, X), but not in Y, is

$$\mathcal{I}_{(Y,X)\backslash Y}(\mu) \equiv \mathcal{I}_{(Y,X)}(\mu) - \mathcal{I}_Y(\mu) = \frac{1}{\sigma_1^2(1-\rho^2)} - \frac{1}{\sigma_1^2} = \frac{1}{\sigma_1^2}\frac{\rho^2}{1-\rho^2}. \tag{18}$$

This could be viewed as the additional information about μ that could be attributed to X alone when (Y, X) are jointly observable. Note that this equals zero when $\rho = 0$, that is, when X and Y

are independent. Also, observe that $\mathcal{I}_{(Y,X)\setminus Y}(\mu)/\mathcal{I}_{(Y,X)}(\mu) = \rho^2$. The quantity ρ^2 is usually called the coefficient of determination.

Since the conditional distribution of X, given $Y = y$, is

$$X|Y = y \sim N\left(\rho\frac{\sigma_2}{\sigma_1}(y-\mu), \sigma_2^2(1-\rho^2)\right),$$

then applying (13) to this conditional distribution, we find that

$$\mathcal{I}_{X|Y=y}(\mu) = \frac{1}{\sigma_1^2}\frac{\rho^2}{1-\rho^2},$$

which is independent of the value y of the conditioning variable Y. Thus, in this bivariate normal setting,

$$\mathcal{I}_{(Y,X)\setminus Y}(\mu) = \mathrm{E}_Y\{\mathcal{I}_{X|Y}(\mu)\} = \mathcal{I}_{X|Y}(\mu). \tag{19}$$

Observe that the conditional distribution of X, given $Y = y$, does depend on the value y, but the conditional Fisher information $\mathcal{I}_{X|Y=y}(\mu)$ does not depend on y, and this is a consequence of the homoscedastic property of the bivariate normal distribution, which is that the conditional variance is invariant with respect to the conditioning value. We point out that the second equality in (19), and also that in (20) below, need not hold for other distributions since the conditional Fisher information may depend on the conditioning variable. This will be the case for instance with the trinomial distribution. In such cases the general formula in (15) provides the proper identity, that is, there is a need to take the expectation with respect to the conditioning variable of the conditional Fisher information to get the residual Fisher information.

In a similar vein, but to point out the asymmetric nature in the setting considered, since $\mathcal{I}_X(\mu) = 0$, we have that

$$\mathcal{I}_{(Y,X)\setminus X}(\mu) = \mathcal{I}_{(Y,X)}(\mu) - \mathcal{I}_X(\mu) = \mathcal{I}_{(Y,X)}(\mu) = \frac{1}{\sigma_1^2(1-\rho^2)}.$$

This indicates that the additional information about μ provided by Y alone when (Y, X) are jointly observable is in fact the total information available in (Y, X). This certainly is consistent with the fact that, marginally, X has no information about μ. Also, in contrast to the earlier result

where $\mathcal{I}_{(Y,X)\backslash Y}(\mu)/\mathcal{I}_{(Y,X)}(\mu)$ was equal to the coefficient of determination, note in this case that $\mathcal{I}_{(Y,X)\backslash X}(\mu)/\mathcal{I}_{(Y,X)}(\mu) = 1$, so there is an asymmetry in these two results.

The conditional distribution of Y, given $X = x$, is

$$Y|X = x \sim N\left(\mu + \rho\frac{\sigma_1}{\sigma_2}x, \sigma_1^2(1-\rho^2)\right),$$

which we note depends on x. Applying the Fisher information formula, we obtain

$$\mathcal{I}_{Y|X=x}(\mu) = \frac{1}{\sigma_1^2(1-\rho^2)},$$

which does not depend on the value x of the conditioning variable X. Therefore, in this bivariate normal setting,

$$\mathcal{I}_{(Y,X)\backslash X}(\mu) = \mathrm{E}_X\{\mathcal{I}_{Y|X}(\mu)\} = \mathcal{I}_{Y|X}(\mu). \tag{20}$$

5 Extensions

We conclude this note by pointing out several extensions. First, we point out that since ν is assumed known ($\nu = \nu_0 = 0$), instead of utilizing the estimator $\hat{\beta}$, we could have used the estimator given by

$$\tilde{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})X_i}{\sum_{i=1}^n X_i^2}. \tag{21}$$

However, the resulting plug-in estimator of μ given by $\delta_4 = \bar{Y} - \tilde{\beta}\bar{X}$ is biased, and it turns out that it is harder to deal with analytically compared to δ_3, that is, in the context of obtaining closed-form expressions. However, we intuitively surmise that with respect to mean-squared error, δ_4 will tend to be better than δ_3. We also expect that with respect to mean-squared error, δ_4 will tend to be better than δ_1 outside a small neighborhood of zero in ρ-space.

The result in Theorem 1 indicates that $\delta_1 = \bar{Y}$ tends to be better than δ_3 when ρ is close to zero. An idea therefore is to perform a hypothesis test of $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ using for instance Fisher's z-test for correlation (cf., Neter et al. (1996)). An adaptive estimator will then use δ_1 if the test leads to non-rejection of H_0, whereas if H_0 is rejected, then δ_3 is used. This resulting adaptive estimator is called a preliminary test estimator (see, for instance, the monograph by Saleh (2006)), which will be a biased estimator. In constructing such an estimator, there is the important

and non-trivial problem of deciding on the appropriate size or level of significance to use in the test of H_0 versus H_1. As also suggested by J. Tebbs, an interesting exercise for students would be to perform a computer simulation of the properties of such an adaptive estimator, as well as those of δ_4, and to compare them with δ_1 and δ_3.

Finally, we have dealt here with the bivariate normal setting to make things as simple as possible. It is clear that the results could be generalized to the multivariate setting, that is, for the setting where $(\mathbf{Y}_i{}^t, \mathbf{X}_i{}^t)^t, i = 1, 2, \ldots, n$, are IID random vectors from a $(p+q)$-dimensional multivariate normal distribution with partitioned mean vector and covariance matrix given, respectively, by

$$\eta = \begin{bmatrix} \mu \\ \nu \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}. \tag{22}$$

With ν assumed to equal some ν_0, which could be taken to be $\mathbf{0}$, interest would then be on the mean parameter μ and with the elements of the covariance matrix $\mathbf{\Sigma}$ unknown and viewed as nuisance parameters.

Acknowledgements

The authors wish to acknowledge the Statistical and Applied Mathematical Sciences Institute (SAMSI) for allowing them to visit during Fall 2008. As such this material was based upon work partially supported by the National Science Foundation (NSF) under Grant DMS-0635449 to SAMSI. However, any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

E. Peña also wishes to acknowledge the following grants which partially support his research: NSF Grant DMS 0805809, National Institutes of Health (NIH) Grant RR17698, and Environmental Protection Agency (EPA) Grant RD-83241902-0 to University of Arizona with subaward number Y481344 to the University of South Carolina. He also wishes to thank his colleagues Xianzheng (Shan) Huang, Joshua Tebbs, and Lianming Wang, and his students Joshua Habiger and Wensong Wu for very helpful discussions, comments, and criticisms.

References

Casella, G. and Berger, R. L. (1990), *Statistical inference*, The Wadsworth & Brooks/Cole Statistics/Probability Series, Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Dudoit, S. and van der Laan, M. J. (2008), *Multiple testing procedures with applications to genomics*, Springer Series in Statistics, New York: Springer.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), *Applied Linear Statistical Models*, Chicago: Irwin, 4th ed.

Saleh, A. K. M. E. (2006), *Theory of preliminary test and Stein-type estimation with applications*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Therneau, T. M. and Grambsch, P. M. (2000), *Modeling survival data: extending the Cox model*, Statistics for Biology and Health, New York: Springer-Verlag.