Journal of the Korean
Statistical Society
Vol. 19, No. 2, 1990

Posterior Density of Parameters in Multiresponse Regression Analysis with Missing Values in One Response

Gun Seog Kang*

ABSTRACT

In this article we develop the marginal posterior density of the model parameters in the multiresponse regression models when missing values exist only in one response. The resulting density resolves a couple of problems in the estimation approach proposed by Box, Draper, and Hunter(1970) and provides a general interpretation for relationship between the estimates of the missing values and the parameters.

1. Introduction

When we deal with the multivariate data, it is common to have some type of missing values and the statistical techniques developed for the complete data set should be modified to handle that situation. Box, Draper, and Hunter(1970) proposed an approach to estimate the model parameters when missing values occur in some of responses in multiresponse regression analysis. Treating the missing observations as a part of parameters, they obtained the joint posterior density of the model parameters and the missing values. Since it is not possible in most cases to derive the marginal posterior density of the model parameters, they suggested to estimate both the model parameters and the missing values by maximizing the joint posterior density of these two sets of parameters. One of shortcomings of this approach is that the "parameter" vector can become impractically long if there are many missing values. Also, since the missing observations are not usually of interest in themselves, we prefer to use the marginal density of the model parameters as possible.

In this article we show that it is always possible to derive the marginal posterior density of the model parameters when missing values exist only in one of responses. The resulting density also gives a clear interpretation of the estimates of the missing values. In Box et al. (1970), they noticed in a numerical example with a missing value that the estimate of the missing value can be interpreted as a predicted value from a linear regression using the observed values. We show that this fact generally holds under the data scheme we consider in this paper.

In the next section, we formulate the underlying model and briefly review previous work. We present the main result in Section 3 by deriving the marginal posterior density and show in Section 4 how this marginal posterior density can be used.

2. Model Formulation

We consider data from experiments where there are R responses, $\mathbf{y}_n=(y_{n1}, y_{n2}, \cdots, y_{nR})^T$, measured on the nth experimental run of total N runs and the models for the R responses depend on P parameters, $\theta=(\theta_1, \theta_2, \cdots, \theta_P)^T$. Let the vector $\mathbf{x}_n=(x_{n1}, x_{n2}, \cdots, x_{nK})^T$ represent the values of K independent variables in the nth experimental run. We assume that model functions, $f_r(\mathbf{x}_n, \theta)$, have been postulated as

$$y_{nr}=f_r(\mathbf{x}_n, \theta)+z_{nr} \tag{2.1}$$

where $n=1, \cdots, N$; $r=1, \cdots, R$, and $\mathbf{z}_n=(z_{n1}, z_{n2}, \cdots, z_{nR})^T$ denotes the error term.

Example : AB Box and Draper(1965) considered a chemical reaction system in which there are 2 responses with models given as

$$f_1=e^{-\kappa t}$$
$$f_2=1-e^{-\kappa t}$$

See Box and Draper(1965) for the data and its description. In this example, there are $N=10$ cases, $P=1$ parameter(κ), $R=2$ responses (f_1, f_2), and $K=1$ independent variable, time (t). ■

The model (2.1) can be written in matrix form as

$$\mathbf{Y}=\mathbf{F}(\mathbf{x}, \theta)+\mathbf{Z} \tag{2.2}$$

by collecting all the elements into matrices. The $N\times R$ observation matrix $\mathbf{Y}$ has y_{nr} as the (n, r)th elements, $\mathbf{F}(\mathbf{x}, \theta)$ is the $N\times R$ response matrix with the (n, r)th element $f_r(\mathbf{x}_n, \theta)$ and $\mathbf{Z}$ is the $N\times R$ residual matrix with the (n, r)th element z_{nr}. If the response matrix in (2.2) is given as a linear function of parameters like

$$\mathbf{F}(\mathbf{x}, \mathbf{B})=\mathbf{XB} \tag{2.3}$$

where $\mathbf{X}$ is the $N\times K$ design matrix which is common for all responses and $\mathbf{B}$ is the $K\times R$ parameter matrix, the model (2.2) belongs to the classical multivariate linear regression models(Anderson, 1984). When the structure of (2.3) is destroyed, for example, by the different design matrices for the different responses(Zellner, 1962, Tiao and Zellner, 1964) or by nonlinear relationships between the responses and the parameters(Box and Draper, 1965), the model (2.2) belongs to the multiresponse regression models, which are considered in this paper.

Assuming $\mathbf{z}_n$ follows independently an identical multivariate normal distribution with the zero mean vector and the common variance-covariance matrix Σ, Box and Draper(1965) derived the posterior density of θ as

$$p(\theta \mid \mathbf{Y})\propto \mid \mathbf{Z}^T\mathbf{Z} \mid ^{-N/2} \tag{2.4}$$

for the complete data set, when noninformative priors for θ and Σ are used. Hence, the parameter estimates $\hat{\theta}$ are chosen to minimize $\mid \mathbf{Z}^T\mathbf{Z} \mid$, which is called the determinant criterion. Bates and Watts(1984, 1987) developed a generalized Gauss-Newton method to optimize $\mid \mathbf{Z}^T\mathbf{Z} \mid$ by explicitly deriving the gradient and the Hessian of $\mid \mathbf{Z}^T\mathbf{Z} \mid$ with respect to θ. Further techniques for multiresponse regression analysis can be found in, for example, Box et al.(1973), Stewart and Sorensen(1981), Bates and Watts(1985), Kang and Bates(1990).

Box et al.(1970) modified the determinant criterion to handle the data with missing values. Treating the missing observations as parameters (say, $\mathbf{y}_m$) and using a locally uniform prior for $\mathbf{y}_m$, they showed that the marginal posterior density of θ and $\mathbf{y}_m$, by integrating out Σ, is

$$p(\theta, \mathbf{y}_m \mid \mathbf{Y})\propto \mid \mathbf{Z}^T\mathbf{Z} \mid ^{-N/2} \tag{2.5}$$

where each element of the residual matrix $\mathbf{Z}$ is now a function of both θ and $\mathbf{y}_m$. The estimate $\hat{\theta}$ of θ as well as the estimate $\hat{\mathbf{y}}_m$ of $\mathbf{y}_m$ are obtained by minimizing $|\mathbf{Z}^T\mathbf{Z}|$ with respect to both θ and $\mathbf{y}_m$ simultaneously.

3. Missing Values in One Response

As a special case of their approach, Box et al. (1970) obtained the marginal posterior density of $\theta=\log(\kappa)$ explicitly by integrating out the missing observation in Example : AB, where $y_{10,2}$ is assumed to be missing. In this section we show that this kind of integration can be done generally when there are missing values in only one response. We actually derive the marginal posterior density of θ and show that the process does not depend on the model function forms.

When we estimate the missing values, we suggest to deal with the missing residuals $\mathbf{z}_m$ instead of missing observations $\mathbf{y}_m$. In regression analysis, we have an assumption that the disturbance is independent of the model functions whereas the observed response is not. Hence, if the model specification is correct, estimating the residuals will have less effect on the estimation of the model parameters than estimating the observations. Also, the computational algorithm becomes much simpler if we deal with the missing residuals(Kang, 1988).

By taking the transformation $z_{mr}=y_{mr}-f_r(\mathbf{x}_n, \theta)$ for missing residuals, or by following directly the same procedure as in Box et al. (1970) with a locally uniform prior for $\mathbf{z}_m$, we can show that the marginal posterior density of θ and $\mathbf{z}_m$ is given by

$$p(\theta, \mathbf{z}_m \mid \mathbf{Y}) \propto |\mathbf{Z}^T\mathbf{Z}|^{-N/2} \tag{3.1}$$

Hence, we determine the values of θ and $\mathbf{z}_m$ to minimize $|\mathbf{Z}^T\mathbf{Z}|$. When we have missing values in one response, we can consider the following structure of the residual matrix

$$\mathbf{Z}=\begin{bmatrix}\mathbf{Z}_1 & \mathbf{z}_1\\ \mathbf{Z}_2 & \mathbf{z}_m\end{bmatrix}$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are matrices of sizes $N_1\times(R-1)$ and $N_2\times(R-1)$, respectively, $\mathbf{z}_1$ is an N_1-dimensional vector, and $\mathbf{z}_m$ is an N_2-dimensional vector containing the missing residuals. (We assume that there are N_2 missing residuals.) This structure is obtained by moving the response with the missing values to the last column of the data matrix, and by moving down any cases with a missing value to the end of the data matrix. It is possible to do this because we have assumed that the observations from the different experimental runs are independent. Then we have

$$\mathbf{Z}^T\mathbf{Z}=\begin{bmatrix}\mathbf{Z}_1^T\mathbf{Z}_1+\mathbf{Z}_2^T\mathbf{Z}_2 & \mathbf{Z}_1^T\mathbf{z}_1+\mathbf{Z}_2^T\mathbf{z}_m\\ \mathbf{z}_1^T\mathbf{Z}_1+\mathbf{z}_m^T\mathbf{Z}_2 & \mathbf{z}_1^T\mathbf{z}_1+\mathbf{z}_m^T\mathbf{z}_m\end{bmatrix}$$

Define $\mathbf{S}_{11}=\mathbf{Z}_1^T\mathbf{Z}_1+\mathbf{Z}_2^T\mathbf{Z}_2$ and $\mathbf{s}_{12}=\mathbf{Z}_1^T\mathbf{z}_1+\mathbf{Z}_1^T\mathbf{z}_1+\mathbf{Z}_1^T\mathbf{z}_m$ so the determinant of $\mathbf{Z}^T\mathbf{Z}$ can be written as

$$|\mathbf{Z}^T\mathbf{Z}| = |\mathbf{S}_{11}|\ \{\mathbf{z}_1^T\mathbf{z}_1+\mathbf{z}_m^T\mathbf{z}_m-\mathbf{s}_{12}^T\mathbf{S}_{11}^{-1}\mathbf{s}_{12}\} \tag{3.2}$$

Noting the second factor on the right of (3.2), which is a scalar, is a quadratic function of $\mathbf{z}_m$, we express it as a quadratic form for $\mathbf{z}_m$. If we set

$$\mathbf{z}_1^T\mathbf{z}_1+\mathbf{z}_m^T\mathbf{z}_m-\mathbf{s}_{12}^T\mathbf{S}_{11}^{-1}\mathbf{s}_{12}=c+(\mathbf{z}_m-\mu)^T\mathbf{A}(\mathbf{z}_m-\mu),$$

then $\mathbf{A}$, μ, and c are found as

$$\mathbf{A}=\mathbf{I}_{N_2}-\mathbf{Z}_2\mathbf{S}_{11}^{-1}\mathbf{Z}_2^T$$
$$\mu=\mathbf{A}^{-1}\mathbf{Z}_2\mathbf{S}_{11}^{-1}\mathbf{Z}_1^T\mathbf{z}_1$$
$$c=\mathbf{z}_1^T\mathbf{z}_1-\mathbf{z}_1^T\mathbf{Z}_1\mathbf{S}_{11}^{-1}\mathbf{Z}_1^T\mathbf{z}_1-\mu^T\mathbf{A}\mu$$

To simplify the above expressions further, we note that the inverse matrix of **A** can be given as

$$\mathbf{A}^{-1}=\mathbf{I}_{N_2}+\mathbf{Z}_2(\mathbf{Z}_1^T\mathbf{Z}^T)^{-1}\mathbf{Z}_2^T \tag{3.3}$$

using the formula from Rao(1973, p.33) and the determinant of **A** is

$$\begin{aligned}|\mathbf{A}| &= |\mathbf{S}_{11}|^{-1}\,|\mathbf{Z}_1^T\mathbf{Z}_1| \\ &= |\mathbf{Z}_1^T\mathbf{Z}_1+\mathbf{Z}_2^T\mathbf{Z}_2|^{-1}\,|\mathbf{Z}_1^T\mathbf{Z}_1|\end{aligned} \tag{3.4}$$

using the fact that $|\mathbf{I}_m-\mathbf{CD}|=|\mathbf{I}_n-\mathbf{DC}|$ for an $m\times n$ matrix **C** and an $n\times m$ matrix **D**(Tiao and Zellner, 1964). The expression for μ and c is simplified by using (3.3) to

$$\mu=\mathbf{Z}_2(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\,\mathbf{Z}_1^T\mathbf{z}_1$$
$$c=\mathbf{z}_1^T\mathbf{z}_1-\mathbf{z}_1^T\mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\,\mathbf{Z}_1^T\mathbf{z}_1$$

Note that expressions for **A**, μ, and c do not involve the missing residuals. The joint posterior density of θ and $\mathbf{z}_m$ then can be written as

$$p(\theta,\ \mathbf{z}_m\mid\mathbf{Y})\propto|\mathbf{S}_{11}|^{-N/2}\{c+(\mathbf{z}_m-\mu)^T\,\mathbf{A}(\mathbf{z}_m-\mu)\}^{-N/2} \tag{3.5}$$

The second term on the right of (3.5) is in the form of an N_2-variate Student's t density and so we can integrate out $\mathbf{z}_m$ by comparing with a t density to get the marginal posterior density of θ. It follows that

$$p(\theta\mid\mathbf{Y})\propto|\mathbf{S}_{11}|^{-N/2}\,c^{-N_1/2}\,N\times(R-1) \tag{3.6}$$

To obtain a more simplified expression of c, we define

$$\mathbf{Z}_R=\begin{bmatrix}\mathbf{Z}_1 & \mathbf{z}_1\end{bmatrix}\quad:\ N_1\times R$$
$$\mathbf{Z}_C=\begin{bmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{bmatrix}\ :\ N\times(R-1)$$

Then $\mathbf{S}_{11}=\mathbf{Z}_C^T\mathbf{Z}_C$ and, since

$$\begin{aligned}|\mathbf{Z}_R^T\mathbf{Z}_R| &= |\mathbf{Z}_1^T\mathbf{Z}_1|\ \{\mathbf{z}_1^T\mathbf{z}_1-\mathbf{z}_1^T\mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T\mathbf{z}_1\} \\ &= c\,|\mathbf{Z}_1^T\mathbf{Z}_1|\end{aligned}$$

we have

$$c=\frac{|\mathbf{Z}_R^T\mathbf{Z}_R|}{|\mathbf{Z}_1^T\mathbf{Z}_1|} \tag{3.7}$$

From (3.4), $|\mathbf{A}|=|\mathbf{Z}_C^T\mathbf{Z}_C|^{-1}\,|\mathbf{Z}_1^T\mathbf{Z}_1|$. Plugging this and (3.7) into (3.6), we have

$$p(\theta\mid\mathbf{Y})\propto\frac{|\mathbf{Z}_C^T\mathbf{Z}_C|^{-(N-1)/2}\,|\mathbf{Z}_R^T\mathbf{Z}_R|^{-N_1/2}}{|\mathbf{Z}_1^T\mathbf{Z}_1|^{-(N_1-1)/2}} \tag{3.8}$$

If we apply this equation to the first illustration of Example : AB of Box et al.(1970), we obtain the same expression for the marginal posterior density of $\theta=\log(\kappa)$ as they did.

We also note from equation (3.5) that the conditional distribution of $\mathbf{z}_m$ given θ is a multivariate t distribution with the location parameter μ. Hence, an approximate estimate for the missing residuals $\mathbf{z}_m$ is given by μ evaluated at θ. That is,

$$\mathbf{z}_m=\hat{\mathbf{Z}}_2(\hat{\mathbf{Z}}_1^T\hat{\mathbf{Z}}_1)^{-1}\,\hat{\mathbf{Z}}_1^T\hat{\mathbf{z}}_1 \tag{3.9}$$

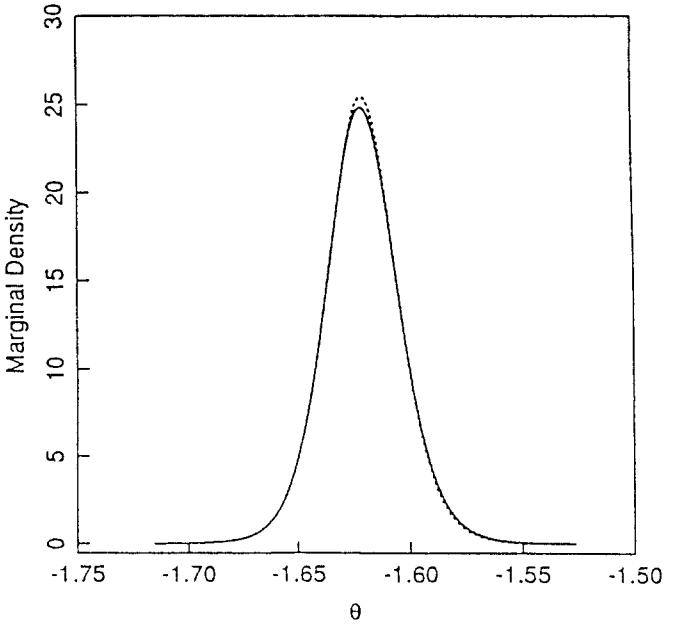
where $\hat{\mathbf{Z}}_1$, $\hat{\mathbf{Z}}_2$, and $\hat{\mathbf{z}}_1$ are $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{z}_1$ evaluated at $\hat{\theta}$, respectively. This explicit expression of the estimate of $\mathbf{z}_m$ also can be found by directly minimizing $|\mathbf{Z}^T\mathbf{Z}|$ with respect to $\mathbf{z}_m$. Equation (3.9)

columns being the independent variables. The vector $(\hat{\mathbf{Z}}_1^T\hat{\mathbf{Z}}_1)^{-1}\hat{\mathbf{Z}}_1^T\hat{\mathbf{z}}_1$ is the estimated regression coefficient vector when we regress $\hat{\mathbf{z}}_1$ on $\hat{\mathbf{Z}}_1$ with N_1 observations, and z_m is the predicted values at the values of $\hat{\mathbf{Z}}_2$. Box et al. (1970) described a similar result when they analyzed Example : AB. Assuming $y_{10.2}$ is missing, they noted that $z_{10.2}=(\Sigma^9_{n=1}\hat{z}_{n1}\hat{z}_{n2})/(\Sigma^9_{n=1}\hat{z}_{n1}^2)\cdot\hat{z}_{10.1}$, which can be obtained directly from (3.9).

4. Applications

The marginal posterior density $p(\theta \mid \mathbf{Y})$ obtained in the previous section can be used in two ways. First we can use it as an estimation criterion for the model parameters. That is, instead of using the joint posterior density (3.1), we can estimate θ by maximizing (3.8) directly with respect to θ. This reduces the number of parameters to be estimated. For this optimization, we can use the generalized Gauss-Newton method(Bates and Watts, 1984, 1987). If we decide to estimate θ by minimizing $-\log p(\theta \mid \mathbf{Y})$, then the objective function is now a linear combination of three parts, each of which is the logarithm of a determinant. As shown in Bates and Watts(1985), we can easily develop the gradient and an approximate Hessian of the logarithm of determinants involved, which, in turn, provides the gradient and an approximate Hessian of $-\log p(\theta \mid \mathbf{Y})$ with respect to θ.

Secondly we can use this exact marginal posterior density of θ to check the validity of any approximation method. For example, the quadratic approximation to the posterior density (2.4) for the complete data set (Bates and Watts, 1985) is a possible approximation method which can also be applied to the joint posterior density (3.1). This approximation seems to work nicely because it makes the analysis easy by using the properties of the multivariate t distribution. The quality of this quadratic approximation can be measured by comparing with the exact marginal posterior density (3.8). We demonstrate this application using Example : AB. Since there is only one parameter in this example, it is possible to plot its posterior density. In Fig. 1, the solid line is the exact marginal posterior density of θ when $y_{10.2}$ is assumed to be missing, and the dotted line is the approximate marginal density obtained from the quadratic approximation. (Each density is normalized numerically.) The closeness of the two densities shows good performance of the quadratic approximation method in this example.



Acknowledgement

The author is grateful for the helpful comments and suggestions of referees.

References

1. Anderson, T.W.(1984), *An Introduction to Multivariate Statistical Analysis*(2nd ed.), Wiley, New York.
2. Bates, D.M. and Watts, D.G.(1984), A Multi-Response Gauss-Newton Algorithm, *Communications in Statistics, Part B-Simulation and Computation*, 13, 705-715.
3. Bates, D.M. and Watts, D.G.(1985), Multiresponse Estimation with Special Application to Linear Systems of Differential Equations(with discussion), *Technometrics*, 27, 329-360.
4. Bates, D.M. and Watts, D.G.(1987), A Generalized Gauss-Newton Procedure for Multi-response Parameter Estimation, *SIMAN Journal of Scientific and Statistical Computing*, 7(1), 49-55.
5. Box, G.E.P. and Draper, N.R.(1965), The Bayesian Estimation of Common Parameters from Several Responses, *Biometrika*, 52, 355-365.
6. Box, G.E.P., Hunter, W.G., MacGregor, J.F., and Erjavec, J.(1973), Some Problems Associated with the Analysis of Multiresponse Models, *Technometrics*, 15, 33-51.
7. Box, M.J., Draper, N.R., and Hunter, W.G.(1970), Missing Values in Multiresponse Nonlinear Data Fitting, *Technometrics*, 12, 613-620.
8. Kang, G.(1988). Ph. D. Dissertation, University of Wisconsin-Madison.
9. Kang, G. and Bates, D.M.(1990), Approximate Inference in Multiresponse Regression Analysis, *Biometrika*, 77, 321-331.
10. Rao, C.R.(1973), *Linear Statistical Inference and Its Applications*(2nd edition), Wiley, New York.
11. Stewart, W.E. and Sorensen, J.P.(1981), Bayesian Estimation of Common Parameters from Multiresponse Data with Missing Observations, *Technometrics*, 23, 131-141.
12. Tiao, G.C. and Zellner, A.(1964), On the Bayesian Estimation of Multivariate Regression, *Journal of the Royal Statistical Society, Series B*, 26, 277-285.
13. Zellner, A.(1962), An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, *Journal of American Statistical Association*, 58, 977-992.