Bias caused by a misspecified model. Suppose some important predictor variables are omitted from the proposed regression model. That is, suppose the true model has $\mathbf{Z} = [\mathbf{Z}_1 \mid \mathbf{Z}_2]$ with rank $r + 1$ and

$$\underset{(n\times 1)}{\mathbf{Y}} = \left[\underset{(n\times(q+1))}{\mathbf{Z}_1} \;\vdots\; \underset{(n\times(r-q))}{\mathbf{Z}_2}\right] \begin{bmatrix} \underset{((q+1)\times 1)}{\boldsymbol{\beta}_{(1)}} \\ \hdashline \underset{((r-q)\times 1)}{\boldsymbol{\beta}_{(2)}} \end{bmatrix} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{7-20}$$

$$= \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. However, the investigator unknowingly fits a model using only the first q predictors by minimizing the error sum of squares $(\mathbf{Y} - \mathbf{Z}_1\boldsymbol{\beta}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\boldsymbol{\beta}_{(1)})$. The least squares estimator of $\boldsymbol{\beta}_{(1)}$ is $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$. Then, unlike the situation when the model is correct,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{(1)}) &= (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'E(\mathbf{Y}) = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'(\mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + E(\boldsymbol{\varepsilon})) \\ &= \boldsymbol{\beta}_{(1)} + (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Z}_2\boldsymbol{\beta}_{(2)} \end{aligned} \tag{7-21}$$

That is, $\hat{\boldsymbol{\beta}}_{(1)}$ is a biased estimator of $\boldsymbol{\beta}_{(1)}$ unless the columns of $\mathbf{Z}_1$ are perpendicular to those of $\mathbf{Z}_2$ (that is, $\mathbf{Z}_1'\mathbf{Z}_2 = \mathbf{0}$). If important variables are missing from the model, the least squares estimates $\hat{\boldsymbol{\beta}}_{(1)}$ may be misleading.

7.7 Multivariate Multiple Regression

In this section, we consider the problem of modeling the relationship between m responses $Y_1, Y_2, \ldots, Y_m$ and a single set of predictor variables $z_1, z_2, \ldots, z_r$. Each response is assumed to follow its own regression model, so that

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}z_1 + \cdots + \beta_{r1}z_r + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}z_1 + \cdots + \beta_{r2}z_r + \varepsilon_2 \\ &\vdots \qquad\qquad \vdots \\ Y_m &= \beta_{0m} + \beta_{1m}z_1 + \cdots + \beta_{rm}z_r + \varepsilon_m \end{aligned} \tag{7-22}$$

The error term $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m]$ has $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Thus, the error terms associated with different responses may be correlated.

To establish notation conforming to the classical linear regression model, let $[z_{j0}, z_{j1}, \ldots, z_{jr}]$ denote the values of the predictor variables for the jth trial, let $\mathbf{Y}_j' = [Y_{j1}, Y_{j2}, \ldots, Y_{jm}]$ be the responses, and let $\boldsymbol{\varepsilon}_j' = [\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm}]$ be the errors. In matrix notation, the design matrix

$$\underset{(n\times(r+1))}{\mathbf{Z}} = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix}$$

is the same as that for the single-response regression model. [See (7-3).] The other matrix quantities have multivariate counterparts. Set

$$\underset{(n\times m)}{\mathbf{Y}} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix} = [\mathbf{Y}_{(1)} \,\vdots\, \mathbf{Y}_{(2)} \,\vdots\, \cdots \,\vdots\, \mathbf{Y}_{(m)}]$$

$$\underset{((r+1)\times m)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} = [\boldsymbol{\beta}_{(1)} \,\vdots\, \boldsymbol{\beta}_{(2)} \,\vdots\, \cdots \,\vdots\, \boldsymbol{\beta}_{(m)}]$$

$$\underset{(n\times m)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = [\boldsymbol{\varepsilon}_{(1)} \,\vdots\, \boldsymbol{\varepsilon}_{(2)} \,\vdots\, \cdots \,\vdots\, \boldsymbol{\varepsilon}_{(m)}]$$

$$= \begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \hdashline \boldsymbol{\varepsilon}_2' \\ \hdashline \vdots \\ \hdashline \boldsymbol{\varepsilon}_n' \end{bmatrix}$$

The *multivariate linear regression model* is

$$\underset{(n\times m)}{\mathbf{Y}} = \underset{(n\times(r+1))}{\mathbf{Z}} \; \underset{((r+1)\times m)}{\boldsymbol{\beta}} + \underset{(n\times m)}{\boldsymbol{\varepsilon}}$$

with (7-23)

$$E(\boldsymbol{\varepsilon}_{(i)}) = \mathbf{0} \quad \text{and} \quad \operatorname{Cov}(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(k)}) = \sigma_{ik}\mathbf{I} \qquad i, k = 1, 2, \ldots, m$$

The m observations on the jth trial have covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ik}\}$, but observations from different trials are uncorrelated. Here $\boldsymbol{\beta}$ and σ_{ik} are unknown parameters; the design matrix $\mathbf{Z}$ has jth row $[z_{j0}, z_{j1}, \ldots, z_{jr}]$.

very important!

Simply stated, the ith response $\mathbf{Y}_{(i)}$ follows the linear regression model

$$\mathbf{Y}_{(i)} = \mathbf{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \qquad i = 1, 2, \ldots, m \tag{7-24}$$

with $\operatorname{Cov}(\boldsymbol{\varepsilon}_{(i)}) = \sigma_{ii}\mathbf{I}$. However, the errors for *different* responses on the *same* trial can be correlated.

Given the outcomes $\mathbf{Y}$ and the values of the predictor variables $\mathbf{Z}$ with full column rank, we determine the least squares estimates $\hat{\boldsymbol{\beta}}_{(i)}$ exclusively from the observations $\mathbf{Y}_{(i)}$ on the ith response. In conformity with the single-response solution, we take

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}_{(i)} \tag{7-25}$$

Collecting these univariate least squares estimates, we obtain

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\beta}}_{(m)}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[\mathbf{Y}_{(1)} \mid \mathbf{Y}_{(2)} \mid \cdots \mid \mathbf{Y}_{(m)}]$$

or

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \tag{7-26}$$

For any choice of parameters $\mathbf{B} = [\mathbf{b}_{(1)} \mid \mathbf{b}_{(2)} \mid \cdots \mid \mathbf{b}_{(m)}]$, the matrix of errors is $\mathbf{Y} - \mathbf{ZB}$. The error sum of squares and cross products matrix is

$$(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB})$$

$$= \begin{bmatrix} (\mathbf{Y}_{(1)} - \mathbf{Zb}_{(1)})'(\mathbf{Y}_{(1)} - \mathbf{Zb}_{(1)}) & \cdots & (\mathbf{Y}_{(1)} - \mathbf{Zb}_{(1)})'(\mathbf{Y}_{(m)} - \mathbf{Zb}_{(m)}) \\ \vdots & & \vdots \\ (\mathbf{Y}_{(m)} - \mathbf{Zb}_{(m)})'(\mathbf{Y}_{(1)} - \mathbf{Zb}_{(1)}) & \cdots & (\mathbf{Y}_{(m)} - \mathbf{Zb}_{(m)})'(\mathbf{Y}_{(m)} - \mathbf{Zb}_{(m)}) \end{bmatrix} \tag{7-27}$$

The selection $\mathbf{b}_{(i)} = \hat{\boldsymbol{\beta}}_{(i)}$ minimizes the ith diagonal sum of squares $(\mathbf{Y}_{(i)} - \mathbf{Zb}_{(i)})'(\mathbf{Y}_{(i)} - \mathbf{Zb}_{(i)})$. Consequently, $\text{tr}[(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB})]$ is minimized by the choice $\mathbf{B} = \hat{\boldsymbol{\beta}}$. Also, the generalized variance $|(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB})|$ is minimized by the least squares estimates $\hat{\boldsymbol{\beta}}$. (See Exercise 7.11 for an additional generalized sum of squares property.)

Using the least squares estimates $\hat{\boldsymbol{\beta}}$, we can form the matrices of

Predicted values:	$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$
Residuals:	$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y}$

(7-28)

The orthogonality conditions among the residuals, predicted values, and columns of $\mathbf{Z}$, which hold in classical linear regression, hold in multivariate multiple regression. They follow from $\mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'] = \mathbf{Z}' - \mathbf{Z}' = \mathbf{0}$. Specifically,

$$\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} = \mathbf{0} \tag{7-29}$$

so the residuals $\hat{\boldsymbol{\varepsilon}}_{(i)}$ are perpendicular to the columns of $\mathbf{Z}$. Also,

$$\hat{\mathbf{Y}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}'\mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} = \mathbf{0} \tag{7-30}$$

confirming that the predicted values $\hat{\mathbf{Y}}_{(i)}$ are perpendicular to all residual vectors $\hat{\boldsymbol{\varepsilon}}_{(k)}$. Because $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}$,

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \mathbf{0} + \mathbf{0}'$$

or

$$\begin{array}{ccccc} \mathbf{Y}'\mathbf{Y} & = & \hat{\mathbf{Y}}'\hat{\mathbf{Y}} & + & \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ \left(\begin{array}{c}\text{total sum of squares}\\ \text{and cross products}\end{array}\right) & = & \left(\begin{array}{c}\text{predicted sum of squares}\\ \text{and cross products}\end{array}\right) & + & \left(\begin{array}{c}\text{residual (error) sum}\\ \text{of squares and}\\ \text{cross products}\end{array}\right) \end{array} \tag{7-31}$$

The residual sum of squares and cross products can also be written as

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\beta}} \tag{7-32}$$

Example 7.8 (Fitting a multivariate straight-line regression model) To illustrate the calculations of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, and $\hat{\boldsymbol{\varepsilon}}$, we fit a straight-line regression model (see Panel 7.2),

$$Y_{j1} = \beta_{01} + \beta_{11}z_{j1} + \varepsilon_{j1}$$
$$Y_{j2} = \beta_{02} + \beta_{12}z_{j1} + \varepsilon_{j2}, \qquad j = 1, 2, \ldots, 5$$

to two responses Y_1 and Y_2 using the data in Example 7.3. These data, augmented by observations on an additional response, are as follows:

z_1	0	1	2	3	4
y_1	1	4	3	8	9
y_2	−1	−1	2	3	2

The design matrix $\mathbf{Z}$ remains unchanged from the single-response problem. We find that

$$\overset{2\times 5}{\mathbf{Z}'} = \begin{bmatrix} \overbrace{1 \quad 1 \quad 1 \quad 1 \quad 1}^{\beta_{0j}\text{ coefficients}} \\ \underbrace{0 \quad 1 \quad 2 \quad 3 \quad 4}_{\beta_{1j}\text{ coefficients}} \end{bmatrix} \qquad (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix}$$

PANEL 7.2 SAS ANALYSIS FOR EXAMPLE 7.8 USING PROC. GLM. area of interest

```
title 'Multivariate Regression Analysis';
data mra;
infile 'Example 7-8 data;
input y1 y2 z1;
proc glm data = mra;
model y1 y2 = z1/ss3;
manova h = z1/printe;
```

PROGRAM COMMANDS

General Linear Models Procedure

Dependent Variable: Y1

OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40.00000000	40.00000000	20.00	0.0208
Error	3	6.00000000	2.00000000		
Corrected Total	4	46.00000000			

R-Square	C.V.	Root MSE	Y1 Mean
0.869565	28.28427	1.414214	5.00000000

(continues on next page)

PANEL 7.2 *(continued)*

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Z1	1	40.00000000	40.00000000	20.00	0.0208

Parameter	Estimate	T for H0: Parameter = 0	Pr > ITI	Std Error of Estimate
INTERCEPT	1.000000000	0.91	0.4286	1.09544512
Z1	2.000000000	4.47	0.0208	0.44721360

Dependent Variable: Y2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.00000000	10.00000000	7.50	0.0714
Error	3	4.00000000	1.33333333		
Corrected Total	4	14.00000000			

R-Square	C.V.	Root MSE	Y2 Mean
0.714286	115.4701	1.154701	1.00000000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Z1	1	10.00000000	10.00000000	7.50	0.0714

Parameter	Estimate	T for H0: Parameter = 0	Pr > ITI	Std Error of Estimate
INTERCEPT	−1.000000000	−1.12	0.3450	0.89442719
Z1	1.000000000	2.74	0.0714	0.36514837

E = Error SS & CP Matrix

	Y1	Y2
Y1	6	−2
Y2	−2	4

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall Z1 Effect
H = Type III SS&CP Matrix for Z1 E = Error SS&CP Matrix
S = 1 M = 0 N = 0

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.06250000	15.0000	2	2	0.0625
Pillai's Trace	0.93750000	15.0000	2	2	0.0625
Hotelling-Lawley Trace	15.00000000	15.0000	2	2	0.0625
Roy's Greatest Root	15.00000000	15.0000	2	2	0.0625

and

$$\mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ 2 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 20 \end{bmatrix}$$

so

$$\hat{\boldsymbol{\beta}}_{(2)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix} \begin{bmatrix} 5 \\ 20 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

From Example 7.3,

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(1)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Hence,

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)}] = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[\mathbf{y}_{(1)} \mid \mathbf{y}_{(2)}]$$

The fitted values are generated from $\hat{y}_1 = 1 + 2z_1$ and $\hat{y}_2 = -1 + z_2$. Collectively,

$$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix}$$

and

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{bmatrix}'$$

Note that

$$\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{Y}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Since

$$\mathbf{Y}'\mathbf{Y} = \begin{bmatrix} 1 & 4 & 3 & 8 & 9 \\ -1 & -1 & 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix} = \begin{bmatrix} 171 & 43 \\ 43 & 19 \end{bmatrix}$$

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \begin{bmatrix} 165 & 45 \\ 45 & 15 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}$$

the sum of squares and cross-products decomposition

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

is easily verified. ■

Result 7.9. For the least squares estimator $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\beta}}_{(m)}]$ determined under the multivariate multiple regression model (7-23) with full rank $(\mathbf{Z}) = r + 1 < n$,

$$E(\hat{\boldsymbol{\beta}}_{(i)}) = \boldsymbol{\beta}_{(i)} \quad \text{or} \quad E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

and

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}, \qquad i, k = 1, 2, \ldots, m$$

The residuals $\hat{\boldsymbol{\varepsilon}} = [\hat{\boldsymbol{\varepsilon}}_{(1)} \mid \hat{\boldsymbol{\varepsilon}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\varepsilon}}_{(m)}] = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$ satisfy $E(\hat{\boldsymbol{\varepsilon}}_{(i)}) = \mathbf{0}$ and $E(\hat{\boldsymbol{\varepsilon}}'_{(i)}\hat{\boldsymbol{\varepsilon}}_{(k)}) = (n - r - 1)\sigma_{ik}$, so

$$E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \quad \text{and} \quad E\left(\frac{1}{n - r - 1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = \boldsymbol{\Sigma}$$

Also, $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated.

Proof. The ith response follows the multiple regression model

$$\mathbf{Y}_{(i)} = \mathbf{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \quad E(\boldsymbol{\varepsilon}_{(i)}) = \mathbf{0}, \quad \text{and} \quad E(\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}'_{(i)}) = \sigma_{ii}\mathbf{I}$$

Also, from (7-24),

$$\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}_{(i)} - \boldsymbol{\beta}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}_{(i)} \tag{7-33}$$

and

$$\hat{\boldsymbol{\varepsilon}}_{(i)} = \mathbf{Y}_{(i)} - \hat{\mathbf{Y}}_{(i)} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y}_{(i)} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\boldsymbol{\varepsilon}_{(i)}$$

so $E(\hat{\boldsymbol{\beta}}_{(i)}) = \boldsymbol{\beta}_{(i)}$ and $E(\hat{\boldsymbol{\varepsilon}}_{(i)}) = \mathbf{0}$.

Next,

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) &= E(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})' \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E(\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}'_{(k)})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}\end{aligned}$$

Using Result 4.9, with $\mathbf{U}$ any random vector and $\mathbf{A}$ a fixed matrix, we have that $E[\mathbf{U}'\mathbf{A}\mathbf{U}] = E[\text{tr}(\mathbf{A}\mathbf{U}\mathbf{U}')] = \text{tr}[\mathbf{A}E(\mathbf{U}\mathbf{U}')]$. Consequently, from the proof of Result 7.1 and using Result 2A.12

$$\begin{aligned}E(\hat{\boldsymbol{\varepsilon}}'_{(i)}\hat{\boldsymbol{\varepsilon}}_{(k)}) &= E(\boldsymbol{\varepsilon}'_{(i)}(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\boldsymbol{\varepsilon}_{(k)}) = \text{tr}[(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\sigma_{ik}\mathbf{I}] \\ &= \sigma_{ik}\,\text{tr}[(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')] = \sigma_{ik}(n - r - 1)\end{aligned}$$

Dividing each entry $\hat{\boldsymbol{\varepsilon}}_{(i)}'\hat{\boldsymbol{\varepsilon}}_{(k)}$ of $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ by $n - r - 1$, we obtain the unbiased estimator of $\boldsymbol{\Sigma}$. Finally,

$$\begin{aligned}\operatorname{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\varepsilon}}_{(k)}) &= E[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(k)}'(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')] \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E(\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(k)}')(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\sigma_{ik}\mathbf{I}(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \\ &= \sigma_{ik}((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' - (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') = \mathbf{0}\end{aligned}$$

so each element of $\hat{\boldsymbol{\beta}}$ is uncorrelated with each element of $\hat{\boldsymbol{\varepsilon}}$. ■

The mean vectors and covariance matrices determined in Result 7.9 enable us to obtain the sampling properties of the least squares predictors.

We first consider the problem of estimating the mean vector when the predictor variables have the values $\mathbf{z}_0' = [1, z_{01}, \ldots, z_{0r}]$. The mean of the ith response variable is $\mathbf{z}_0'\boldsymbol{\beta}_{(i)}$, and this is estimated by $\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$, the ith component of the fitted regression relationship. Collectively,

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} = [\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} \mid \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} \mid \cdots \mid \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(m)}] \tag{7-34}$$

is an unbiased estimator $\mathbf{z}_0'\boldsymbol{\beta}$ since $E(\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{z}_0'\boldsymbol{\beta}_{(i)}$ for each component. From the covariance matrix for $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}_{(k)}$, the estimation errors $\mathbf{z}_0'\boldsymbol{\beta}_{(i)} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$ have covariances

$$\begin{aligned}E[\mathbf{z}_0'(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)})(\boldsymbol{\beta}_{(k)} - \hat{\boldsymbol{\beta}}_{(k)})'\mathbf{z}_0] &= \mathbf{z}_0'(E(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)})(\boldsymbol{\beta}_{(k)} - \hat{\boldsymbol{\beta}}_{(k)})')\mathbf{z}_0 \\ &= \sigma_{ik}\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 \end{aligned}\tag{7-35}$$

The related problem is that of forecasting a new observation vector $\mathbf{Y}_0' = [Y_{01}, Y_{02}, \ldots, Y_{0m}]$ at $\mathbf{z}_0$. According to the regression model, $Y_{0i} = \mathbf{z}_0'\boldsymbol{\beta}_{(i)} + \varepsilon_{0i}$ where the "new" error $\boldsymbol{\varepsilon}_0' = [\varepsilon_{01}, \varepsilon_{02}, \ldots, \varepsilon_{0m}]$ is independent of the errors $\boldsymbol{\varepsilon}$ and satisfies $E(\varepsilon_{0i}) = 0$ and $E(\varepsilon_{0i}\varepsilon_{0k}) = \sigma_{ik}$. The *forecast error* for the ith component of $\mathbf{Y}_0$ is

$$\begin{aligned}Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} &= Y_{0i} - \mathbf{z}_0'\boldsymbol{\beta}_{(i)} + \mathbf{z}_0'\boldsymbol{\beta}_{(i)} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \\ &= \varepsilon_{0i} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\end{aligned}$$

so $E(Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}) = E(\varepsilon_{0i}) - \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}) = 0$, indicating that $\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$ is an *unbiased predictor* of Y_{0i}. The forecast errors have covariances

$$\begin{aligned}&E(Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)})(Y_{0k} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(k)}) \\ &\quad = E(\varepsilon_{0i} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}))(\varepsilon_{0k} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})) \\ &\quad = E(\varepsilon_{0i}\varepsilon_{0k}) + \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})'\mathbf{z}_0 \\ &\qquad - \mathbf{z}_0'E((\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\varepsilon_{0k}) - E(\varepsilon_{0i}(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})')\mathbf{z}_0 \\ &\quad = \sigma_{ik}(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)\end{aligned}\tag{7-36}$$

Note that $E((\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\varepsilon_{0k}) = \mathbf{0}$ since $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}_{(i)} + \boldsymbol{\beta}_{(i)}$ is independent of $\boldsymbol{\varepsilon}_0$. A similar result holds for $E(\varepsilon_{0i}(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})')$.

Maximum likelihood estimators and their distributions can be obtained when the errors $\boldsymbol{\varepsilon}$ have a normal distribution.

Result 7.10. Let the multivariate multiple regression model in (7-23) hold with full rank $(\mathbf{Z}) = r + 1$, $n \geq (r + 1) + m$, and let the errors $\boldsymbol{\varepsilon}$ have a normal distribution. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$

is the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ has a normal distribution with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}$. Also, $\hat{\boldsymbol{\beta}}$ is independent of the maximum likelihood estimator of the positive definite $\boldsymbol{\Sigma}$ given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$$

and

$$n\hat{\boldsymbol{\Sigma}} \quad \text{is distributed as} \quad W_{p,n-r-1}(\boldsymbol{\Sigma})$$

The maximized likelihood $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (2\pi)^{-mn/2}|\hat{\boldsymbol{\Sigma}}|^{-n/2}e^{-mn/2}$.

Proof. (See website: www.prenhall.com/statistics) ■

Result 7.10 provides additional support for using least squares estimates. When the errors are normally distributed, $\hat{\boldsymbol{\beta}}$ and $n^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, respectively. Therefore, for large samples, they have nearly the smallest possible variances.

Comment. The multivariate multiple regression model poses no new computational problems. Least squares (maximum likelihood) estimates, $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(i)}$, are computed individually for each response variable. Note, however, that the model requires that the *same* predictor variables be used for all responses.

Once a multivariate multiple regression model has been fit to the data, it should be subjected to the diagnostic checks described in Section 7.6 for the single-response model. The residual vectors $[\hat{\varepsilon}_{j1}, \hat{\varepsilon}_{j2}, \ldots, \hat{\varepsilon}_{jm}]$ can be examined for normality or outliers using the techniques in Section 4.6.

The remainder of this section is devoted to brief discussions of inference for the normal theory multivariate multiple regression model. Extended accounts of these procedures appear in [2] and [18].

Likelihood Ratio Tests for Regression Parameters

The multiresponse analog of (7-12), the hypothesis that the responses do not depend on $\underbrace{z_{q+1}, z_{q+2}, \ldots, z_r}$, becomes

given $z_0, \ldots, z_q$ are false ; q can be adjusted as we need for the est

$$H_0: \underbrace{\boldsymbol{\beta}_{(2)} = \mathbf{0}}_{\text{lower matrix}} \quad \text{where} \quad \boldsymbol{\beta} = \begin{bmatrix} \underset{((q+1)\times m)}{\boldsymbol{\beta}_{(1)}} \\ \hdashline \underset{((r-q)\times m)}{\boldsymbol{\beta}_{(2)}} \end{bmatrix} \tag{7-37}$$

Setting $\mathbf{Z} = \left[\underset{(n\times(q+1))}{\mathbf{Z}_1} \;\vdots\; \underset{(n\times(r-q))}{\mathbf{Z}_2}\right]$, we can write the general model as

$$E(\mathbf{Y}) = \mathbf{Z}\boldsymbol{\beta} = [\mathbf{Z}_1 \;\vdots\; \mathbf{Z}_2]\begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \hdashline \boldsymbol{\beta}_{(2)} \end{bmatrix} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)}$$

Under H_0: $\boldsymbol{\beta}_{(2)} = \mathbf{0}$, $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$ and the likelihood ratio test of H_0 is based on the quantities involved in the

extra sum of squares and cross products

$$= (\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$$
$$= n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$$

where $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$ and $\hat{\boldsymbol{\Sigma}}_1 = n^{-1}(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})$.

From Result 7.10, the likelihood ratio, Λ, can be expressed in terms of generalized variances:

$$\Lambda = \frac{\max\limits_{\boldsymbol{\beta}_{(1)},\boldsymbol{\Sigma}} L(\boldsymbol{\beta}_{(1)},\boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\beta},\boldsymbol{\Sigma}} L(\boldsymbol{\beta},\boldsymbol{\Sigma})} = \frac{L(\hat{\boldsymbol{\beta}}_{(1)},\hat{\boldsymbol{\Sigma}}_1)}{L(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)^{n/2} \tag{7-38}$$

Equivalently, *Wilks' lambda statistic*

$$\Lambda^{2/n} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}$$

can be used.

Result 7.11. Let the multivariate multiple regression model of (7-23) hold with $\mathbf{Z}$ of full rank $r + 1$ and $(r + 1) + m \le n$. Let the errors $\boldsymbol{\varepsilon}$ be normally distributed. Under H_0: $\boldsymbol{\beta}_{(2)} = \mathbf{0}$, $n\hat{\boldsymbol{\Sigma}}$ is distributed as $W_{p,n-r-1}(\boldsymbol{\Sigma})$ independently of $n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$ which, in turn, is distributed as $W_{p,r-q}(\boldsymbol{\Sigma})$. The likelihood ratio test of H_0 is equivalent to rejecting H_0 for large values of

$$-2\ln\Lambda = -n\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right) = -n\ln\frac{|n\hat{\boldsymbol{\Sigma}}|}{|n\hat{\boldsymbol{\Sigma}} + n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})|}$$

For n large,[5] the modified statistic

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right]\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)$$

has, to a close approximation, a chi-square distribution with $m(r - q)$ d.f.

Proof. (See Supplement 7A.) ■

If $\mathbf{Z}$ is not of full rank, but has rank $r_1 + 1$, then $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'\mathbf{Y}$, where $(\mathbf{Z}'\mathbf{Z})^{-}$ is the *generalized inverse* discussed in [22]. (See also Exercise 7.6.) The distributional conclusions stated in Result 7.11 remain the same, provided that r is replaced by r_1 and $q + 1$ by rank $(\mathbf{Z}_1)$. However, not all hypotheses concerning $\boldsymbol{\beta}$ can be tested due to the lack of uniqueness in the identification of $\boldsymbol{\beta}$ caused by the linear dependencies among the columns of $\mathbf{Z}$. Nevertheless, the generalized inverse allows all of the important MANOVA models to be analyzed as special cases of the multivariate multiple regression model.

[5] Technically, both $n - r$ and $n - m$ should also be large to obtain a good chi-square approximation.

figure out how r_1 & q_1 interact.

Example 7.9 (Testing the importance of additional predictors with a multivariate response) The service in three locations of a large restaurant chain was rated
→ predictors
according to two measures of quality by male and female patrons. The first service-quality index was introduced in Example 7.5. Suppose we consider a regression model that allows for the effects of location, gender, and the location–gender interaction on both service-quality indices. The design matrix (see Example 7.5) remains the same for the two-response situation. We shall illustrate the test of no location-gender interaction in either response using Result 7.11. A computer program provides

extra variables

$$\begin{pmatrix}\text{residual sum of squares}\\ \text{and cross products}\end{pmatrix} = n\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 2977.39 & 1021.72 \\ 1021.72 & 2050.95 \end{bmatrix}$$

$$\begin{pmatrix}\text{extra sum of squares}\\ \text{and cross products}\end{pmatrix} = n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}) = \begin{bmatrix} 441.76 & 246.16 \\ 246.16 & 366.12 \end{bmatrix}$$

Let $\boldsymbol{\beta}_{(2)}$ be the matrix of interaction parameters for the two responses. Although the sample size $n = 18$ is not large, we shall illustrate the calculations involved in the test of $H_0\colon \boldsymbol{\beta}_{(2)} = \mathbf{0}$ given in Result 7.11. Setting $\alpha = .05$, we test H_0 by referring

no. of response variables

no. of predictors

$$-\left[n - r_1 - 1 - \frac{1}{2}(m - r_1 + q_1 + 1)\right]\ln\left(\frac{|n\hat{\boldsymbol{\Sigma}}|}{|n\hat{\boldsymbol{\Sigma}} + n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})|}\right)$$

what are r_1 & q_1?

$$= -\left[18 - 5 - 1 - \frac{1}{2}(2 - 5 + 3 + 1)\right]\ln(.7605) = 3.28$$

$3+2=5$

to a chi-square percentage point with $m(r_1 - q_1) = 2(2) = 4$ d.f. Since $3.28 < \chi^2_4(.05) = 9.49$, we do not reject H_0 at the 5% level. The interaction terms are not needed. ■

Information criterion are also available to aid in the selection of a simple but adequate multivariate multiple regresson model. For a model that includes d predictor variables counting the intercept, let

$$\hat{\boldsymbol{\Sigma}}_d = \frac{1}{n}(\text{residual sum of squares and cross products matrix})$$

Then, the multivariate multiple regression version of the Akaike's information criterion is

$$\text{AIC} = n\ln(|\hat{\boldsymbol{\Sigma}}_d|) - 2p \times d$$

This criterion attempts to balance the generalized variance with the number of parameters. Models with smaller AIC's are preferable.

In the context of Example 7.9, under the null hypothesis of no interaction terms, we have $n = 18$, $p = 2$ response variables, and $d = 4$ terms, so

$$\text{AIC} = n\ln(|\boldsymbol{\Sigma}|) - 2p \times d = 18\ln\left(\left|\frac{1}{18}\begin{bmatrix} 3419.15 & 1267.88 \\ 1267.88 & 2417.07 \end{bmatrix}\right|\right) - 2 \times 2 \times 4$$
$$= 18 \times \ln(20545.7) - 16 = 162.75$$

More generally, we could consider a null hypothesis of the form $H_0\colon \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\Gamma}_0$, where $\mathbf{C}$ is $(r - q) \times (r + 1)$ and is of full rank $(r - q)$. For the choices

$\mathbf{C} = \left[\mathbf{0} \;\vdots\; \underset{(r-q)\times(r-q)}{\mathbf{I}}\right]$ and $\boldsymbol{\Gamma}_0 = \mathbf{0}$, this null hypothesis becomes $H_0\colon \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\beta}_{(2)} = \mathbf{0}$, the case considered earlier. It can be shown that the extra sum of squares and cross products generated by the hypothesis H_0 is

$$n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}) = (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\Gamma}_0)'(\mathbf{C}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\Gamma}_0)$$

Under the null hypothesis, the statistic $n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$ is distributed as $W_{r-q}(\boldsymbol{\Sigma})$ independently of $\hat{\boldsymbol{\Sigma}}$. This distribution theory can be employed to develop a test of $H_0\colon \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\Gamma}_0$ similar to the test discussed in Result 7.11. (See, for example, [18].)

Other Multivariate Test Statistics

Tests other than the likelihood ratio test have been proposed for testing $H_0\colon \boldsymbol{\beta}_{(2)} = \mathbf{0}$ in the multivariate multiple regression model.

Popular computer-package programs routinely calculate four multivariate test statistics. To connect with their output, we introduce some alternative notation. Let $\mathbf{E}$ be the $p \times p$ *error*, or residual, sum of squares and cross products matrix

$$\mathbf{E} = n\hat{\boldsymbol{\Sigma}}$$

that results from fitting the full model. The $p \times p$ *hypothesis*, or extra, sum of squares and cross-products matrix

$$\mathbf{H} = n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$$

The statistics can be defined in terms of $\mathbf{E}$ and $\mathbf{H}$ directly, or in terms of the nonzero eigenvalues $\eta_1 \ge \eta_2 \ge \dots \ge \eta_s$ of $\mathbf{HE}^{-1}$, where $s = \min(p, r-q)$. Equivalently, they are the roots of $|(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}) - \eta\hat{\boldsymbol{\Sigma}}| = 0$. The definitions are

$$\text{Wilks' lambda} = \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$$

$$\text{Pillai's trace} = \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \operatorname{tr}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}]$$

$$\text{Hotelling–Lawley trace} = \sum_{i=1}^{s} \eta_i = \operatorname{tr}[\mathbf{HE}^{-1}]$$

$$\text{Roy's greatest root} = \frac{\eta_1}{1+\eta_1}$$

Roy's test selects the coefficient vector $\mathbf{a}$ so that the univariate F-statistic based on a $\mathbf{a}'\mathbf{Y}_j$ has its maximum possible value. When several of the eigenvalues η_i are moderately large, Roy's test will perform poorly relative to the other three. Simulation studies suggest that its power will be best when there is only one large eigenvalue.

Charts and tables of critical values are available for Roy's test. (See [21] and [17].) Wilks' lambda, Roy's greatest root, and the Hotelling–Lawley trace test are nearly equivalent for large sample sizes.

If there is a large discrepancy in the reported P-values for the four tests, the eigenvalues and vectors may lead to an interpretation. In this text, we report Wilks' lambda, which is the likelihood ratio test.

Area of interest!

Predictions from Multivariate Multiple Regressions

Suppose the model $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with normal errors $\boldsymbol{\varepsilon}$, has been fit and checked for any inadequacies. If the model is adequate, it can be employed for predictive purposes.

One problem is to predict the mean responses corresponding to fixed values $\mathbf{z}_0$ of the predictor variables. Inferences about the mean responses can be made using the distribution theory in Result 7.10. From this result, we determine that

$$\hat{\boldsymbol{\beta}}'\mathbf{z}_0 \quad \text{is distributed as} \quad N_m(\boldsymbol{\beta}'\mathbf{z}_0, \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\,\boldsymbol{\Sigma})$$

and

$$n\hat{\boldsymbol{\Sigma}} \quad \text{is independently distributed as} \quad W_{n-r-1}(\boldsymbol{\Sigma})$$

The unknown value of the regression function at $\mathbf{z}_0$ is $\boldsymbol{\beta}'\mathbf{z}_0$. So, from the discussion of the T^2-statistic in Section 5.2, we can write

$$T^2 = \left(\frac{\hat{\boldsymbol{\beta}}'\mathbf{z}_0 - \boldsymbol{\beta}'\mathbf{z}_0}{\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0}}\right)' \left(\frac{n}{n-r-1}\hat{\boldsymbol{\Sigma}}\right)^{-1} \left(\frac{\hat{\boldsymbol{\beta}}'\mathbf{z}_0 - \boldsymbol{\beta}'\mathbf{z}_0}{\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0}}\right) \tag{7-39}$$

and the $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}'\mathbf{z}_0$ is provided by the inequality

$$(\boldsymbol{\beta}'\mathbf{z}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0)'\left(\frac{n}{n-r-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\beta}'\mathbf{z}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0) \le \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\left[\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)\right] \tag{7-40}$$

where $F_{m,n-r-m}(\alpha)$ is the upper (100α)th percentile of an F-distribution with m and $n-r-m$ d.f.

The $100(1-\alpha)\%$ *simultaneous* confidence intervals for $E(Y_i) = \mathbf{z}_0'\boldsymbol{\beta}_{(i)}$ are

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)}\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\left(\frac{n}{n-r-1}\hat{\sigma}_{ii}\right)}, \qquad i = 1,2,\ldots,m \tag{7-41}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the ith column of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_{ii}$ is the ith diagonal element of $\hat{\boldsymbol{\Sigma}}$.

Stick with these in terms of showing reliability of estimates

The second prediction problem is concerned with forecasting new responses $\mathbf{Y}_0 = \boldsymbol{\beta}'\mathbf{z}_0 + \boldsymbol{\varepsilon}_0$ at $\mathbf{z}_0$. Here $\boldsymbol{\varepsilon}_0$ is independent of $\boldsymbol{\varepsilon}$. Now,

$$\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{z}_0 + \boldsymbol{\varepsilon}_0 \quad \text{is distributed as} \quad N_m(\mathbf{0}, (1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)\boldsymbol{\Sigma})$$

independently of $n\hat{\boldsymbol{\Sigma}}$, so the $100(1-\alpha)\%$ *prediction ellipsoid* for $\mathbf{Y}_0$ becomes

$$(\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0)'\left(\frac{n}{n-r-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0) \le (1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)\left[\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)\right] \tag{7-42}$$

The $100(1-\alpha)\%$ *simultaneous prediction intervals* for the individual responses Y_{0i} are

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)}\sqrt{(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)\left(\frac{n}{n-r-1}\hat{\sigma}_{ii}\right)}, \qquad i = 1,2,\ldots,m \tag{7-43}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$, $\hat{\sigma}_{ii}$, and $F_{m,n-r-m}(\alpha)$ are the same quantities appearing in (7-41). Comparing (7-41) and (7-43), we see that the prediction intervals for the *actual* values of the response variables are wider than the corresponding intervals for the *expected* values. The extra width reflects the presence of the random error ε_{0i}.

Example 7.10 (Constructing a confidence ellipse and a prediction ellipse for bivariate responses) A second response variable was measured for the computer-requirement problem discussed in Example 7.6. Measurements on the response Y_2, disk input/output capacity, corresponding to the z_1 and z_2 values in that example were

$$\mathbf{y}_2' = [301.8, 396.1, 328.2, 307.4, 362.4, 369.5, 229.1]$$

Obtain the 95% confidence ellipse for $\boldsymbol{\beta}'\mathbf{z}_0$ and the 95% prediction ellipse for $\mathbf{Y}_0' = [Y_{01}, Y_{02}]$ for a site with the configuration $\mathbf{z}_0' = [1, 130, 7.5]$.

Computer calculations provide the fitted equation

$$\hat{y}_2 = 14.14 + 2.25z_1 + 5.67z_2$$

with $s = 1.812$. Thus, $\hat{\boldsymbol{\beta}}_{(2)}' = [14.14, 2.25, 5.67]$. From Example 7.6,

$$\hat{\boldsymbol{\beta}}_{(1)}' = [8.42, 1.08, 42], \quad \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} = 151.97, \quad \text{and} \quad \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 = .34725$$

We find that

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} = 14.14 + 2.25(130) + 5.67(7.5) = 349.17$$

and

$$n\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} (\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)}) & (\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)}) \\ (\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)})'(\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)}) & (\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)})'(\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)}) \end{bmatrix}$$

$$= \begin{bmatrix} 5.80 & 5.30 \\ 5.30 & 13.13 \end{bmatrix}$$

Since

$$\hat{\boldsymbol{\beta}}'\mathbf{z}_0 = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{(1)}' \\ \hdashline \hat{\boldsymbol{\beta}}_{(2)}' \end{bmatrix} \mathbf{z}_0 = \begin{bmatrix} \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} \\ \hdashline \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} \end{bmatrix} = \begin{bmatrix} 151.97 \\ 349.17 \end{bmatrix}$$

$n = 7$, $r = 2$, and $m = 2$, a 95% confidence ellipse for $\boldsymbol{\beta}'\mathbf{z}_0 = \begin{bmatrix} \mathbf{z}_0'\boldsymbol{\beta}_{(1)} \\ \hdashline \mathbf{z}_0'\boldsymbol{\beta}_{(2)} \end{bmatrix}$ is, from (7-40), the set

$$[\mathbf{z}_0'\boldsymbol{\beta}_{(1)} - 151.97, \mathbf{z}_0'\boldsymbol{\beta}_{(2)} - 349.17](4) \begin{bmatrix} 5.80 & 5.30 \\ 5.30 & 13.13 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{z}_0'\boldsymbol{\beta}_{(1)} - 151.97 \\ \mathbf{z}_0'\boldsymbol{\beta}_{(2)} - 349.17 \end{bmatrix}$$

$$\leq (.34725)\left[\left(\frac{2(4)}{3}\right)F_{2,3}(.05)\right]$$

with $F_{2,3}(.05) = 9.55$. This ellipse is centered at $(151.97, 349.17)$. Its orientation and the lengths of the major and minor axes can be determined from the eigenvalues and eigenvectors of $n\hat{\boldsymbol{\Sigma}}$.

Comparing (7-40) and (7-42), we see that the only change required for the calculation of the 95% prediction ellipse is to replace $\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 = .34725$ with

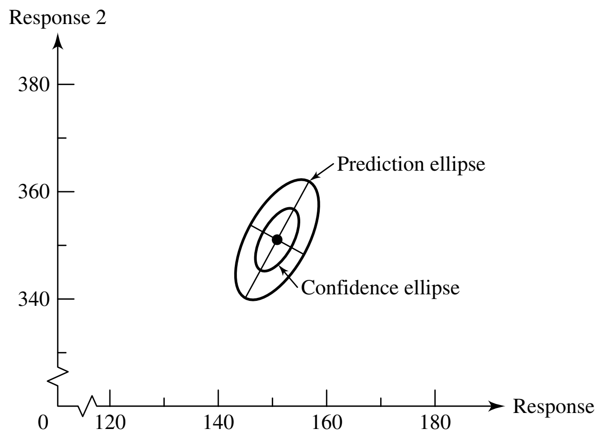

Figure 7.5 95% confidence and prediction ellipses for the computer data with two responses.

$1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 = 1.34725$. Thus, the 95% prediction ellipse for $\mathbf{Y}_0' = [Y_{01}, Y_{02}]$ is also centered at $(151.97, 349.17)$, but is larger than the confidence ellipse. Both ellipses are sketched in Figure 7.5.

It is the *prediction* ellipse that is relevant to the determination of computer requirements for a particular site with the given $\mathbf{z}_0$. ■

7.8 The Concept of Linear Regression

The classical linear regression model is concerned with the association between a single dependent variable Y and a collection of predictor variables $z_1, z_2, \ldots, z_r$. The regression model that we have considered treats Y as a random variable whose mean depends upon *fixed* values of the z_i's. This mean is assumed to be a linear function of the regression *coefficients* $\beta_0, \beta_1, \ldots, \beta_r$.

The linear regression model also arises in a different setting. Suppose all the variables $Y, Z_1, Z_2, \ldots, Z_r$ are random and have a joint distribution, not necessarily normal, with mean vector $\underset{(r+1)\times 1}{\boldsymbol{\mu}}$ and covariance matrix $\underset{(r+1)\times(r+1)}{\boldsymbol{\Sigma}}$. Partitioning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in an obvious fashion, we write

$$\boldsymbol{\mu} = \left[\begin{array}{c} \underset{(1\times 1)}{\mu_Y} \\ \hdashline \underset{(r\times 1)}{\boldsymbol{\mu}_{\mathbf{Z}}} \end{array}\right] \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c:c} \underset{(1\times 1)}{\sigma_{YY}} & \underset{(1\times r)}{\boldsymbol{\sigma}_{\mathbf{Z}Y}'} \\ \hdashline \underset{(r\times 1)}{\boldsymbol{\sigma}_{\mathbf{Z}Y}} & \underset{(r\times r)}{\boldsymbol{\Sigma}_{\mathbf{ZZ}}} \end{array}\right]$$

with

$$\boldsymbol{\sigma}_{\mathbf{Z}Y}' = [\sigma_{YZ_1}, \sigma_{YZ_2}, \ldots, \sigma_{YZ_r}] \tag{7-44}$$

$\boldsymbol{\Sigma}_{\mathbf{ZZ}}$ can be taken to have full rank.[6] Consider the problem of predicting Y using the

$$\textit{linear predictor} = b_0 + b_1 Z_1 + \cdots + b_r Z_r = b_0 + \mathbf{b}'\mathbf{Z} \tag{7-45}$$

[6] If $\boldsymbol{\Sigma}_{\mathbf{ZZ}}$ is not of full rank, one variable—for example, Z_k—can be written as a linear combination of the other Z_i's and thus is redundant in forming the linear regression function $\mathbf{Z}'\boldsymbol{\beta}$. That is, $\mathbf{Z}$ may be replaced by any subset of components whose nonsingular covariance matrix has the same rank as $\boldsymbol{\Sigma}_{\mathbf{ZZ}}$.

For a given predictor of the form of (7-45), the error in the prediction of Y is

$$\textit{prediction error} = Y - b_0 - b_1 Z_1 - \cdots - b_r Z_r = Y - b_0 - \mathbf{b}'\mathbf{Z} \tag{7-46}$$

Because this error is random, it is customary to select b_0 and $\mathbf{b}$ to minimize the

$$\textit{mean square error} = E(Y - b_0 - \mathbf{b}'\mathbf{Z})^2 \tag{7-47}$$

Now the mean square error depends on the joint distribution of Y and $\mathbf{Z}$ only through the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It is possible to express the "optimal" linear predictor in terms of these latter quantities.

Result 7.12. The linear predictor $\beta_0 + \boldsymbol{\beta}'\mathbf{Z}$ with coefficients

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}, \qquad \beta_0 = \mu_Y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{Z}}$$

has minimum mean square among all *linear* predictors of the response Y. Its mean square error is

$$E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z})^2 = E(Y - \mu_Y - \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}))^2 = \sigma_{YY} - \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}$$

Also, $\beta_0 + \boldsymbol{\beta}'\mathbf{Z} = \mu_Y + \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})$ is the linear predictor having maximum correlation with Y; that is,

$$\begin{aligned}\operatorname{Corr}(Y, \beta_0 + \boldsymbol{\beta}'\mathbf{Z}) &= \max_{b_0, \mathbf{b}} \operatorname{Corr}(Y, b_0 + \mathbf{b}'\mathbf{Z}) \\ &= \sqrt{\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\beta}}{\sigma_{YY}}} = \sqrt{\frac{\boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}}\end{aligned}$$

Proof. Writing $b_0 + \mathbf{b}'\mathbf{Z} = b_0 + \mathbf{b}'\mathbf{Z} + (\mu_Y - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}}) - (\mu_Y - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}})$, we get

$$\begin{aligned}E(Y - b_0 - \mathbf{b}'\mathbf{Z})^2 &= E[Y - \mu_Y - (\mathbf{b}'\mathbf{Z} - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}}) + (\mu_Y - b_0 - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}})]^2 \\ &= E(Y - \mu_Y)^2 + E(\mathbf{b}'(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}))^2 + (\mu_Y - b_0 - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}})^2 \\ &\qquad - 2E[\mathbf{b}'(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})(Y - \mu_Y)] \\ &= \sigma_{YY} + \mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\mathbf{b} + (\mu_Y - b_0 - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}})^2 - 2\mathbf{b}'\boldsymbol{\sigma}_{\mathbf{Z}Y}\end{aligned}$$

Adding and subtracting $\boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}$, we obtain

$$\begin{aligned}E(Y - b_0 - \mathbf{b}'\mathbf{Z})^2 &= \sigma_{YY} - \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y} + (\mu_Y - b_0 - \mathbf{b}'\boldsymbol{\mu}_{\mathbf{Z}})^2 \\ &\qquad + (\mathbf{b} - \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y})'\boldsymbol{\Sigma}_{\mathbf{ZZ}}(\mathbf{b} - \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y})\end{aligned}$$

The mean square error is minimized by taking $\mathbf{b} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \boldsymbol{\beta}$, making the last term zero, and then choosing $b_0 = \mu_Y - (\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y})'\boldsymbol{\mu}_{\mathbf{Z}} = \beta_0$ to make the third term zero. The minimum mean square error is thus $\sigma_{YY} - \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}$.

Next, we note that $\operatorname{Cov}(b_0 + \mathbf{b}'\mathbf{Z}, Y) = \operatorname{Cov}(\mathbf{b}'\mathbf{Z}, Y) = \mathbf{b}'\boldsymbol{\sigma}_{\mathbf{Z}Y}$ so

$$[\operatorname{Corr}(b_0 + \mathbf{b}'\mathbf{Z}, Y)]^2 = \frac{[\mathbf{b}'\boldsymbol{\sigma}_{\mathbf{Z}Y}]^2}{\sigma_{YY}(\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\mathbf{b})}, \qquad \text{for all } b_0, \mathbf{b}$$

Employing the extended Cauchy–Schwartz inequality of (2-49) with $\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}$, we obtain

$$(\mathbf{b}'\boldsymbol{\sigma}_{\mathbf{Z}Y})^2 \le \mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\mathbf{b}\,\boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}$$

or

$$[\text{Corr}(b_0 + \mathbf{b}'\mathbf{Z}, Y)]^2 \le \frac{\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}$$

with equality for $\mathbf{b} = \boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \boldsymbol{\beta}$. The alternative expression for the maximum correlation follows from the equation $\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\beta} = \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\beta}$. ■

The correlation between Y and its best linear predictor is called the *population multiple correlation coefficient*

$$\rho_{Y(\mathbf{Z})} = +\sqrt{\frac{\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}} \tag{7-48}$$

The square of the population multiple correlation coefficient, $\rho^2_{Y(\mathbf{Z})}$, is called the *population coefficient of determination*. Note that, unlike other correlation coefficients, the multiple correlation coefficient is a *positive* square root, so $0 \le \rho_{Y(\mathbf{Z})} \le 1$.

The population coefficient of determination has an important interpretation. From Result 7.12, the mean square error in using $\beta_0 + \boldsymbol{\beta}'\mathbf{Z}$ to forecast Y is

$$\sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \sigma_{YY} - \sigma_{YY}\left(\frac{\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}\right) = \sigma_{YY}(1 - \rho^2_{Y(\mathbf{Z})}) \tag{7-49}$$

If $\rho^2_{Y(\mathbf{Z})} = 0$, there is no predictive power in $\mathbf{Z}$. At the other extreme, $\rho^2_{Y(\mathbf{Z})} = 1$ implies that Y can be predicted with no error.

Example 7.11 (Determining the best linear predictor, its mean square error, and the multiple correlation coefficient) Given the mean vector and covariance matrix of Y, Z_1, Z_2,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \hline \boldsymbol{\mu}_{\mathbf{Z}} \end{bmatrix} = \begin{bmatrix} 5 \\ \hline 2 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \sigma_{YY} & \boldsymbol{\sigma}'_{\mathbf{Z}Y} \\ \hline \boldsymbol{\sigma}_{\mathbf{Z}Y} & \boldsymbol{\Sigma}_{\mathbf{ZZ}} \end{array}\right] = \left[\begin{array}{r|rr} 10 & 1 & -1 \\ \hline 1 & 7 & 3 \\ -1 & 3 & 2 \end{array}\right]$$

determine (a) the best linear predictor $\beta_0 + \beta_1 Z_1 + \beta_2 Z_2$, (b) its mean square error, and (c) the multiple correlation coefficient. Also, verify that the mean square error equals $\sigma_{YY}(1 - \rho^2_{Y(\mathbf{Z})})$.

First,

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} .4 & -.6 \\ -.6 & 1.4 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$\beta_0 = \mu_Y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{Z}} = 5 - [1, -2]\begin{bmatrix} 2 \\ 0 \end{bmatrix} = 3$$

so the best linear predictor is $\beta_0 + \boldsymbol{\beta}'\mathbf{Z} = 3 + Z_1 - 2Z_2$. The mean square error is

$$\sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = 10 - [1, -1]\begin{bmatrix} .4 & -.6 \\ -.6 & 1.4 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 10 - 3 = 7$$

and the multiple correlation coefficient is

$$\rho_{Y(\mathbf{Z})} = \sqrt{\frac{\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}} = \sqrt{\frac{3}{10}} = .548$$

Note that $\sigma_{YY}(1 - \rho^2_{Y(\mathbf{Z})}) = 10\left(1 - \frac{3}{10}\right) = 7$ is the mean square error. ■

It is possible to show (see Exercise 7.5) that

$$1 - \rho^2_{Y(\mathbf{Z})} = \frac{1}{\rho^{YY}} \tag{7-50}$$

where ρ^{YY} is the upper-left-hand corner of the inverse of the correlation matrix determined from $\boldsymbol{\Sigma}$.

The restriction to linear predictors is closely connected to the assumption of normality. Specifically, if we take

$$\begin{bmatrix} Y \\ Z_1 \\ Z_2 \\ \vdots \\ Z_r \end{bmatrix} \quad \text{to be distributed as} \quad N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

then the conditional distribution of Y with $z_1, z_2, \ldots, z_r$ fixed (see Result 4.6) is

$$N(\mu_Y + \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}), \sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y})$$

The mean of this conditional distribution is the linear predictor in Result 7.12. That is,

$$\begin{aligned} E(Y|z_1, z_2, \ldots, z_r) &= \mu_Y + \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}) \\ &= \beta_0 + \boldsymbol{\beta}'\mathbf{z} \end{aligned} \tag{7-51}$$

and we conclude that $E(Y|z_1, z_2, \ldots, z_r)$ is the best linear predictor of Y when the population is $N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The conditional expectation of Y in (7-51) is called the *regression function*. For normal populations, it is linear.

When the population is *not* normal, the regression function $E(Y \mid z_1, z_2, \ldots, z_r)$ need not be of the form $\beta_0 + \boldsymbol{\beta}'\mathbf{z}$. Nevertheless, it can be shown (see [22]) that $E(Y \mid z_1, z_2, \ldots, z_r)$, whatever its form, predicts Y with the smallest mean square error. Fortunately, this wider optimality among all estimators is possessed by the *linear* predictor when the population is normal.

Result 7.13. Suppose the joint distribution of Y and $\mathbf{Z}$ is $N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \overline{Y} \\ \hline \overline{\mathbf{Z}} \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \left[\begin{array}{c|c} s_{YY} & \mathbf{s}'_{\mathbf{Z}Y} \\ \hline \mathbf{s}_{\mathbf{Z}Y} & \mathbf{S}_{\mathbf{ZZ}} \end{array}\right]$$

be the sample mean vector and sample covariance matrix, respectively, for a random sample of size n from this population. Then the maximum likelihood estimators of the coefficients in the linear predictor are

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y}, \qquad \hat{\beta}_0 = \overline{Y} - \mathbf{s}'_{\mathbf{Z}Y}\mathbf{S}_{\mathbf{ZZ}}^{-1}\overline{\mathbf{Z}} = \overline{Y} - \hat{\boldsymbol{\beta}}'\overline{\mathbf{Z}}$$

Consequently, the maximum likelihood estimator of the linear regression function is

$$\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{z} = \overline{Y} + \mathbf{s}'_{\mathbf{Z}Y}\mathbf{S}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \overline{\mathbf{Z}})$$

and the maximum likelihood estimator of the mean square error $E[Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z}]^2$ is

$$\hat{\sigma}_{YY\cdot\mathbf{Z}} = \frac{n-1}{n}(s_{YY} - \mathbf{s}'_{\mathbf{Z}Y}\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y})$$

Proof. We use Result 4.11 and the invariance property of maximum likelihood estimators. [See (4-20).] Since, from Result 7.12,

$$\beta_0 = \mu_Y - (\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y})'\boldsymbol{\mu}_{\mathbf{Z}},$$
$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}, \qquad \beta_0 + \boldsymbol{\beta}'\mathbf{z} = \mu_Y + \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})$$

and

$$\text{mean square error} = \sigma_{YY\cdot\mathbf{Z}} = \sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}$$

the conclusions follow upon substitution of the maximum likelihood estimators

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \overline{Y} \\ \hline \overline{\mathbf{Z}} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \left[\begin{array}{c|c} \hat{\sigma}_{YY} & \hat{\boldsymbol{\sigma}}'_{\mathbf{Z}Y} \\ \hline \hat{\boldsymbol{\sigma}}_{\mathbf{Z}Y} & \hat{\boldsymbol{\Sigma}}_{\mathbf{ZZ}} \end{array}\right] = \left(\frac{n-1}{n}\right)\mathbf{S}$$

for

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \hline \boldsymbol{\mu}_{\mathbf{Z}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \sigma_{YY} & \boldsymbol{\sigma}'_{\mathbf{Z}Y} \\ \hline \boldsymbol{\sigma}_{\mathbf{Z}Y} & \boldsymbol{\Sigma}_{\mathbf{ZZ}} \end{array}\right] \qquad \blacksquare$$

It is customary to change the divisor from n to $n - (r + 1)$ in the estimator of the mean square error, $\sigma_{YY\cdot\mathbf{Z}} = E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z})^2$, in order to obtain the *unbiased* estimator

$$\left(\frac{n-1}{n-r-1}\right)(s_{YY} - \mathbf{s}'_{\mathbf{Z}Y}\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y}) = \frac{\sum_{j=1}^{n}(Y_j - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'\mathbf{Z}_j)^2}{n-r-1} \tag{7-52}$$

Example 7.12 (Maximum likelihood estimate of the regression function—single response) For the computer data of Example 7.6, the $n = 7$ observations on Y (CPU time), Z_1 (orders), and Z_2 (add–delete items) give the sample mean vector and sample covariance matrix:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \bar{y} \\ \hline \bar{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} 150.44 \\ \hline 130.24 \\ 3.547 \end{bmatrix}$$

$$\mathbf{S} = \left[\begin{array}{c|c} s_{YY} & \mathbf{s}'_{\mathbf{Z}Y} \\ \hline \mathbf{s}_{\mathbf{Z}Y} & \mathbf{S}_{\mathbf{ZZ}} \end{array}\right] = \left[\begin{array}{c|cc} 467.913 & 418.763 & 35.983 \\ \hline 418.763 & 377.200 & 28.034 \\ 35.983 & 28.034 & 13.657 \end{array}\right]$$

Assuming that Y, Z_1, and Z_2 are jointly normal, obtain the estimated regression function and the estimated mean square error.

Result 7.13 gives the maximum likelihood estimates

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y} = \begin{bmatrix} .003128 & -.006422 \\ -.006422 & .086404 \end{bmatrix}\begin{bmatrix} 418.763 \\ 35.983 \end{bmatrix} = \begin{bmatrix} 1.079 \\ .420 \end{bmatrix}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{z}} = 150.44 - [1.079, .420]\begin{bmatrix} 130.24 \\ 3.547 \end{bmatrix} = 150.44 - 142.019$$
$$= 8.421$$

and the estimated regression function

$$\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{z} = 8.42 - 1.08z_1 + .42z_2$$

The maximum likelihood estimate of the mean square error arising from the prediction of Y with this regression function is

$$\left(\frac{n-1}{n}\right)(s_{YY} - \mathbf{s}_{\mathbf{Z}Y}'\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y})$$
$$= \left(\frac{6}{7}\right)\left(467.913 - [418.763, 35.983]\begin{bmatrix} .003128 & -.006422 \\ -.006422 & .086404 \end{bmatrix}\begin{bmatrix} 418.763 \\ 35.983 \end{bmatrix}\right)$$
$$= .894$$

■

Prediction of Several Variables

The extension of the previous results to the prediction of several responses $Y_1, Y_2, \ldots, Y_m$ is almost immediate. We present this extension for normal populations.

Suppose

$$\begin{bmatrix} \underset{(m\times 1)}{\mathbf{Y}} \\ \hdashline \underset{(r\times 1)}{\mathbf{Z}} \end{bmatrix} \quad \text{is distributed as} \quad N_{m+r}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\mu} = \begin{bmatrix} \underset{(m\times 1)}{\boldsymbol{\mu}_{\mathbf{Y}}} \\ \hdashline \underset{(r\times 1)}{\boldsymbol{\mu}_{\mathbf{Z}}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c:c} \underset{(m\times m)}{\boldsymbol{\Sigma}_{\mathbf{YY}}} & \underset{(m\times r)}{\boldsymbol{\Sigma}_{\mathbf{YZ}}} \\ \hdashline \underset{(r\times m)}{\boldsymbol{\Sigma}_{\mathbf{ZY}}} & \underset{(r\times r)}{\boldsymbol{\Sigma}_{\mathbf{ZZ}}} \end{array}\right]$$

By Result 4.6, the conditional expectation of $[Y_1, Y_2, \ldots, Y_m]'$, given the fixed values $z_1, z_2, \ldots, z_r$ of the predictor variables, is

$$E[\mathbf{Y} \mid z_1, z_2, \ldots, z_r] = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}) \tag{7-53}$$

This conditional expected value, considered as a function of $z_1, z_2, \ldots, z_r$, is called the *multivariate regression* of the vector $\mathbf{Y}$ on $\mathbf{Z}$. It is composed of m univariate regressions. For instance, the first component of the conditional mean vector is $\mu_{Y_1} + \boldsymbol{\Sigma}_{Y_1\mathbf{Z}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}) = E(Y_1 \mid z_1, z_2, \ldots, z_r)$, which minimizes the mean square error for the prediction of Y_1. The $m \times r$ matrix $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}$ is called the matrix of *regression coefficients*.

The error of prediction vector

$$\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})$$

has the expected squares and cross-products matrix

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{YY\cdot Z}} &= E[\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})]\,[\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})]' \\ &= \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{YZ}})' - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}} + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{YZ}})' \qquad (7\text{-}54) \\ &= \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}}\end{aligned}$$

Because $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are typically unknown, they must be estimated from a random sample in order to construct the multivariate linear predictor and determine expected prediction errors.

Result 7.14. Suppose $\mathbf{Y}$ and $\mathbf{Z}$ are jointly distributed as $N_{m+r}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the regression of the vector $\mathbf{Y}$ on $\mathbf{Z}$ is

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{z} = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\mu}_{\mathbf{Z}} + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\mathbf{z} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})$$

The expected squares and cross-products matrix for the errors is

$$E(\mathbf{Y} - \boldsymbol{\beta}_0 - \boldsymbol{\beta}\mathbf{Z})(\mathbf{Y} - \boldsymbol{\beta}_0 - \boldsymbol{\beta}\mathbf{Z})' = \boldsymbol{\Sigma}_{\mathbf{YY\cdot Z}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}}$$

Based on a random sample of size n, the maximum likelihood estimator of the regression function is

$$\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}\mathbf{z} = \overline{\mathbf{Y}} + \mathbf{S}_{\mathbf{YZ}}\mathbf{S}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \overline{\mathbf{Z}})$$

and the maximum likelihood estimator of $\boldsymbol{\Sigma}_{\mathbf{YY\cdot Z}}$ is

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{YY\cdot Z}} = \left(\frac{n-1}{n}\right)(\mathbf{S}_{\mathbf{YY}} - \mathbf{S}_{\mathbf{YZ}}\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{S}_{\mathbf{ZY}})$$

Proof. The regression function and the covariance matrix for the prediction errors follow from Result 4.6. Using the relationships

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\mu}_{\mathbf{Z}}, \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}$$

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{z} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})$$

$$\boldsymbol{\Sigma}_{\mathbf{YY\cdot Z}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\beta}'$$

we deduce the maximum likelihood statements from the invariance property [see (4-20)] of maximum likelihood estimators upon substitution of

$$\hat{\boldsymbol{\mu}} = \left[\begin{array}{c}\overline{\mathbf{Y}} \\ \hdashline \overline{\mathbf{Z}}\end{array}\right]; \quad \hat{\boldsymbol{\Sigma}} = \left[\begin{array}{c:c}\hat{\boldsymbol{\Sigma}}_{\mathbf{YY}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{YZ}} \\ \hdashline \hat{\boldsymbol{\Sigma}}_{\mathbf{ZY}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{ZZ}}\end{array}\right] = \left(\frac{n-1}{n}\right)\mathbf{S} = \left(\frac{n-1}{n}\right)\left[\begin{array}{c:c}\mathbf{S}_{\mathbf{YY}} & \mathbf{S}_{\mathbf{YZ}} \\ \hdashline \mathbf{S}_{\mathbf{ZY}} & \mathbf{S}_{\mathbf{ZZ}}\end{array}\right]$$

■

It can be shown that an unbiased estimator of $\boldsymbol{\Sigma}_{\mathbf{YY\cdot Z}}$ is

$$\begin{aligned}&\left(\frac{n-1}{n-r-1}\right)(\mathbf{S}_{\mathbf{YY}} - \mathbf{S}_{\mathbf{YZ}}\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{S}_{\mathbf{ZY}}) \\ &\qquad = \frac{1}{n-r-1}\sum_{j=1}^{n}(\mathbf{Y}_j - \hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}\mathbf{Z}_j)(\mathbf{Y}_j - \hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}\mathbf{Z}_j)' \qquad (7\text{-}55)\end{aligned}$$

Example 7.13 (Maximum likelihood estimates of the regression functions—two responses) We return to the computer data given in Examples 7.6 and 7.10. For Y_1 = CPU time, Y_2 = disk I/O capacity, Z_1 = orders, and Z_2 = add–delete items, we have

$$\hat{\boldsymbol{\mu}} = \left[\begin{array}{c}\bar{\mathbf{y}} \\ \hdashline \bar{\mathbf{z}}\end{array}\right] = \left[\begin{array}{r}150.44 \\ 327.79 \\ \hdashline 130.24 \\ 3.547\end{array}\right]$$

and

$$\mathbf{S} = \left[\begin{array}{c:c}\mathbf{S_{YY}} & \mathbf{S_{YZ}} \\ \hdashline \mathbf{S_{ZY}} & \mathbf{S_{ZZ}}\end{array}\right] = \left[\begin{array}{rr:rr}467.913 & 1148.556 & 418.763 & 35.983 \\ 1148.556 & 3072.491 & 1008.976 & 140.558 \\ \hdashline 418.763 & 1008.976 & 377.200 & 28.034 \\ 35.983 & 140.558 & 28.034 & 13.657\end{array}\right]$$

Assuming normality, we find that the estimated regression function is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}\mathbf{z} &= \bar{\mathbf{y}} + \mathbf{S_{YZ}}\mathbf{S_{ZZ}^{-1}}(\mathbf{z} - \bar{\mathbf{z}}) \\
&= \begin{bmatrix}150.44 \\ 327.79\end{bmatrix} + \begin{bmatrix}418.763 & 35.983 \\ 1008.976 & 140.558\end{bmatrix} \\
&\qquad\qquad \times \begin{bmatrix}.003128 & -.006422 \\ -.006422 & .086404\end{bmatrix}\begin{bmatrix}z_1 - 130.24 \\ z_2 - 3.547\end{bmatrix} \\
&= \begin{bmatrix}150.44 \\ 327.79\end{bmatrix} + \begin{bmatrix}1.079\,(z_1 - 130.24) + .420\,(z_2 - 3.547) \\ 2.254\,(z_1 - 130.24) + 5.665\,(z_2 - 3.547)\end{bmatrix}
\end{aligned}$$

Thus, the minimum mean square error predictor of Y_1 is

$$150.44 + 1.079(z_1 - 130.24) + .420(z_2 - 3.547) = 8.42 + 1.08z_1 + .42z_2$$

Similarly, the best predictor of Y_2 is

$$14.14 + 2.25z_1 + 5.67z_2$$

The maximum likelihood estimate of the expected squared errors and cross-products matrix $\boldsymbol{\Sigma}_{\mathbf{YY}\cdot\mathbf{Z}}$ is given by

$$\begin{aligned}
&\left(\frac{n-1}{n}\right)(\mathbf{S_{YY}} - \mathbf{S_{YZ}}\mathbf{S_{ZZ}^{-1}}\mathbf{S_{ZY}}) \\
&\quad = \left(\frac{6}{7}\right)\left(\begin{bmatrix}467.913 & 1148.536 \\ 1148.536 & 3072.491\end{bmatrix}\right. \\
&\qquad \left. - \begin{bmatrix}418.763 & 35.983 \\ 1008.976 & 140.558\end{bmatrix}\begin{bmatrix}.003128 & -.006422 \\ -.006422 & .086404\end{bmatrix}\begin{bmatrix}418.763 & 1008.976 \\ 35.983 & 140.558\end{bmatrix}\right) \\
&\quad = \left(\frac{6}{7}\right)\begin{bmatrix}1.043 & 1.042 \\ 1.042 & 2.572\end{bmatrix} = \begin{bmatrix}.894 & .893 \\ .893 & 2.205\end{bmatrix}
\end{aligned}$$

The first estimated regression function, $8.42 + 1.08z_1 + .42z_2$, and the associated mean square error, .894, are the same as those in Example 7.12 for the single-response case. Similarly, the second estimated regression function, $14.14 + 2.25z_1 + 5.67z_2$, is the same as that given in Example 7.10.

We see that the data enable us to predict the first response, Y_1, with smaller error than the second response, Y_2. The positive covariance .893 indicates that overprediction (underprediction) of CPU time tends to be accompanied by overprediction (underprediction) of disk capacity. ■

Comment. Result 7.14 states that the assumption of a joint normal distribution for the whole collection $Y_1, Y_2, \ldots, Y_m, Z_1, Z_2, \ldots, Z_r$ leads to the prediction equations

$$\begin{aligned}\hat{y}_1 &= \hat{\beta}_{01} + \hat{\beta}_{11}z_1 + \cdots + \hat{\beta}_{r1}z_r \\ \hat{y}_2 &= \hat{\beta}_{02} + \hat{\beta}_{12}z_1 + \cdots + \hat{\beta}_{r2}z_r \\ &\vdots \qquad\quad \vdots \\ \hat{y}_m &= \hat{\beta}_{0m} + \hat{\beta}_{1m}z_1 + \cdots + \hat{\beta}_{rm}z_r\end{aligned}$$

We note the following:

1. The same values, $z_1, z_2, \ldots, z_r$ are used to predict each Y_i.
2. The $\hat{\beta}_{ik}$ are estimates of the (i,k)th entry of the regression coefficient matrix $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}$ for $i, k \geq 1$.

We conclude this discussion of the regression problem by introducing one further correlation coefficient.

Partial Correlation Coefficient

Consider the pair of errors

$$\begin{aligned}&Y_1 - \mu_{Y_1} - \boldsymbol{\Sigma}_{Y_1\mathbf{Z}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}) \\ &Y_2 - \mu_{Y_2} - \boldsymbol{\Sigma}_{Y_2\mathbf{Z}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})\end{aligned}$$

obtained from using the best linear predictors to predict Y_1 and Y_2. Their correlation, determined from the error covariance matrix $\boldsymbol{\Sigma}_{\mathbf{YY}\cdot\mathbf{Z}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}}$, measures the association between Y_1 and Y_2 after eliminating the effects of $Z_1, Z_2, \ldots, Z_r$.

We define the *partial correlation coefficient* between Y_1 and Y_2, eliminating $Z_1, Z_2, \ldots, Z_r$, by

$$\rho_{Y_1Y_2\cdot\mathbf{Z}} = \frac{\sigma_{Y_1Y_2\cdot\mathbf{Z}}}{\sqrt{\sigma_{Y_1Y_1\cdot\mathbf{Z}}}\sqrt{\sigma_{Y_2Y_2\cdot\mathbf{Z}}}} \tag{7-56}$$

where $\sigma_{Y_iY_k\cdot\mathbf{Z}}$ is the (i,k)th entry in the matrix $\boldsymbol{\Sigma}_{\mathbf{YY}\cdot\mathbf{Z}} = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YZ}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ZY}}$. The corresponding *sample partial correlation coefficient* is

$$r_{Y_1Y_2\cdot\mathbf{Z}} = \frac{s_{Y_1Y_2\cdot\mathbf{Z}}}{\sqrt{s_{Y_1Y_1\cdot\mathbf{Z}}}\sqrt{s_{Y_2Y_2\cdot\mathbf{Z}}}} \tag{7-57}$$