THE EFFECTS AND DETECTION OF COLLINEARITY IN AN ANALYSIS OF COVARIANCE

by

JO JANE GIACOMINI

B.S., Colorado State University, 1977

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

Approved by

James J. Higgins
Major Professor

ACKNOWLEDGMENTS

I wish to thank the members of my committee, Dr. James Higgins, Dr. George Milliken and Dr. Michael Deaton. Special appreciation is extended to Dr. James Schwenke for his valuable suggestions during the course of this research.

I also wish to thank the entire Statistics Department for making my stay at Kansas State a pleasant experience. I am also thankful to Sharon Luce for doing an excellent job of typing this report.

Finally I wish to thank my family and friends for their encouragement and understanding.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Analysis of covariance is used to reduce the variability due to experimental error and thus increase precision in randomized experiments. The procedure is to measure a covariate, a source of variation that cannot or has not been controlled by the experimental design, along with the response variable. This covariate, X, is related to the response variable, Y. The relationship between X and Y may be different for each level of the treatment giving different regression lines. Through the analysis of covariance, a comparison of these regression lines is made.

The purpose of the analysis of covariance is to test whether or not there are significant distances between the regression lines, or in more common terminology, whether or not there are treatment effects. The difference between analysis of covariance and the analysis of variance is that analysis of variance is based on the unconditional distribution of Y, the response variable, whereas analysis of covariance is based on the conditional distribution of Y given X, the covariate. The analysis of covariance model combines into one model, both the regression and analysis of variance models, by having different regression models for each population.

In the analysis of covariance, different testing strategies may be employed to determine an appropriate form of the covariance model. One strategy may assume the covariate is necessary and that the slope parameter is constant across all treatments. The only test of interest is one for equal intercepts of the parallel lines. Another strategy might initially test whether or not the covariate is necessary, then based on the conclusion of that test do another test.

One of the assumptions made in analysis of covariance is independence of covariate and treatment, or noncollinearity between covariate and treatment. When a treatment is adjusted for the covariate, it is as if the treatment mean is changed to the value it would be expected to have if all observations had the same covariate value. However, if the treatment affects the covariate, the adjustment removes part of the treatment effect and may lead to incorrect conclusions.

If the design of the experiment is simple, that is, a one-way analysis of covariance with one covariate, then the problem of a collinearity between treatment and covariate can be detected by looking at a scatterplot. Once a design becomes more complicated then the existence of a collinearity will be more difficult to detect.

The purpose of this study is to investigate the effects of a collinearity in an analysis of covariance model and the importance of a testing strategy. This report is divided into two parts. The first part discusses the effects of a collinearity between covariate and treatment. These effects are examined under different tests of hypothesis, assuming different true models. The second part concerns the diagnostic procedures employed to detect a collinearity in analysis of covariance.

CHAPTER 2

Effects of Collinearity in an Analysis of Covariance Model

2.1 Statement of the Problem

In the standard use of the analysis of covariance model, the assumption is made that the range of values for the covariate is the same for all levels of the treatment. When this assumption is violated, the tests of hypothesis in a covariance analysis are affected and the F-tests may likely lead to the wrong conclusion concerning the presence of treatment effects. This problem could be described as a collinearity between the covariate and treatment, where collinearity is a term normally used in regression analysis. A broad definition of this term is that, if two variables lie "almost" on the same line, then they are collinear (Belsley, Kuh, and Welsch). By this definition, the covariate and treatment are collinear if the values of the X's are similar to each other within treatment levels but are dissimilar between treatment levels.

The extent of this problem is determined by the severity of the collinearity, the underlying true population model, and the testing strategy employed. How these factors influence the problem is the subject of this chapter.

In addition to determining the effects of the collinearity, there is also the problem of detecting the collinearity. Depending on the complexity of the analysis of covariance model, this may either be a simple, or a difficult task. If the model contains multiple covariates and/or multiple factors in the treatment structure it is possible that more than one collinearity may exist. In Chapter 3, methods of detecting and assessing the collinearity will be discussed.

2.2 Notation and Models

Consider a completely randomized design with t treatments and one covariate where the response and covariate are measured on each experimental unit. A linear model denoting a relationship between the response variable and the covariate is

$$y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij} \tag{2.2.1}$$

$$i = 1,\ldots,t \qquad j = 1,\ldots,n_i$$

where $\varepsilon_{ij} \sim N(0,\sigma^2)$ and is the random experimental error

α_i denotes the intercept of the ith regression line

β_i denotes the slope for the ith regression line

X_{ij} denotes the value of the covariate for the jth individual in the ith treatment

y_{ij} denotes the jth observation from the ith treatment.

There are many variations of model (2.2.1). For example assume that only one slope is needed for the t regression lines, that is $\beta_1 = \beta_2 = \ldots = \beta_t = \beta$. This model can be expressed as

$$y_{ij} = \alpha_i + \beta X_{ij} + \varepsilon_{ij} \tag{2.2.2}$$

where β represents the common slope parameters.

In this chapter, five analysis of covariance models are assumed to be the true population models. Using each of the models, different hypotheses of interest are examined. A noncentrality parameter is computed based on the true model and the sums of squares due to deviations from H_o for each test of hypothesis.

In addition to models (2.2.1) and (2.2.2), the other three linear models considered as true models are

$$y_{ij} = \alpha_i + \varepsilon_{ij} \tag{2.2.3}$$

$$y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij} \tag{2.2.4}$$

$$y_{ij} = \alpha + \beta_i X_{ij} + \varepsilon_{ij} \tag{2.2.5}$$

where α denotes a common intercept. The other variables and parameters have been described previously.

In order to express models (2.2.1) through (2.2.5) in matrix notation, the following five matrices and five vectors are defined

$$\underline{X}_1 = \begin{bmatrix} \underline{j}_{n_1} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{j}_{n_2} & & \\ \vdots & & \ddots & \\ \underline{0} & \cdots & & \underline{j}_{n_t} \end{bmatrix} \qquad \underline{b}_1 = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_t \end{bmatrix} \tag{2.2.6}$$

$$\underline{X}_2 = \begin{bmatrix} \underline{j}_{n_1} & \underline{x}_1 \\ \underline{j}_{n_2} & \underline{x}_2 \\ \vdots & \vdots \\ \underline{j}_{n_t} & \underline{x}_t \end{bmatrix} \qquad \underline{b}_2 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{2.2.7}$$

$$\underline{X}_3 = \begin{bmatrix} \underline{j}_{n_1} & \underline{0} & \cdots & \underline{0} & \underline{x}_1 \\ \underline{0} & \underline{j}_{n_2} & & \vdots & \underline{x}_2 \\ \vdots & & \ddots & \vdots & \vdots \\ \underline{0} & \cdots & \cdots & \underline{j}_{n_t} & \underline{x}_t \end{bmatrix} \qquad \underline{b}_3 = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_t \\ \beta \end{bmatrix} \tag{2.2.8}$$

$$\underline{X}_4 = \begin{bmatrix} \underline{j}_{n_1} & \underline{x}_1 & \underline{0} & \cdots & \underline{0} \\ \underline{j}_{n_2} & \underline{0} & \underline{x}_2 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \underline{j}_{n_t} & \underline{0} & \cdots & \cdots & \underline{x}_t \end{bmatrix} \qquad \underline{b}_4 = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_t \end{bmatrix} \tag{2.2.9}$$

$$\underline{X}_5 = \begin{bmatrix} \underline{j}_{n_1} & \underline{0} & \cdots & \underline{0} & \underline{x}_1 & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{j}_{n_2} & & \vdots & \underline{0} & \underline{x}_2 & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ \underline{0} & \cdots & \cdots & \underline{j}_{n_t} & \underline{0} & \cdots & \cdots & \underline{x}_t \end{bmatrix} \quad \underline{b}_5 = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_t \\ \beta_1 \\ \vdots \\ \beta_t \end{bmatrix} \tag{2.2.10}$$

where $\underline{j}_{n_i}$ denotes an $n_i \times 1$ vector of ones and $\underline{x}_i$ is the $n_i \times 1$ vector of covariate values for treatment i.

Using the matrices defined above, models (2.2.1) through (2.2.5) may be expressed, respectively, in matrix notation, as

$$\underline{y} = \underline{X}_5\underline{b}_5 + \underline{\varepsilon} \qquad (2.2.11)$$

$$\underline{y} = \underline{X}_3\underline{b}_3 + \varepsilon \qquad (2.2.12)$$

$$\underline{y} = \underline{X}_1\underline{b}_1 + \varepsilon \qquad (2.2.13)$$

$$\underline{y} = \underline{X}_2\underline{b}_2 + \varepsilon \qquad (2.2.14)$$

$$\underline{y} = \underline{X}_4\underline{b}_4 + \varepsilon \qquad (2.2.15)$$

where $\underline{y}$ is the N x 1 vector of observations ordered by treatment, and $\underline{\varepsilon}$ is the N x 1 vector of random experimental errors and $N = \sum_{i=1}^{t} n_i$.

2.3 Tests of Hypothesis

The main objective of a covariance analysis is to compare the regression lines associated with each population. Different testing strategies may be employed in reaching this objective. If certain assumptions are made, one need only do a single test of hypothesis whereas, under another set of assumptions, a series of tests is necessary. For example, most textbooks will assume a common slope model is adequate and the only test of hypothesis conducted is one of equal intercepts of the parallel lines. In Milliken and Johnson (1983), the first step is to test whether the mean of Y given X depends on the value of X, that is, whether $\beta_1 = \beta_2 = \ldots = \beta_t = 0$. In this case, no assumption is made concerning a single slope. If this hypothesis is rejected then they recommend testing the null hypothesis that the common slope model is adequate. The rejection of this hypothesis implies that each of the regression lines has a different slope and the lines are not parallel. In this case, a comparison of the intercepts is just a comparison of the lines at X = 0 which will only be meaningful when X = 0 is included in or is close to the range of covariate values in the experiment. When the lines are not parallel, a comparison among the lines should be made at several different values of the covariate. If the hypothesis corresponding to a particular covariate value is rejected, then there is a difference between at least two of the regression lines.

Five different tests of hypothesis are considered in this paper. Any combination of the five constitute a testing strategy to be used by the investigator. The five tests are

$$\begin{aligned} H_{01}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha_i \\ H_{a1}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha_i + \beta_i x_{ij} \end{aligned} \tag{2.3.1}$$

$$\begin{aligned} H_{02}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha_i + \beta x_{ij} \\ H_{a2}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha_i + \beta_i x_{ij} \end{aligned} \tag{2.3.2}$$

$$\begin{aligned} H_{03}&: E(y_{ij} \mid \underline{X} = x_{ij}) = \alpha_i \\ H_{a3}&: E(y_{ij} \mid \underline{X} = x_{ij}) = \alpha_i + \beta x_{ij} \end{aligned} \tag{2.3.3}$$

$$\begin{aligned} H_{04}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha + \beta x_{ij} \\ H_{a4}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}) = \alpha_i + \beta x_{ij} \end{aligned} \tag{2.3.4}$$

$$\begin{aligned} H_{05}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}, x_o) = {}^{\alpha}x_o + \beta_i (x_{ij} - x_o) \\ H_{a5}&: E(y_{ij} \mid \underline{X}_{ij} = x_{ij}, x_o) = \alpha_i + \beta_i (x_{ij} - x_o) \end{aligned} \tag{2.3.5}$$

where x_o is a constant.

The first three tests deal with investigating the slope parameter(s) and the last two deal with comparing regression lines. As mentioned previously, the main objective in a covariance analysis is to do tests (2.3.4) and

(2.3.5). Note test (2.3.5) is the comparison of nonparallel lines described previously and this test should be conducted for several values of x_0.

The primary goal of this section is to compute the noncentrality parameter of each test statistic by assuming each model from section 2.2 is the true model. The ratio, U/V, of two independent random variables, where U is a noncentral chi-square random variable divided by its degrees of freedom and V is a central chi-square random variable divided by its degrees of freedom, is a noncentral F random variable (Graybill, 1976). The parameters associated with this noncentral F random variable are its degrees of freedom n_1 and n_2, and a noncentrality parameter, λ. If $\lambda = 0$, then the ratio U/V is a central F random variable.

To calculate λ, the principle of conditional error provides an excellent approach. In order to use this method, the models under the null and alternative hypotheses are expressed in matrix notation and the tests are

$$H_{01}: E(\underline{y}) = \underline{X}_1\underline{b}_1 \qquad (2.3.6)$$
$$H_{a1}: E(\underline{y}) = \underline{X}_5\underline{b}_5$$

$$H_{02}: E(\underline{y}) = \underline{X}_3\underline{b}_3 \qquad (2.3.7)$$
$$H_{a2}: E(\underline{y}) = \underline{X}_5\underline{b}_5$$

$$H_{03}: E(\underline{y}) = \underline{X}_1\underline{b}_1 \qquad (2.3.8)$$
$$H_{a3}: E(\underline{y}) = \underline{X}_3\underline{b}_3$$

$$H_{04}: E(\underline{y}) = \underline{X}_2\underline{b}_2 \qquad (2.3.9)$$
$$H_{a4}: E(\underline{y}) = \underline{X}_3\underline{b}_3$$

$$H_{05}\text{: } E(\underline{y}) = \underline{X}_{4_0}\underline{b}_{4_0} \qquad (2.3.10)$$
$$H_{a5}\text{: } E(\underline{y}) = \underline{X}_5\underline{b}_5$$

where the matrices and vectors are defined in (2.2.6) through (2.2.8) and (2.2.10). The matrix $\underline{X}_{4_0}$ and vector $\underline{b}_{4_0}$ are defined to be the conditional matrix given X_0 and corresponding vector. The matrix is similar to (2.2.9) except the vectors $\underline{x}_i$ are equal to $\underline{x}_i - x_0\underline{j}_{n_i}$ for all i. The change in $\underline{b}_{4_0}$ from $\underline{b}_4$ is that α is equal to α_{x_0}, the intersection point of the regression lines for the covariate value, x_0. The matrix $\underline{X}_{4_0}$ and vector $\underline{b}_4$ is different for every different value of x_0 that the test is done for.

To calculate λ by the principle of conditional error method, first compute the sums of squares residual for the full model, or the model under H_1 and denote it by SSRes(H_1) . The second step is to compute the sums of squares residual for the reduced model, or the model under H_0 and denote it as SSRes(H_0) . Then, the sums of squares due to deviations from H_0, SSH_0, is equal to SSRes(H_0) - SSRes(H_1). The degrees of freedom for SSH_0 is equal to the degrees of freedom for SSRes(H_0) minus the degrees of freedom for SSRes(H_1). The ratio

$$F_c = \frac{SSH_0/n_1}{SSRes(H_1)/n_2} \quad \sim \quad F'_{(n_1,n_2;\lambda)} \qquad (2.3.11)$$

where n_1 is the degrees of freedom for SSH_0,

n_2 is the degrees of freedom for SSRes(H_1),

F' denotes the noncentral F distribution and

λ is the noncentrality parameter.

The noncentrality parameter, λ, is calculated as

$$\lambda = \frac{1}{2\sigma^2} \ \underline{\mu}' \ A \ \underline{\mu} \qquad (2.3.12)$$

where σ^2 is the variance of each experimental unit

$\underline{\mu}$ is the vector of expected values of $\underline{y}$ under the true model

$\underline{A}$ is the matrix of the quadratic form, SSH_0.

Using the approach discussed above, assume model (2.2.13) is the true model and use test (2.3.6). The model under the alternative hypothesis is (2.2.11), thus the residual sums of squares for the model under H_{a1} is

$$\begin{aligned} SSRes(H_{a1}) &= (\underline{y} - \underline{X_5 X_5^- y})'(\underline{y} - \underline{X_5 X_5^- y}) \\ &= \underline{y}'(\underline{I} - \underline{X_5 X_5^-})\underline{y} \qquad (2.3.13) \end{aligned}$$

where $\underline{X_5^-}$ is the generalized inverse of $\underline{X_5}$ and $\underline{X_5^- y}$ is the least squares estimator of $\underline{b_5}$. The matrix $\underline{I}$ denotes the identity matrix.

The model under the null hypothesis is (2.2.13) and the residual sums of squares for the model under H_{01} is

$$\begin{aligned} SSRes(H_{01}) &= (\underline{y} - \underline{X_1 X_1^- y})' \ (\underline{y} - \underline{X_1 X_1^- y}) \\ &= \underline{y}' \ (\underline{I} - \underline{X_1 X_1^-})\underline{y} \qquad (2.3.14) \end{aligned}$$

where $\underline{X_1^-}$ is the generalized inverse of $\underline{X_1}$ and $\underline{X_1^- y}$ is the least squares estimator of $\underline{b_1}$.

Using equations (2.3.13) and (2.3.14), the sums of squares due to deviations from H_0 is

$$\begin{aligned} SSH_0 &= SSRes(H_{01}) - SSRes(H_{a1}) \\ &= \underline{y}' \ (\underline{I} - \underline{X_1 X_1^-})\underline{y} - \underline{y}' \ (\underline{I} - \underline{X_5 X_5^-})\underline{y} \\ &= \underline{y}' \ (\underline{X_5 X_5^-} - \underline{X_1 X_1^-})\underline{y} \qquad (2.3.15) \end{aligned}$$

The degrees of freedom for SSRes(H_{01}) is N - t, and for SSRes(H_{a1}) is N - 2t. Thus, the degrees of freedom for SSH_0 is (N-t)-(N-2t) = t. The test statistic for (2.3.6) is

$$F_c = \frac{SSH_{01}/t}{SSRes(H_{a1})/N-2t} \sim F'_{(t,N-2t;\lambda)} \qquad (2.3.16)$$

The noncentrality parameter, λ, will equal 0 if the model under the null hypothesis is the true model or if a perfect collinearity is present. To compute λ, the expected value of $\underline{y}$ under the true model and the matrix of the quadratic form SSH_0, is needed. Assume model (2.2.13) is the true model which implies the expected value of $\underline{y}$ is

$$E(\underline{y}) = \underline{\mu} = \underline{X}_1\underline{b}_1 \qquad (2.3.17)$$

The matrix of the quadratic form in 2.3.15 is

$$\underline{A} = (\underline{X}_5\underline{X}_5^- - \underline{X}_1\underline{X}_1^-) \qquad (2.3.18)$$

Therefore, based on the equation in (2.3.12), λ is

$$\begin{aligned}\lambda &= \frac{1}{2\sigma^2}\ \underline{\mu}'\ A\ \underline{\mu} \\ &= \frac{1}{2\sigma^2}\ (\underline{X}_1\underline{b}_1)'(\underline{X}_5\underline{X}_5^- - \underline{X}_1\underline{X}_1^-)(\underline{X}_1\underline{b}_1) \\ &= \frac{1}{2\sigma^2}\ \underline{b}_1'(\underline{X}_1\underline{X}_5\underline{X}_5^-\underline{X}_1 - \underline{X}_1'\underline{X}_1)\underline{b}_1 \qquad (2.3.19)\end{aligned}$$

By definition of a generalized inverse, it is known that $\underline{X}_1\underline{X}_1^-\underline{X}_1 = \underline{X}_1$, so $\underline{X}_1'\underline{X}_1\underline{X}_1^-\underline{X}_1 = \underline{X}_1'\underline{X}_1$. Also note that the column space of $\underline{X}_1$ is contained in the column space of $\underline{X}_5$, which implies $\underline{X}_5\underline{X}_5^-$ is a projection operator of $\underline{X}_1$ (Graybill, 1969) and $\underline{X}_1'\underline{X}_5\underline{X}_5^-\underline{X}_1 = \underline{X}_1'\underline{X}_1$. Therefore, equation (2.3.19) becomes

$$\lambda = \frac{1}{2\sigma^2} \; \underline{b}_1'(\underline{X}_1'\underline{X}_1 - \underline{X}_1'\underline{X}_1)\underline{b}_1$$
$$= 0$$

Since $\lambda = 0$, the model under the null hypothesis is the true model if there is not a perfect collinearity. Recall the model under H_{01} is $\underline{y} = \underline{X}_1\underline{b}_1 + \underline{\varepsilon}$ and model (2.2.13) is $\underline{y} = \underline{X}_1\underline{b}_1 + \underline{\varepsilon}$. Therefore the model under H_{01} is the true model.

For test (2.3.7), the sums of squares due to deviations from H_{02} is

$$SSH_{02} = \underline{y}'(\underline{X}_5\underline{X}_5^- - \underline{X}_3\underline{X}_3^-)\underline{y}$$

Again, assume model (2.2.13) is the true model so the expected value of $\underline{y}$ is shown in (2.3.17). Therefore the noncentrality parameter for test (2.3.7) for this true model is

$$\lambda = \frac{1}{2\sigma^2} \; \underline{b}_1'(\underline{X}_1'\underline{X}_5\underline{X}_5^-\underline{X}_1 - \underline{X}_1'\underline{X}_3\underline{X}_3^-\underline{X}_1)\underline{b}_1$$
$$= \frac{1}{2\sigma^2} \; \underline{b}_1'(\underline{X}_1'\underline{X}_1 - \underline{X}_1'\underline{X}_1)\underline{b}_1$$
$$= 0$$

Note, the column space of $\underline{X}_1$ is contained in the column space of $\underline{X}_3$. For this test of hypothesis, the model under H_{02}, $\underline{y} = \underline{X}_3\underline{b}_3 + \underline{\varepsilon}$, is not the true model, $\underline{y} = \underline{X}_1\underline{b}_1 + \underline{\varepsilon}$, but is a special case of the true model. The model under H_{02} is the same as the true model only when β is equal to zero.

Tables 2.3.1 and 2.3.2 show the final value for λ computed for each of the five tests (2.3.6) through (2.3.10), assuming each of the models (2.2.11) though (2.2.15), is the true model.

2.4 The Effect of a Collinearity on λ

One of the objectives of hypothesis testing in the analysis of covariance is to determine the best model that fits the data. If a collinearity is present, the testing strategy is even more important because an investigator may make the wrong conclusion at step one and never get close to the true model. In the case of a collinearity, using different strategies may yield different conclusions which is an indication of the collinearity. This section explains how a collinearity affects λ and in turn influences the testing strategy.

If no collinearity exists between the treatment and the covariate, then the calculation of λ in tables 2.3.1 and 2.3.2 indicate which hypothesis or model will be accepted the majority of the time. As explained in the previous section, the model under the null hypothesis may not be the true model even when $\lambda = 0$. This is actually a problem of misspecification of the assumed model and would be avoided if an appropriate testing strategy is used. A good testing strategy would insure that the true model would eventually be the model under the null hypothesis. For example, assume $\underline{y} = \underline{X}_2\underline{b}_2 + \underline{\varepsilon}$ (2.2.14) is the true model and test (2.3.7) is conducted. From table (2.3.1), it is seen that $\lambda = 0$ so the null hypothesis is not rejected yet the model under H_{02} is not the true model. The next step might be to do test (2.3.8). For this test, $\lambda \neq 0$ so the null hypothesis would be rejected and test (2.3.9) conducted. At this step, $\lambda = 0$ and the model under H_{04} is the true model. Using an appropriate testing strategy eventually led to the true model.

TABLE 2.3.1

Calculation of λ

				TRUE MODELS	
	TEST	SSH_0	(2.2.13) $\underline{y}=\underline{X}_1\underline{b}_1+\underline{\varepsilon}$	(2.2.14) $\underline{y}=\underline{X}_2\underline{b}_2+\underline{\varepsilon}$	(2.2.12) $\underline{y}=\underline{X}_3\underline{b}_3+\underline{\varepsilon}$
(2.3.6)	$H_{01}:E(\underline{y})=\underline{X}_1\underline{b}_1$ $H_{a1}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$\underline{y}'(\underline{X}_5\underline{X}_5^- - \underline{X}_1\underline{X}_1^-)\underline{y}$	0	$k\underline{b}_2'(\underline{X}_2'\underline{X}_2-$ $\underline{X}_2\underline{X}_1\underline{X}_1^-\underline{X}_2)\underline{b}_2$	$k\underline{b}_3'(\underline{X}_3'\underline{X}_3-$ $\underline{X}_3'\underline{X}_1\underline{X}_1^-\underline{X}_3)\underline{b}_3$
(2.3.7)	$H_{02}:E(\underline{y})=\underline{X}_3\underline{b}_3$ $H_{a2}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$\underline{y}'(\underline{X}_5\underline{X}_5^- - \underline{X}_3\underline{X}_3^-)\underline{y}$	0	0	0
(2.3.8)	$H_{03}:E(\underline{y})=\underline{X}_1\underline{b}_1$ $H_{a3}:E(\underline{y})=\underline{X}_3\underline{b}_3$	$\underline{y}'(\underline{X}_3\underline{X}_3^- - \underline{X}_1\underline{X}_1^-)\underline{y}$	0	$k\underline{b}_2'(\underline{X}_2'\underline{X}_2-$ $\underline{X}_2'\underline{X}_1\underline{X}_1^-\underline{X}_2)\underline{b}_2$	$k\underline{b}_3'(\underline{X}_3'\underline{X}_3-$ $\underline{X}_3'\underline{X}_1\underline{X}_1^-\underline{X}_3)\underline{b}_3$
(2.3.9)	$H_{04}:E(\underline{y})=\underline{X}_2\underline{b}_2$ $H_{a4}:E(\underline{y})=\underline{X}_3\underline{b}_3$	$\underline{y}'(\underline{X}_3\underline{X}_3^- - \underline{X}_2\underline{X}_2^-)\underline{y}$	$k\underline{b}_1'(\underline{X}_1'\underline{X}_1-$ $\underline{X}_1'\underline{X}_2\underline{X}_2^-\underline{X}_1)\underline{b}_1$	0	$k\underline{b}_3'(\underline{X}_3'\underline{X}_3-$ $\underline{X}_3'\underline{X}_2\underline{X}_2^-\underline{X}_3)\underline{b}_3$
(2.3.10)	$H_{05}:E(\underline{y})=\underline{X}_{4_0}\underline{b}_{4_0}$ $H_{a5}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$y'(X_5X_5^- - X_{4_0}X_{4_0}^-)\underline{y}$	$k\underline{b}_1'(\underline{X}_1'\underline{X}_1-$ $\underline{X}_1\underline{X}_{4_0}\underline{X}_{4_0}^-\underline{X}_1)\underline{b}_1$	0	$k\underline{b}_3'(\underline{X}_3'\underline{X}_3-$ $\underline{X}_3'\underline{X}_{4_0}\underline{X}_{4_0}^-\underline{X}_3)\underline{b}_3$

NOTE: $k = \frac{1}{2\sigma^2}$

TABLE 2.3.2

Calculation of λ

			TRUE MODELS	
	TEST	SSH_0	2.2.15 $\underline{y}=\underline{X}_4\underline{b}_4+\underline{\varepsilon}$	2.2.11 $\underline{y}=\underline{X}_5\underline{b}_5+\underline{\varepsilon}$
(2.3.6)	$H_{01}:E(\underline{y})=\underline{X}_1\underline{b}_1$ $H_{a1}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$\underline{y}'(\underline{X}_5\underline{X}_5^- - \underline{X}_1\underline{X}_1^-)\underline{y}$	$k\underline{b}_4(\underline{X}_4'\underline{X}_4 - \underline{X}_4'\underline{X}_1\underline{X}_1^-\underline{X}_4)\underline{b}_4$	$k\underline{b}_5'(\underline{X}_5'\underline{X}_5 - \underline{X}_5'\underline{X}_1\underline{X}_1^-\underline{X}_5)\underline{b}_5$
(2.3.7)	$H_{02}:E(\underline{y})=\underline{X}_3\underline{b}_3$ $H_{a2}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$\underline{y}'(\underline{X}_5\underline{X}_5^- - \underline{X}_3\underline{X}_3^-)\underline{y}$	$k\underline{b}_4(\underline{X}_4'\underline{X}_4 - \underline{X}_4\underline{X}_3\underline{X}_3^-\underline{X}_4)\underline{b}_4$	$k\underline{b}_5'(\underline{X}_5'\underline{X}_5 - \underline{X}_5\underline{X}_3\underline{X}_3^-\underline{X}_5)\underline{b}_5$
(2.3.8)	$H_{03}:E(\underline{y})=\underline{X}_1\underline{b}_1$ $H_{a3}:E(\underline{y})=\underline{X}_3\underline{b}_3$	$\underline{y}'(\underline{X}_3\underline{X}_3^- - \underline{X}_1\underline{X}_1^-)\underline{y}$	$kb_4(X_4'X_3X_3^-X_4 - X_4'X_1X_1^-X_4)b_4$	$kb_5'(X_5'X_3X_3^-X_5 - X_5'X_1X_1^-X_5)b_5$
(2.3.9)	$H_{04}:E(\underline{y})=\underline{X}_2\underline{b}_2$ $H_{a4}:E(\underline{y})=\underline{X}_3\underline{b}_3$	$\underline{y}(\underline{X}_3\underline{X}_3^- - \underline{X}_2\underline{X}_2^-)\underline{y}$	$k\underline{b}_4(\underline{X}_4'\underline{X}_3\underline{X}_3^-\underline{X}_4 - \underline{X}_4\underline{X}_2\underline{X}_2^-\underline{X}_4)\underline{b}_4$	$k\underline{b}_5'(\underline{X}_5'\underline{X}_3\underline{X}_3^-\underline{X}_5 - \underline{X}_5'\underline{X}_2\underline{X}_2^-\underline{X}_5)\underline{b}_5$
(2.3.10)	$H_{05}:E(\underline{y})=\underline{X}_{4_0}\underline{b}_{4_0}$ $H_{a5}:E(\underline{y})=\underline{X}_5\underline{b}_5$	$\underline{y}(\underline{X}_5\underline{X}_5^- - \underline{X}_{4_0}\underline{X}_{4_0}^-)\underline{y}$	0	$k\underline{b}_5'(\underline{X}_5'\underline{X}_5 - \underline{X}_5'\underline{X}_{4_0}\underline{X}_{4_0}^-\underline{X}_5)\underline{b}_5$

NOTE: $k = \frac{1}{2\sigma^2}$

When a collinearity does exist between the treatment and the covariate, problems will arise in the tests of hypothesis where λ does not equal 0. This does not mean the collinearity does not cause any problems for these cases where λ equals 0. In Chapter 3, it will be shown that a collinearity results in an ill-conditioned design matrix which in turn affects the parameter estimates. Thus, even though the correct decision will probably be made, one would have to be careful in the interpretation of the results.

When the model under the null hypothesis is not the true model or a special case of the true model, then one would want λ to be large. For λ sufficiently large, the null hypothesis will be rejected. It is shown that when a collinearity does exist, λ will get closer to 0, depending on the severity of the collinearity and this will tend to lead to the wrong conclusion about the nature of the true model.

From table 2.3.1, it can be seen for model (2.2.12) that λ is not equal to 0 for four of the tests. Two of the tests, (2.3.8) and (2.3.9), are examined to show how λ is affected by a collinearity. The explanations involved with these two tests are applicable to the other models and tests as well.

Consider first test (2.3.8) with true model (2.2.12). The noncentrality parameter in this case is

$$\lambda = \frac{1}{2\sigma^2} \underline{b}_3'(\underline{X}_3'\underline{X}_3-\underline{X}_3\underline{X}_1\underline{X}_1^-\underline{X}_3)\underline{b}_3 \qquad (2.4.1)$$

Obviously, it is desired to have λ nonzero because the model under the alternative hypothesis is the true model. Therefore, if λ is close to 0, the test would have low power which would likely lead to an incorrect conclusion.

Note that matrix $\underline{X}_1$ (2.2.6) is just the design matrix for the one-way analysis of variance and matrix $\underline{X}_3$ (2.2.8) is $\underline{X}_1$ augmented with the vector of covariate values. The rank of $\underline{X}_1$ is t and the rank of $\underline{X}_3$ is t+1, if no collinearity exists. Now assume a perfect collinearity between treatment and covariate is present. Perfect collinearity implies that two variables lie on the same line. Therefore, $\underline{x}_i = \alpha_i \underline{j}_{n_i}$ and $\underline{X}_3$ is now defined as

$$\underline{X}_3 = \begin{bmatrix} \underline{j}_{n_1} & \underline{0} & \cdots & \underline{0} & \alpha_1 \underline{j}_{n_1} \\ \underline{0} & \underline{j}_{n_2} & & & \alpha_2 \underline{j}_{n_2} \\ \vdots & & \ddots & & \vdots \\ \underline{0} & \cdots & \cdots & \underline{j}_{n_t} & \alpha_t \underline{j}_{n_t} \end{bmatrix} . \qquad (2.4.2)$$

The last column of $\underline{X}_3$ is a linear combination of the other t columns and the rank of $\underline{X}_3$ is equal to t rather than t+1. Furthermore, the column space of $\underline{X}_3$ is the same as the column space of $\underline{X}_1$ so that $\underline{X}_3'\underline{X}_1\underline{X}_1'\underline{X}_3 = \underline{X}_3'\underline{X}_3$. Therefore (2.4.1) is now

$$\lambda = \frac{1}{2\sigma^2} \; \underline{b}_3'(\underline{X}_3'\underline{X}_3 - \underline{X}_3'\underline{X}_3)\underline{b}_3 = 0$$

Matrix (2.4.2) results when a perfect collinearity occurs, which is very rare, but perfect collinearity is not necessary for a problem to exist. Suppose the values in $\underline{x}_i$ are not $\alpha_i \underline{j}_{n_i}$ but rather $(\alpha_i + \delta_{ij})\underline{j}_{n_i}$ where δ_{ij} is a random component and is small compared to α_i, then there is a "near" dependency among the columns of (2.2.8). The stronger the collinearity is between treatment and covariate, that is, the closer it is to a perfect collinearity, the closer λ gets to zero. Thus, the presence of a collinearity may cause the null hypothesis not to be rejected which is an incorrect conclusion in this case.

In the other test, (2.3.9) and assuming (2.2.12) to be the true model, then,

$$\lambda = \frac{1}{2\sigma^2} \underline{b}_3'(\underline{X}_3'\underline{X}_3 - \underline{X}_3'\underline{X}_2\underline{X}_2^-\underline{X}_3)\underline{b}_3 \qquad (2.4.3)$$

Recall the matrix $\underline{X}_2$ (2.2.7) is a column of ones and a column of the covariate values. The discussion given above explaining why λ approaches 0 in the presence of a collinearity, would not apply in this situation. The column space of $\underline{X}_3$ (2.4.2) is not the same as the column space of $\underline{X}_2$ even when a perfect collinearity exists.

Let $\underline{V} = \underline{X}_3'\underline{X}_3 - \underline{X}_3'\underline{X}_2\underline{X}_2^-\underline{X}_3$. The number of linearly independent solutions to $\underline{V}\underline{b}_3 = 0$ is n-r where n is the number of columns in $\underline{V}$ and r is the rank of $\underline{V}$. The rank of $\underline{V}$ is given in the following proof.

Theorem 2.4.1

The rank of $\underline{V}$ is t-1 when no collinearity exists.

Proof:

Let $\underline{X} = [\underline{X}_3,\underline{X}_2]$ then rank $(\underline{X})$ = rank $(\underline{XX}^-)$ = t+1 (2.4.4)

$\underline{XX}^-$ can be expressed as (Milliken, 1971)

$$\underline{XX}^- = \underline{X}_2\underline{X}_2^- + (\underline{I} - \underline{X}_2\underline{X}_2^-)\underline{X}_3[(\underline{I} - \underline{X}_2\underline{X}_2^-)\underline{X}_3]^- .$$

Rank $(\underline{XX}^-)$ = rank $(\underline{X}_2\underline{X}_2^-)$ + rank $[(\underline{I} - \underline{X}_2\underline{X}_2^-)\underline{X}_3]$ because the column space of $\underline{X}_2\underline{X}_2^-$ is orthogonal to the column space of $(I - X_2X_2^-)$.

$$\text{Rank } [(\underline{I} - \underline{X}_2\underline{X}_2^-)\underline{X}_3] = \text{rank } (\underline{XX}^-) - \text{rank}(\underline{X}_2\underline{X}_2^-)$$

$$= (t+1) - 2$$

$$= t - 1$$

$$\underline{V} = \underline{X}_3'\underline{X}_3 - \underline{X}_3'\underline{X}_2\underline{X}_2^-\underline{X}_3 = \underline{X}_3'(\underline{I}-\underline{X}_2\underline{X}_2^-)\underline{X}_3.$$

Thus rank $(\underline{V})$ = rank $[(\underline{I} - \underline{X}_2\underline{X}_2^-)\underline{X}_3]$ = t - 1. (2.4.5)

The number of linearly independent solutions to $\underline{Vb}_1 = 0$ is n-r = (t+1) - (t-1) = 2. Therefore, there are two linearly independent solutions of $\underline{b}_3$ that satisfy the equation (2.4.3) for λ=0, even when there is no collinearity. One class of solutions is when all the α_i's are nonzero and are equal to each other for any value of β. The other class of solutions is where the α_i's are zero and β is any constant. Recall test (2.3.9) is testing for equal treatment means given x. Therefore, the value of β does not influence the computation of λ. In the case where the α_i's are equal to each other, the model under H_0 is the true model and one would want λ to be zero.

Suppose a perfect collinearity between covariate and treatment is present. Then in (2.4.4), the rank of $\underline{X} = [X_3,X_2]$ is t rather than t+1. This implies that (2.4.5) is equal to t - 2, the rank of $\underline{V}$. Thus, the number of linearly independent solutions to $\underline{Vb}_3 = 0$ is three. As stated previously, a perfect collinearity is a rare event but if a collinearity does exist, there is a third solution which will cause λ to be small. The third solution is the case in which the α_i's in $\underline{b}_3$ are a linear function of the means of the covariate values within each treatment level. Therefore, the null hypothesis may not be rejected even though the true model is specified under the alternative.

The other nonzero λ's in tables 2.3.1 and 2.3.2 may approach zero when there is a collinearity by the reasons given above.

2.5 Testing Strategy and Collinearity

When there is a collinearity between the treatment and the covariate, the final covariance model selected is influenced by the testing strategy. Two different strategies used on the same data set may yield different results when a collinearity is present. This section considers three different testing strategies and what happens under each true model. Refer to tables 2.3.1 and 2.3.2 for the calculations of λ when no collinearity exists.

The first strategy consists of first testing whether β is equal to zero (test 2.3.8). If the null hypothesis is not rejected then one would treat the problem as a one-way analysis of variance. Otherwise, the next test would be 2.3.9 which is testing whether there are significant differences between the parallel lines. Note in this test strategy, the assumption is made that if a covariate effect is present, a common slope is adequate for all regression lines.

For model 2.2.13, which is a one-way analysis of variance model, the null hypothesis would not be rejected for test 2.3.8. Thus, the next step would be to treat the model as a one-way analysis of variance model.

The second true model considered (2.2.14) is a simple linear regression model. For this case, the null hypothesis should be rejected in the first step and then test 2.3.9 done. Table 2.3.1 shows that λ is not equal to zero for test 2.3.8 under the true model 2.2.14 but as explained in section 2.4, λ gets close to zero when a collinearity is present. Therefore, the null hypothesis may not be rejected and the one-way analysis of variance model may be used for this true model.

Model 2.2.12 is also affected at test 2.3.8 when a collinearity is present because λ may get close to zero. In this case, the null hypothesis may not be rejected and then the model would be treated as a one-way analysis of variance model. Even if the null hypothesis is rejected in the first step, the collinearity may still cause λ to approach zero for test 2.3.9. For this test and true model , the null hypothesis should be rejected but if λ is small enough, the wrong decision is made.

With models 2.2.15 and 2.2.11, the same problems occur at tests 2.3.8 and 2.3.9, that is the null hypothesis may not be rejected when it should be. This testing strategy is a poor strategy for these true models even if no collinearity existed since neither of these models has a common slope parameter.

The next testing strategy first tests whether the regression lines are parallel (test 2.3.7). Depending on the decision at this step, the next test is either test 2.3.9 or test 2.3.10. If the null hypothesis is rejected at step one, then test 2.3.10 is done. Otherwise, test 2.3.9 is done. In this testing strategy, a covariance effect is assumed.

For model 2.2.13, λ is equal to zero for test 2.3.7 so the collinearity will not affect λ. The null hypothesis is not rejected so the next step is test 2.3.9. At this step, the null hypothesis should be rejected, because even though the model under the alternative hypothesis is not the true model, it is a better choice than the model under the null hypothesis. If a collinearity is present, though, λ may get close to zero so the null hypothesis is not rejected and the wrong decision is made for test 2.3.9.

If the true model is 2.2.14, λ is equal to zero for both test 2.3.7 and 2.3.9. Thus the collinearity does not affect λ for this model. Model 2.2.12 is also not affected by the collinearity at the first step of this testing strategy but at the second step, λ may approach zero and the wrong decision made.

For models 2.2.15 and 2.2.11, the collinearity may cause problems at step one of this testing strategy. If the null hypothesis is not rejected at this step, then the next test is 2.3.9. This test is not appropriate for either of these true models. If the null hypothesis is rejected at the first step, then test 2.3.10 is the next step. For this test a collinearity may lead to a wrong decision for model 2.2.11 but not for model 2.2.15.

The last strategy starts with test 2.3.6, which tests whether all the slopes are equal to zero. The next steps in this strategy are dependent on the outcome of this test. For example, if the null hypothesis is not rejected then the model is treated as a one-way analysis of variance model. Otherwise the next step is test 2.3.7. Based on the decision of test 2.3.7, the next step is either test 2.3.9 or 2.3.10. Thus, a wrong decision at one of these steps may lead a researcher to choose an inadequate model.

If the true model is 2.2.13, λ is zero for test 2.3.6 so the model would be treated as a one-way analysis of variance model, which is the true model. For model 2.2.14, the null hypothesis should be rejected but may not be if a collinearity is present. If the null hypothesis is rejected, then the next step is test 2.3.7 where λ is zero so the collinearity does not affect λ. This is also true for test 2.3.9 which is the next step when the null hypothesis is not rejected for test 2.3.7. Thus, the collinearity only affects model 2.2.14 at the first step.

Model 2.2.12 is affected by a collinearity similarly to model 2.2.14. The difference is at the last step, test 2.3.9, where the null hypothesis should be rejected for model 2.2.12. This may not happen if a collinearity is present.

For model 2.2.11, a collinearity may cause problems at every step of this testing strategy. In this case, the model decided on may not even be close to the true model. This also applies to model 2.2.15 since the collinearity may affect λ at all but the last step of this testing strategy.

These three testing strategies and how each true model is affected under each strategy are examples of how a collinearity influences the analysis of covariance results. This also indicates that using two different strategies on the same set of data may yield different results when a collinearity is present. If this should occur in an analysis of covariance, then a collinearity may be present.

2.6 Example of a Collinearity

A data set was generated in which a collinearity between treatment and covariate was imposed. The purpose of this example is to investigate what happens to the p-value for the test, the noncentrality parameter and the power of the test. The true model selected for this example was (2.2.12), the same model that was discussed in section 2.4. Using data that were generated under this model, tests (2.3.8) and (2.3.9) were examined and the values of the noncentrality parameter, λ, were calculated for each test.

The data were generated as

$$\begin{aligned} x_{ij} &= U_{ij}{*}W + \alpha_i \\ y_{ij} &= \alpha_i + \beta{*}x_{ij} + \varepsilon_{ij}{*}\sigma \qquad (2.6.1) \\ i = 1,2,3 & \qquad\qquad j = 1,2,\ldots,10 \end{aligned}$$

where U_{ij} is a value randomly selected from a Uniform (0,1)

W denotes the width, took on values of 1,2, or 3

ε_{ij} is a value randomly selected from a N(0,1)

σ is the standard deviation, took on value of 1, 1.5,2,3,4, or 6

α_i denotes the ith treatment effect

β denotes the slope

x_{ij} denotes the generated covariate values

y_{ij} denotes the "observed" values

The parameter values selected for the model were

$$\begin{aligned} \alpha_1 &= 1 \\ \alpha_2 &= 5 \\ \alpha_3 &= 9 \\ \beta &= 3 \end{aligned} \quad \text{or } \underline{b}_3 = \begin{bmatrix} 1 \\ 5 \\ 9 \\ 3 \end{bmatrix} \tag{2.6.2}$$

Eighteen different data sets were generated using all combinations of W and σ. The seed number for each of these different sets remained the same. A stronger collinearity would result from a small W value. Equation (2.6.1) shows the obvious collinearity since x_{ij} is a function of α_i.

Recall for model (2.2.2) that tests (2.3.8) and (2.3.9) resulted in λ being nonzero. This implies the null hypothesis should be rejected. The value of λ for each data set and for each test was computed based on the values in (2.6.2) and the matrix equations in table (2.3.1). By using this λ value in the noncentral F distribution, the power of the test of hypothesis was evaluated. The power, $\Pi(\lambda)$, is evaluated as

$$\Pi(\lambda) = \int_{F_{.95;n_1n_2}}^{\infty} F'(w : n_1,n_2;\lambda)\, dw \tag{2.6.3}$$

where $F'(w : n_1, n_2; \lambda)$ denotes the noncentral F distribution,

λ is the calculated noncentrality parameter,

n_1 and n_2 are the degrees of freedom associated with the test, and

$F_{.95;n_1n_2}$ is the upper 95th percentile of a central F distribution.

Table 2.6.1 gives the results for each of the eighteen data sets and each of the two tests. The estimated standard deviation, $\hat{\sigma}$, the calculated F value, F_c, and the probability value, $Pr > F_c$ were taken off the analysis of covariance table. The computation of the power, $\Pi(\lambda)$, was done using an approximation algorithm (Handbook of Mathematical Functions). The degrees of freedom associated with tests (2.3.8) and (2.3.9) are 1 and 26, and 2 and 26, respectively.

The results in Table (2.6.1) give an indication of how the p-value of the test, the noncentrality parameter and the power of the test are affected in an analysis of covariance when a collinearity exists between the treatment and the covariate. The following conclusions are based on this particular example and should not be assumed to be true in general.

This example indicates that the value of λ gets smaller as the width from which the covariate values are selected gets smaller and the standard deviation of the ε_{ij}'s gets larger. Recall a smaller width indicates a stronger collinearity. This smaller λ is expected since a strong collinearity causes λ to approach zero as explained in section 2.4. The λ value in turn affects the power of the test since $\Pi(\lambda)$ is a direct function of λ.

When λ is small, this implies the null hypothesis will not be rejected.

Table 2.6.1

Collinearity Example

			Test (2.3.8)				Test (2.3.9)			
σ	w	$\hat{\sigma}$	F_c	$Pr>F_c$	λ	$\Pi(\lambda)$	F_c	$Pr>F_c$	λ	$\Pi(\lambda)$
1	1	.980	8.72	.0066	11.4408	.998	6.60	.0048	1.4068	.264
	2	.934	120.76	.0001	44.6391	1.000	5.10	.0135	5.3954	.806
	3	1.129	141.61	.0001	100.313	1.000	11.43	.0003	12.1223	.994
1.5	1	1.567	8.27	.0080	5.0848	.882	0.80	.4597	.6252	.138
	2	1.119	98.19	.0001	19.8396	1.000	0.99	.3834	2.3979	.430
	3	1.451	92.29	.0001	44.5834	1.000	6.43	.0054	5.3877	.805
2	1	1.960	0.26	.6126	2.8602	.634	4.02	.0302	.3517	.097
	2	1.868	35.18	.0001	11.1598	.998	1.20	.3170	1.3488	.245
	3	2.257	31.64	.0001	25.0782	1.000	4.08	.0288	3.0306	.530
3	1	3.134	1.82	.1891	1.2712	.320	0.31	.7376	.1563	.070
	2	2.237	32.33	.0001	4.9599	.874	0.14	.8687	.5995	.134
	3	2.903	22.35	.0001	11.1458	.998	2.64	.0907	1.3469	.254
4	1	3.920	0.50	.4856	0.7150	.197	3.00	.0671	.0879	.061
	2	3.737	11.58	.0022	2.7900	.623	0.41	.6671	.3372	.095
	3	4.515	6.19	.0196	6.2695	.942	2.19	.1324	.7576	.158
6	1	6.268	0.34	.5632	.3178	.112	0.17	.8455	.0391	.054
	2	4.474	12.77	.0014	1.2400	.314	0.52	.5987	.1499	.069
	3	5.806	5.23	.0305	2.7864	.622	1.82	.1823	.3367	.095

NOTE: True model is $y_{ij} = \alpha_i + 3X_{ij} + \varepsilon_{ij}$

For both tests, the null hypothesis should be rejected so a large λ is desired. The results from this example show that when a strong collinearity exists, the value of λ and $\Pi(\lambda)$ decreased, causing the wrong decision to be made in the tests of hypothesis. For example, letting $\sigma = 6$ and $W = 1$ in test 2.3.8, resulted in λ equal to 0.3178 and the power of the test equal ot 0.112. Thus, the null hypothesis will only be rejected 11.2% of the time. The collinearity therefore decreased the power of the test because the collinearity caused λ to be small. The power also decreases as the error variance increases. It will be shown in Chapter 3 that a collinearity may cause the estimate of the error to be overestimated.

From Table 2.6.1, one also notes that the p-value for both tests increase as the width decreases and σ increases. In both tests, a small p-value is wanted. Therefore, when the collinearity is strong, the null hypothesis may not be rejected. For example, letting $\sigma = 4$ and $W = 1$, the p-value for test 2.3.8 is .4856 so the null hypothesis would not be rejected.

This example gives an indication of the type of results that may occur in the presence of varying degrees of collinearity. The p-values, λ and the power of the test are affected as expected. Therefore the wrong conclusion is reached for this example for certain values of W and σ.

CHAPTER 3

COLLINEARITY DIAGNOSTICS

3.1 Introduction

Collinearity is a term usually associated with linear regression, and many procedures have been widely used to detect collinearity. The method explained in this chapter is based on a technique described by Belsley, Kuh and Welsch (1980). Though the methods they discuss deal with a linear regression problem, the same procedures may be applied to the analysis of covariance problem which involves several linear regression models.

When there is a relationship between the covariate and treatment, problems will exist in the analysis of covariance as shown in Chapter 2. In this chapter, the detection of this relationship will be discussed. The two principle tools of analysis used in the diagnostic techniques are the singular-value decomposition (SVD) of the design matrix $\underline{X}$ and the decomposition of the estimated variance in a manner corresponding to the SVD.

Before introducing these diagnostic procedures, first note that the covariate data in $\underline{X}$ should be scaled to have unit column length prior to the SVD of $\underline{X}$. The reason for scaling is that scale changes can affect the numerical properties of the data matrix and give very different SVD's. If the columns are scaled to unit length then comparisons of certain diagnostics can be made and provide more stable information. Scaling does not imply that centering should also be done. Centering the data within each treatment level will mask any collinearity between the treatment and covariate.

3.2 Detection of Collinearity

The SVD of a n x p matrix $\underline{X}$ is $\underline{X} = \underline{UDV}'$ where $\underline{U}'\underline{U} = \underline{V}'\underline{V} = \underline{I}_p$ and $\underline{D}$ is a diagonal matrix of nonnegative elements μ_k, k=1,...,p. These μ_k's are known as singular values. These singular values are also the square roots of the characteristic roots of $\underline{X}'\underline{X}$ and the columns of $\underline{V}$ are the corresponding characteristic vectors of $\underline{X}'\underline{X}$. Earlier work in the diagnostic procedures used the characteristic roots and vectors of $\underline{X}'\underline{X}$ to detect collinearity. Belsley, Kuh and Welsch recommend using the SVD of a matrix $\underline{X}$ because the SVD of $\underline{X}$ can be computed with greater numerical stability than finding the characteristic roots and vectors of $\underline{X}'\underline{X}$.

Suppose there are exact linear dependencies in $\underline{X}$, then the rank of $\underline{X}$ is r < p. Since $\underline{U}$ and $\underline{V}$ are orthogonal, then they must be full rank, that is, the rank of $\underline{V}$ and $\underline{U}$ is p. Thus, the rank of $\underline{X}$ is the same as the rank of $\underline{D}$ which implies there are p - r zero elements on the diagonal of $\underline{D}$. Partition $\underline{X}$ as

$$\underline{X} = \underline{UDV}' = [\underline{U}_1, \underline{U}_2] \begin{bmatrix} \underline{D}_{11} & \underline{0} \\ \underline{0} & \underline{0} \end{bmatrix} \begin{bmatrix} \underline{V}_1' \\ \underline{V}_2' \end{bmatrix} \qquad (3.2.1)$$

where $\underline{D}_{11}$ is r x r, $\underline{V}_1$ is p x r and $\underline{V}_2$ is p x (p-r). Then $\underline{XV}_1 = \underline{U}_1\underline{D}_{11}$ and $\underline{XV}_2 = \underline{0}$. The last equation indicates the p - r linear dependencies in $\underline{X}$. Therefore, given p - r linear dependencies, $\underline{D}$ will have p - r zero elements and the variables involved in the dependencies correspond to the nonzero elements of $\underline{V}_2$. In practice, exact collinearities seldom exist. Hence, a small singular value in $\underline{D}$ may indicate a collinearity. The problem is to decide what constitutes a small value.

An extension of this idea involves a measure called the condition number of some n x n nonsingular matrix $\underline{B}$. The larger this condition number, the more *ill-conditioned* $\underline{B}$ is. The spectral norm, $||\underline{B}||$, is defined as

$$||\underline{B}|| \equiv \sup_{||\underline{z}||=1} ||\underline{Bz}||$$

where $||\underline{z}|| \equiv (\underline{z}'\underline{z})^{\frac{1}{2}}$, the Euclidean norm of an n-vector $\underline{z}$. The spectral norm of $\underline{B}$ is equal to the maximum singular value of $\underline{B}$, μ_{max} and $||\underline{B}^{-1}||$ $= 1/\mu_{min}$. The value $||\underline{B}|| \cdot ||\underline{B}^{-1}||$ is defined to be the condition number of $\underline{B}$ and is denoted by $\kappa(\underline{B})$. This number, $\kappa(\underline{B})$, gives a measure of the potential sensitivity of the solution vector $\underline{z}$ of the linear system $\underline{Bz} = \underline{c}$ to small changes in $\underline{B}$ and $\underline{c}$.

To apply this concept to a rectangular matrix $\underline{X}$, recall $\underline{X} = \underline{UDV}'$. Let $\underline{X}^-$ be the generalized inverse of $\underline{X}$ and $\underline{X}^- = \underline{VD}^-\underline{U}'$ where $\underline{D}^-$ is the generalized inverse of $\underline{D}$. The diagonal elements of $\underline{D}^-$ are the nonzero elements of $\underline{D}$ inverted. Therefore, the singular values of $\underline{X}^-$ are the reciprocals of the singular values of $\underline{X}$, and

$$\kappa(\underline{X}) = \frac{\mu_{max}}{\mu_{min}} \geq 1$$

If a near dependency exists in $\underline{X}$, then $\underline{X}$ will have a "small" singular value. This value compared to μ_{max} will then give a measure of the degree of ill-conditioning in $\underline{X}$. If more than one collinearity exists, then there would be the same number of "small" singular values in $\underline{X}$. Define the kth condition index of $\underline{X}$ as

$$\eta_k \equiv \frac{\mu_{max}}{\mu_k} \qquad k=1,\ldots,p. \tag{3.2.2}$$

Therefore, the number of high condition indexes indicates the number of collinearities in $\underline{X}$.

Before the problem in detecting a collinearity was deciding how small a singular value should be before it is an indication of a collinearity. Now, the problem is in deciding how large a condition index should be. Based on some experimental studies, condition indexes around 5 to 10 are associated with a weak collinearity and condition indexes greater than 30 imply moderate to strong collinearities.

The previous discussion dealt with how a collinearity may be detected. The next step it to determine where the collinearity exists. It was stated previously that the variables involved in the collinearity correspond to the nonzero elements of $\underline{V}_2$, where $\underline{V}_2$ is shown in the partition of $\underline{X}$, (3.2.1) If there is a small singular value in $\underline{D}$ or a high condition index, then one collinearity might exist in $\underline{X}$. One possible method to determine what variables might be involved in the collinearity is to check the columns of $\underline{V}'$ that correspond to the small singular value for nonzero elements.

Rather than just using the $\underline{V}'$ matrix though, elements from $\underline{V}'$ and $\underline{D}$ will be used and will provide information on the extent to which near dependencies degrade the estimate of the variance of the parameters. Another measure is introduced that used the variance-covariance matrix of the least-squares estimator of $\underline{b}$ where $\underline{y} = \underline{Xb} + \underline{\varepsilon}$. The least-squares estimator of $\underline{b}$ is

$$\hat{\underline{b}} = \underline{X}^{-}\underline{y} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$$

Using the SVD of $\underline{X}$, the variance-covariance matrix of $\hat{\underline{b}}$ may be written as

$$\mathrm{Var}(\hat{b}) = \sigma^2(\underline{X}'\underline{X})^{-1} = \sigma^2\underline{VD}^{-2}\underline{V}' \qquad (3.2.3)$$

where σ^2 is the variance of the elements of $\underline{\varepsilon}$.

Note that (3.2.3) may be expressed for each kth element of $\hat{\underline{b}}$ as

$$\mathrm{Var}(\hat{\alpha}_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} \qquad (3.2.4)$$

where μ_j's are the singular values of $\underline{X}$ and v_{ij}'s are the elements of $\underline{V}$. The var(α_k) in (3.2.4) is therefore a sum of elements, each associated with one of the p singular values μ_j. If there is a near dependency, that is a small singular value, then the jth element of (3.2.4) would be large compared to the other components, other things being equal. Thus, a high proportion of the variance of two or more parameters associated with the same small singular value suggests that the corresponding near dependency is causing problems.

Define ϕ_{kj} and ϕ_k as

$$\phi_{kj} \equiv \frac{v_{kj}^2}{\mu_j^2} \qquad \text{and} \qquad \phi_k \equiv \sum_{j=1}^{p} \phi_{kj} \qquad k=1,\ldots,p$$

Then, the kth variance-decomposition proportion is

$$\pi_{jk} \equiv \frac{\phi_{kj}}{\phi_k} \qquad \begin{array}{l} k=1,\ldots,p \\ j=1,\ldots,p \end{array} \qquad (3.2.5)$$

which is the proportion of the variance of the kth parameter associated with the jth element of it decomposition (3.2.4). To determine which parameters are then associated with a collinearity, look at the π_{jk}, for k=1,...,p corresponding to the small singular value μ_j or the large condition index η_j. The larger values of π_{jk} will indicate the parameters that are related. Define the p x p matrix $\underline{\Pi}$ as the matrix containing the elements π_{jk}.

3.3 Applications of Diagnostics With Generated Data

The data used in this section were generated as shown in (2.5.2). To apply the techniques described previously, and examination of the $\underline{X}_3$ (2.2.8) and the $\underline{X}_5$ (2.2.10) matrices was done. Recall different severities of collinearity between the covariate and treatment were imposed by varying W. Table 3.3.1 gives the values generated for $\underline{x}_1$, $\underline{x}_2$ and $\underline{x}_3$ for W=1,2 and 3. The dependency between treatment and the covariate is obvious.

After scaling the data so the columns containing the x_{ij} values have unit length, the SVD of $\underline{X}_3$ and $\underline{X}_5$ was computed and the conditions indexes, η_j and variance-decomposition proportions, π_{jk}, were evaluated for each W. Tables (3.3.2), (3.3.3) and (3.3.4) show the condition indexes and the variance-decomposition proportions of both $\underline{X}_3$ and $\underline{X}_5$ for W=1,2, and 3 respectively.

First, look at the results for $\underline{X}_3$ with W=1. Note that $\eta_1 = 76.71$, which indicates a strong collinearity. The next step is to look at row 1 of $\underline{\Pi}$ to determine which parameters are associated with the collinearity. All the values in row 1 are large relative to the other values in the table, therefore the dependency is between all the parameters. This says that the three levels of the treatment are related to the covariate. This is how the data were generated.

The results in tables 3.3.3 and 3.3.4 indicate that the severity of the collinearity decreases as W increases. This was also seen in the calculations of λ in chapter 2.

The second part of table 3.3.2 indicates a slightly different interpretation of the collinearity. For $\underline{X}_5$, there are three high values of the condition indexes, which indicate three collinearities in this matrix. To determine where the dependencies exist, look at the first three rows of the $\underline{\Pi}$ matrix. The first row shows parameters 3 and 6 are involved in one of the collinearities. This says the third level of the treatment and the covariate values for this third treatment level are related. This is what

Table 3.3.1

Generated Covariate Values for W = 1, 2, and 3

	W = 1	W = 2	W = 3
$\underline{x}_1$	1.773	1.554	3.937
	1.166	1.869	3.507
	1.957	2.680	2.457
	1.829	1.568	3.696
	1.976	2.708	3.561
	1.793	1.283	3.995
	1.793	1.784	3.900
	1.683	2.421	2.156
	1.194	2.988	1.453
	1.530	1.390	1.851
$\underline{x}_2$	5.399	5.104	7.531
	5.272	5.181	6.316
	5.455	5.850	6.892
	5.819	6.106	6.015
	5.632	5.720	5.502
	5.040	5.193	6.480
	5.450	5.175	6.817
	5.057	5.240	5.435
	5.966	5.960	5.471
	5.220	6.934	7.337
$\underline{x}_3$	9.952	9.264	10.227
	9.136	9.215	11.454
	9.565	9.497	9.086
	9.129	9.172	11.607
	9.161	10.430	10.415
	9.500	10.689	11.704
	9.864	10.437	9.971
	9.970	9.120	11.182
	9.518	9.599	11.704
	9.681	9.219	9.495

Table 3.3.2

Condition Indexes and Variance-Decomposition Matrices

For W = 1

Results for $\underline{X}_3$

$$\underline{\eta} = \begin{bmatrix} 76.71 \\ 1.05 \\ 1.05 \\ 1 \end{bmatrix} \qquad \Pi = \begin{bmatrix} .92 & .99 & 1.00 & 1.00 \\ .08 & .00 & .00 & 0 \\ .00 & .00 & .00 & 0 \\ .00 & .00 & .00 & .00 \end{bmatrix}$$

Results for $\underline{X}_5$

$$\underline{\eta} = \begin{bmatrix} 107.42 \\ 65.17 \\ 21.56 \\ 1.00 \\ 1.00 \\ 1 \end{bmatrix} \qquad \Pi = \begin{bmatrix} 0 & 0 & 1.00 & 0 & 0 & 1.00 \\ 0 & 1.00 & 0 & 0 & 1.00 & 0 \\ .98 & 0 & 0 & 1.00 & 0 & 0 \\ .02 & 0 & 0 & .00 & 0 & 0 \\ 0 & .00 & 0 & 0 & .00 & 0 \\ 0 & 0 & .00 & 0 & 0 & .00 \end{bmatrix}$$

Table 3.3.3

Condition Indexes and Variance-Decomposition Matrices

For W = 2

Results for $\underline{X}_3$

$$\underline{\eta} = \begin{bmatrix} 39.85 \\ 1.05 \\ 1.05 \\ 1 \end{bmatrix} \qquad \underline{\Pi} = \begin{bmatrix} .81 & .97 & .99 & 1.00 \\ .00 & .02 & .00 & 0 \\ .19 & .00 & .00 & 0 \\ .00 & .01 & .01 & .00 \end{bmatrix}$$

Results for $\underline{X}_4$

$$\underline{\eta} = \begin{bmatrix} 58.03 \\ 35.33 \\ 12.49 \\ 1.00 \\ 1.00 \\ 1 \end{bmatrix} \qquad \underline{\Pi} = \begin{bmatrix} 0 & 0 & 1.00 & 0 & 0 & 1.00 \\ 0 & .99 & 0 & 0 & 1.00 & 0 \\ .94 & 0 & 0 & 1.00 & 0 & 0 \\ .06 & 0 & 0 & .00 & 0 & 0 \\ 0 & .01 & 0 & 0 & .00 & 0 \\ 0 & 0 & .00 & 0 & 0 & .00 \end{bmatrix}$$

Table 3.3.4

Condition Indexes and Variance-Decomposition Matrices

For W = 3

Results for $\underline{X}_3$

$$\underline{\eta} = \begin{bmatrix} 30.02 \\ 1.05 \\ 1.05 \\ 1 \end{bmatrix} \quad \underline{\Pi} = \begin{bmatrix} .81 & .95 & .98 & 1.00 \\ .00 & .04 & .01 & 0 \\ .18 & .00 & .00 & 0 \\ .01 & .01 & .01 & .00 \end{bmatrix}$$

Results for $\underline{X}_4$

$$\underline{\eta} = \begin{bmatrix} 40.37 \\ 30.56 \\ 12.03 \\ 1.00 \\ 1.00 \\ 1 \end{bmatrix} \quad \underline{\Pi} = \begin{bmatrix} 0 & 0 & .99 & 0 & 0 & 1.00 \\ 0 & .99 & 0 & 0 & 1.00 & 0 \\ .93 & 0 & 0 & 1.00 & 0 & 0 \\ .07 & 0 & 0 & .00 & 0 & 0 \\ 0 & .01 & 0 & 0 & .00 & 0 \\ 0 & 0 & .01 & 0 & 0 & .00 \end{bmatrix}$$

is expected because the data were generated to have this relationship. Rows two and three of $\underline{\Pi}$ show dependencies between the second and first levels of the treatment with their respective covariate values.

The results for the other widths show the same relationships, but again the condition indexes decrease as W increases.

These diagnostics may be extended to a more complicated model such as a two-way cross classification structure with one covariate. In the previous one-way model, detecting a collinearity may be done visually by plotting the data, whereas in a more complicated model, that may not be the case.

For a two-way cross classification structure with one covariate and parallel regression lines, the mean model is

$$y_{ijk} = \mu_{ij} + \beta X_{ijk} + \varepsilon_{ijk} \qquad (3.3.1)$$

$$i=1,\ldots I \qquad j=1,\ldots J \qquad k=1,\ldots K$$

where $\varepsilon_{ijk} \sim N(0,\sigma^2)$ for all i, j, k,

ε_{ijk} is the random experimental error,

μ_{ij} denotes the intercept of the line for level i of factor A and level j of factor B.

β denotes the regression slope,

X_{ijk} denotes the value of the covariate for the kth individual in the ij^th treatment combination,

y_{ijk} denotes the kth observation from the ijth treatment combination.

The types of collinearities that might occur in (3.3.1) are where the main effect A or B is related to the covariate or where certain cells are related with the covariate, that is the interaction and covariate are collinear. Let I = 3 and J = 3, then the first colinnearity situation may be depicted as

	B1	B2	B3
A1	L	L	L
A2	M	M	M
A3	H	H	H

(3.3.2)

where L, M, and H denote low, medium and high values of the covariate, and Ai and Bj denote the ith and jth level of factors A and B respectively. In (3.3.2), the collinearity is between the three levels of factor A and the covariates.

In the second type of collinearity stated above, the situation may be depicted as

	B1	B2	B3
A1	H	L	R
A2	L	H	R
A3	R	R	R

(3.3.3)

where L and H denote low and high values of the covariate and R denotes random values of the covariate.

Data were generated for the analysis of covariance model in which the collinearities follow the pattern in (3.3.3). For this situation, x_{ijk} values were generated for I = 3, J = 3 and K = 5, and are shown in table (3.3.5). The diagnostic routines were run on two different matrices. The first matrix, $\underline{M}$, contains the usual design part, that is 1's and 0's, augmented with a vector of the generated x_{ijk} values. The second matrix, $\underline{N}$, also consists of the design matrix but the x_{ijk} values are now represented in a 45 x 9 matrix with $\underline{x}_{ij.}$ vectors on the diagonals. Which matrix is used in the actual covariance analysis depends on whether a different slope is necessary for every treatment combination.

For matrix $\underline{M}$, the highest condition index is $\eta_1 = 8.029$. The other condition indexes are less than or equal to 1.087. This implies one weak collinearity. Following is the first row of $\underline{\Pi}$

$$\pi_{ij}: \quad .734 \quad .048 \quad .412 \quad .050 \quad .732 \quad .262 \quad .374 \quad .363 \quad .556 \quad .997 \tag{3.3.4}$$

where the treatment combination A1B1 is denoted by the first column, A1B2 is denoted by the second column, etc. The last column denotes the covariate.

The large values in row (3.3.4), as compared to the other values are in columns 1, 5 and 10. This implies that the collinearity is between treatment combination A1B1 and A2B2 and the covariate. The imposed collinearity was this plus the treatment combinations A2B1 and A1B2. Thus, even though a weak collinearity is detected, not all the dependencies were

Table 3.3.5

Generated Covariate Values for Situation (3.3.3)

	B1	B2	B3
A1	10.55 10.28 10.02 10.50 10.36	1.87 1.12 1.30 1.82 1.16	9.59 4.80 1.61 6.45 4.48
A2	1.17 1.51 1.75 1.24 1.79	10.92 10.01 10.11 10.01 10.41	6.32 2.67 0.02 8.16 2.07
A3	0.30 3.50 2.05 9.60 9.44	5.89 7.63 4.78 4.31 1.74	1.84 9.48 9.27 9.52 5.57

identified. It is interesting to note that the values in $\underline{\Pi}$ corresponding to A2B1 and A1B2 are the smallest values in row 1.

For matrix $\underline{N}$, all condition indexes greater than 5 are

$$\underline{\eta}' = [146.86 \quad 79.57 \quad 16.04 \quad 12.31 \quad 7.15 \quad 6.77 \quad 6.05] \qquad (3.3.5)$$

From these results, there appears to be two strong collinearities and two moderate collinearities, and three weak collinearities. The first four rows of $\underline{\Pi}$ indicate that the moderate to strong dependencies are between columns 1 and 10, 5 and 14, 4 and 13, and 2 and 11. Recall columns 1 through 9 denote the AiBj treatment combinations and columns 10 through 18 are the corresponding $\underline{x}_{ij\cdot}$ vectors. The information from the first four rows of $\underline{\Pi}$ indicates the collinearities that are shown in (3.3.3). The weak collinearities indicated in (3.3.5) were also investigated. Rows 5, 6 and 7 in $\underline{\Pi}$ indicated these weak dependencies are between columns 8 and 17, 9 and 18, and 3 and 12. The x_{ijk} values in columns 17, 18 and 12 were randomly generated from a uniform (0,10), and from table 3.3.5, there does not appear to be any strong similarities in the five values of these cells.

By using the matrix $\underline{M}$, a collinearity was detected but not all the relationships were identified. The diagnostics for the larger design matrix, $\underline{N}$, detected and identified the imposed collinearities, but other weak relationships were indicated. This suggests that both types of matrices should be examined to aid the investigator in detecting and identifying collinearities.

To further illustrate the detection techniques, consider the pattern of collinearity in (3.3.6).

	B1	B2	B3
A1	H	L	H
A2	L	H	L
A3	H	L	H

(3.3.6)

The x_{ijk} values that were generated for (3.3.6) are reported in table (3.3.6). Again the diagnostic methods were applied to design matrix $\underline{M}$ and the larger design matrix $\underline{N}$.

The results for $\underline{M}$ indicate that one strong collinearity exists in this matrix. The highest condition index is 72.740 and the other indexes are less that 1.10. All the values in the first row of $\underline{\Pi}$ are greater than 0.7 which implies the dependencies are among all the columns of $\underline{M}$. This is the case for the situation depicted in (3.3.6).

Following are the condition indexes computed for matrix $\underline{N}$, that are greater than 1.003.

$$\underline{\eta} = [146.86\ 108.12\ 94.53\ 79.58\ 73.29\ \ 20.65\ 16.04\ 12.79\ 12.31] \qquad (3.3.7)$$

The first nine rows of the corresponding $\underline{\Pi}$ matrix indicated the dependencies are between columns 1 and 10, 3 and 12, 9 and 18, 5 and 14, 7 and 16, 8 and 17, 4 and 13, 6 and 15, and 2 and 11. These are the collinearities shown in (3.3.6).

Table 3.3.6

Generated Covariate Value For Situation in (3.3.6)

	B1	B2	B3
A1	10.55 10.28 10.02 10.50	1.87 1.12 1.30 1.82	10.96 10.48 10.16 10.64
A2	1.17 1.51 1.75 1.24	10.92 10.01 10.11 10.01	1.63 1.27 1.00 1.82
A3	10.03 10.35 10.20 10.96 10.94	1.59 1.76 1.48 1.43 1.17	10.18 10.95 10.93 10.95 10.56

Chapter 4

Examples of Collinearity in Analysis of Covariance

4.1 Introduction

In this chapter two data sets are discussed to illustrate the diagnostic procedures explained in Chapter 3. One data set is an exercise in a popular statistical methods book. The other came from a study conducted by an education researcher at Kansas State University.

4.2 Example I

The data for this example appear in a statistical textbook by Ott (1977, p.621). The purpose of the experiment given in the exercise was to study the effects of three different antidepressants (A, B and C) on patients. Three patients were randomly selected from each of six age-sex combinations. A pretreatment rating of depression was taken from each patient on the day the study was to begin. The assigned antidepressants were then given for one week, and a second rating (posttreatment) was taken at the end of the week. The covariate in this experiment was the pretreatment rating. The data set is given in Table 4.2.1.

Since drug A was administered to patients with high pretreatment ratings and drug C was administered to patients with low pretreatment ratings, there appears to be a collinearity between the treatment and the covariate.

First, the analysis of covariance results are presented, then the

Table 4.2.1

Data For Example 4.2

		Pretreatment			Posttreatment		
Sex	Age (years)	A	B	C	A	B	C
F	<20	48	36	31	21	25	17
F	20-40	43	31	28	22	21	19
F	>40	44	35	29	18	24	18
M	<20	42	38	29	26	20	17
M	20-40	37	34	28	21	24	15
M	>40	41	36	26	18	24	19

Table 4.2.2

Analysis of Covariance Table
Example 4.2

Source	df	SS	MS	F	Pr>F
Drug	2	67.48	33.74	6.28	.0113
Covariate	1	0.28	0.28	.05	.8219
Error	14	75.22	5.37		

results from the diagnostic procedures are given. The six age-sex combinations are blocks, so this model is a randomized block design with one factor and one covariate. Since a preliminary analysis indicated that the blocking factor was not significant, it was eliminated from the model. Thus, this example is consistent with models discussed previously.

The first model fitted to the data was the unequal slope model given in (2.2.1). The calculated F for the test of equal slopes is F = 0.11 with a p-value of 0.8997. Therefore, one cannot reject the null hypothesis of equal slopes. The next fitted model was the parallel line or equal slope model (2.2.2). The two tests of interest associated with this model, tests (2.3.3) and (2.3.4) as summarized in Table 4.2.2. Since the appropriate test statistic for test (2.3.3) has a p-value of .8219, one cannot reject the null hypothesis that β is equal to 0 given there is a treatment difference. This implies that a covariate is not necessary and a one-way analysis of variance model is adequate. With the covariate in the model, the test shows that there are significant differences in the regression lines, thus indicating treatment effects. When the analysis is done without the covariate, the calculated F value is 9.24 with a p-value of 0.0024, again indicating significant treatment differences but with a lower p-value.

Based on the analysis above, a covariate is not necessary in the model. However, if the test of β equal to 0 is done prior to assuming any treatment effect, the results are quite different. For this test the calculated F value is 4.80 with a p-value of .0458. This implies there is a relationship between the dependent variable Y and the covariate X. Thus the order of testing makes a difference in deciding whether or not to include a covariate in the model. In such situations one may suspect that a collinearity is causing the conflicting conclusions from the two analyses.

By applying the diagnostic techniques from Chapter 3, the condition indexes and the variance-decomposition proportion matrix shown in Table 4.2.3 were obtained for the full design matrix. There are three large condition indexes from this data set, implying three strong collinearities. From the first three rows of $\underline{\Pi}$, the collinearities are between each level of the treatment and the corresponding covariate. This is clearly seen in the data.

Since these collinearities exist, the first test of hypothesis may be yielding incorrect results. In Chapter 2, it was pointed out that when a collinearity is present, the noncentrality parameter for that test gets closer to zero. Therefore, not rejecting the hypothesis that each regression line has an equal slope, may be the incorrect conclusion. The same reasoning also applies when the null hypothesis of $\beta = 0$, given there is a difference between regression lines, is not rejected.

There are some alternatives to the analysis done above. It is known that centering and standardizing the covariate data within each treatment group will remove the collinearity. Doing the analysis with the centered data may be a viable option, but it is an alternative analysis that needs further investigation. The analysis with the centered and standardized data for this example gave the same conclusions as the analysis on the raw data, which may imply that the covariate is not necessary in this problem. A plot of the data in Figure 4.2.1 also indicates this. Within each

Table 4.2.3

Condition Indexes and Π -Matrix for Example 4.2

$$\underline{\eta} = \begin{bmatrix} 54.3538 \\ 46.3701 \\ 36.8503 \\ 1.0002 \\ 1.0001 \\ 1 \end{bmatrix}$$

$$\underline{\Pi} = \begin{bmatrix} 0 & 0 & .998 & 0 & 0 & 1.000 \\ 0 & .997 & 0 & 0 & 1.000 & 0 \\ .996 & 0 & 0 & 1.000 & 0 & 0 \\ .004 & 0 & 0 & .000 & 0 & 0 \\ 0 & .003 & 0 & 0 & .000 & 0 \\ 0 & 0 & .002 & 0 & 0 & .000 \end{bmatrix}$$

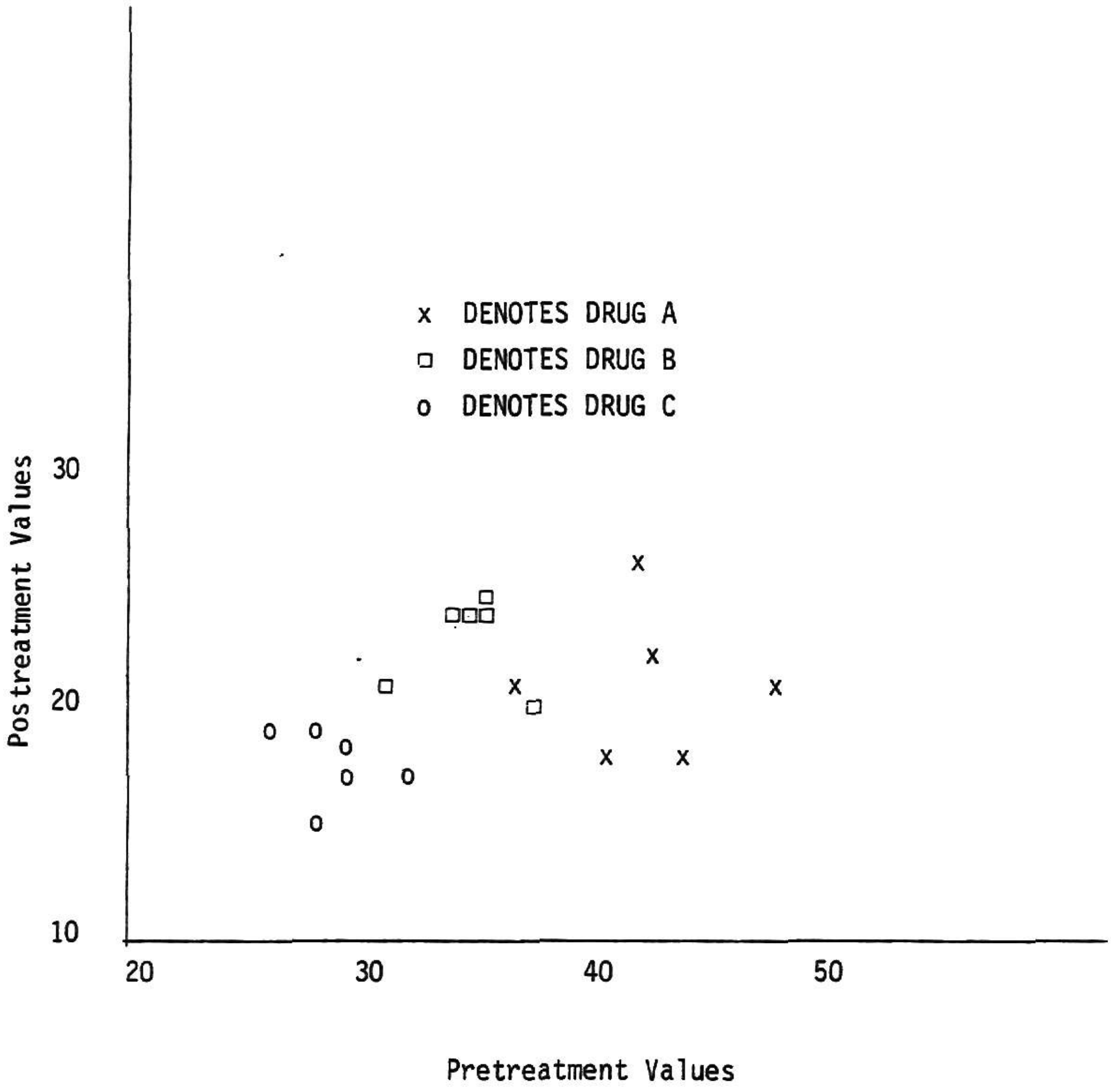

FIGURE 4.2.1 PLOT OF RAW DATA FROM EXAMPLE 4.2

treatment group, there does not appear to be any relationship between the dependent variable Y and the covariate X.

Another option one might consider when a collinearity exists is to collect additional data that would provide the needed independence between the covariate and treatment. This may not always be possible, though. Either the cost and time involved in collecting the additional data may be prohibitive or else the collinearity may be a physical phenomenon that could not be removed by the collection of more data.

Other method of analyzing a data set with collinearity problems are Bayesian-type techniques or true-score analysis of covariance. The first method has been suggested in the case of a collinearity in regression analysis and encompasses the technique of ridge regression. The second analysis is explained in Huitema (1980) and involves using estimated true covariance values rather than observed covariate scores. Both of these methods need more investigation before any comment can be made on their validity in solving the collinearity problem in the analysis of covariance model.

4.3 Example II

This example is a result of an education research project where five different teaching conditions were used and scores for the five were recorded. Blocking was done on five schools and classes within schools. The students were classified by sex and age (two levels each). Each student was administered two tests after hearing a story under one of the five conditions.

One test was given immediately after the story was told and the other was given 48 hours later. The order of the two tests were random, so half the students took test 1 immediately whereas the other half took test 2 immediately. On each test, there was an explicit part and an implicit part. The two covariates in this problem were the implicit and explicit immediate test scores (IMEXP and IMIMP). The two response variables were the scores on the explicit and implicit parts of the delayed test.

The analysis of this data set is not presented here due to the complexity and length of the problem. The reader is referred to Hathaway (1981) for a discussion of the analysis and the results.

The interesting thing about this example from the point of view of the present study is that the results were not consistent with what the researcher expected. A collinearity may explain the inconsistencies in the analysis but since this is such a complicated model, visual inspection of the data does not give any information on the existence of any collinearities.

The problem is to determine whether there are any collinearities between either of the two covariates and any treatment combination. The four factors that make up the treatment combinations are teaching condition, sex, age, and the test that was given immediately after the story was told. The experiment was designed such that every treatment combination had two observations each. Thus the design matrix used for the diagnostic procedure had 42 columns; the first 40 columns represented the treatment combinations and the last 2 contained the two covariates.

There were two condition indexes out of the 42 that were greater than

1.4. One was 11.543 and the other was 10.1105 which indicates two possible collinearities. The first row of $\underline{\Pi}$ showed that the first collinearity is between covariate IMEXP and the following two treatment combinations: condition 5, sex 1, age 1, test 2 and condition 1, sex 1, age 1, test 2. The second collinearity is between the covariate IMIMP and 17 different treatment combinations.

In a design as complex as this, determining exactly what is occurring between the treatment combinations and the covariate may be impossible. The important aspect of this exercise is to know there are two possible collinearities and that these may influence the analysis of covariance results. Other methods of analysis, like those mentioned in section 4.2, may be applied to these data to resolve the inconsistencies found in the initial analysis.

Chapter 5

Summary

In this report the problem of a collinearity between the treatment and the covariate in an analysis of covariance model was investigated. The detection of this collinearity was also discussed.

Five models were assumed to be the true model and five different tests of hypothesis were examined under each true model. Using the principle of conditional error approach, λ, the noncentrality parameter, was calculated for each model and test combination. It was shown in section 2.4 that a collinearity may cause a nonzero λ to approach zero in two ways. In one case, the collinearity causes the design matrix for the true model to be in the column space or "almost" in the column space of the two design matrices for the models in the hypotheses. The other way λ is affected is that a class of solutions to $\underline{Vb} = \lambda$ for λ close to zero exists because of collinearity. This solution does not satisfy the null hypothesis.

A testing strategy may be comprised of any combinations of the five tests of hypothesis examined. The purpose of an analysis of covariance is to determine the appropriate model. This entails deciding if a covariate should be included in the model and if it is, then determine whether the regression lines are parallel or not. Finally, one tests whether differences between the lines are significant. Thus, a testing strategy is employed to determine the best covariance model.

It was shown that when a collinearity exists, some strategies will lead to wrong conclusions concerning covariate effects and differences between regression lines. This is due to the fact that a collinearity may cause λ to approach zero even when the model under the null hypothesis is not the true model or a special case of the true model. Suppose, if in the first step of the testing strategy, the null hypothesis was not rejected and a collinearity was present. One would not know if this decision was a result of the collinearity or because the model under the null hypothesis was the true model or a special case of the true model.

Several data sets were generated for a particular model and using these data sets, λ and the power were calculated for two of the tests. The results from these data showed that λ and the power decreased as the collinearity became more severe and the variance increased.

The diagnostic methods used to detect a collinearity were presented in Section 3.2. It was shown that a collinearity in a data set results in an ill-conditioned design matrix and condition indexes are calculated to indicate the severity of the ill-conditioning. These diagnostic procedures were applied to computer generated data with imposed collinearities. The results from a one-factor parallel line model with one covariate indicated the collinearity and the parameters involved in the collinearity. For the parallel line model with two factors and one covariate, the diagnostic procedures detected the collinearity but failed to identify all the parameters involved with the collinearity. In another example with the same model but a different collinearity, the diagnostic procedures detected the collinearity and identified the parameters associated with the collinearity.

These diagnostic procedures are important to a researcher with an analysis of covariance problem because he can determine if a collinearity exists. From this he knows whether the results from his analysis of covariance are valid.

Two examples were presented in Chapter 4. Possible collinearities were detected in these data and other methods of analyses were mentioned.

Good design practices should insure that the covariate and treatments are independent. This study explained why independence is important and how one can check whether their experiment satisfies this condition. The problem still exists concerning how the data should be analyzed when this condition of independence does not exist. This problem needs additional research and a starting point are those methods mentioned in Section 4.2.

References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, Inc., New York.

Cochran, W.G. and Cox, G.M. (1957) Experimental Designs. John Wiley and Sons, Inc., New York.

Graybill, F. (1969). Introduction to Matrices with Applications in Statistics. Wadsworth Publishing Company, Belmont, California.

Graybill, F. (1976). Theory and Application of the Linear Model. Duxbury Press, Belmont, California.

Handbook of Mathematical Functions, (1964). Edited by M. Abramowitz and I.A. Stegun. National Bureau of Standards Applied Mathematics Series 55.

Hathaway, B.K. (1981). The Effects of Visual and Motor Elaborations on Preschool Children's Recall and Comprehension of Prose. Doctoral Dissertation, Kansas State University, Manhattan, Kansas.

Huitema, B.E. (1980). The Analysis of Covariance and Alternatives. John Wiley and Sons, Inc., New York.

Milliken, G.A. (1971) New Criteria for Estimability for Linear Models. Ann. Math. Stat., 42:1588-1594.

Milliken, G.A. and Johnson, D.E. (1983) The Analysis of Messy Data. Lifetime Publications, Belmont, California.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. McGraw-Hill, Inc., New York.

Ott, L. (1977) An Introduction to Statistical Methods and Data Analysis. Duxbury Press, Belmont, California.

SAS User's Guide, 1979 Edition (1979). SAS Institute Inc., Raleigh, North Carolina.

Searle, S.R. Linear Models. John Wiley and Sons, Inc., New York.

THE EFFECTS AND DETECTION OF COLLINEARITY IN AN ANALYSIS OF COVARIANCE

by

JO JANE GIACOMINI

B.S., Colorado State University, 1977

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

ABSTRACT

This report considers the problem of collinearity between the treatment and the covariate in an analysis of covariance model. The effects of the collinearity are examined by looking at the noncentrality parameter under certain true models for five different tests of hypothesis. The effect of the collinearities on the noncentrality parameter is explained and examples are given to show how the power of the test decreases as the collinearity gets stronger.

Methods for detecting the collinearity between the treatment and the covariate are given. These methods are based on the singular-value decomposition of the design matrix. One measure, the condition index, indicates the presence of a collinearity, while a matrix of variance-decomposition proportions shows which parameters are associated with the collinearity. The diagnostic methods are applied to simulated data with different collinearities.

Two real data sets are presented in which the assumption of independence between the treatment and covariate appear to be violated. The diagnostic procedure gave results that indicate that collinearities exist among the treatments and covariates.