CHAPTER 9

ANALYSIS OF COLLINEAR DATA

9.1 INTRODUCTION

Interpretation of the multiple regression equation depends implicitly on the assumption that the predictor variables are not strongly interrelated. It is usual to interpret a regression coefficient as measuring the change in the response variable when the corresponding predictor variable is increased by one unit and all other predictor variables are held constant. This interpretation may not be valid if there are strong linear relationships among the predictor variables. It is always conceptually possible to increase the value of one variable in an estimated regression equation while holding the others constant. However, there may be no information about the result of such a manipulation in the estimation data. Moreover, it may be impossible to change one variable while holding all others constant in the process being studied. When these conditions exist, simple interpretation of the regression coefficient as a marginal effect is lost.

When there is a complete absence of linear relationship among the predictor variables, they are said to be *orthogonal.* In most regression applications the predictor variables are not orthogonal. Usually, the lack of orthogonality is not serious enough to affect the analysis. However, in some situations the predictor variables are so strongly interrelated that the regression results are ambiguous.

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

Typically, it is impossible to estimate the unique effects of individual variables in the regression equation. The estimated values of the coefficients are very sensitive to slight changes in the data and to the addition or deletion of variables in the equation. The regression coefficients have large sampling errors, which affect both inference and forecasting that is based on the regression model.

The condition of severe nonorthogonality is also referred to as the problem of collinear data, or *multicollinearity*. The problem can be extremely difficult to detect. It is not a specification error that may be uncovered by exploring regression residual. In fact, multicollinearity is not a modeling error. It is a condition of deficient data. In any event, it is important to know when multicollinearity is present and to be aware of its possible consequences. It is recommended that one should be very cautious about any and all substantive conclusions based on a regression analysis in the presence of multicollinearity.

This chapter focuses on three questions:

1. How does multicollinearity affect statistical inference and forecasting?
2. How can multicollinearity be detected?
3. What can be done to resolve the difficulties associated with multicollinearity?

When analyzing data, these questions cannot be answered separately. If multicollinearity is a potential problem, the three issues must be treated simultaneously by necessity.

The discussion begins with two examples. They have been chosen to demonstrate the effects of multicollinearity on inference and forecasting, respectively. A treatment of methods for detecting multicollinearity follows and the chapter concludes with a presentation of methods for resolving problems of multicollinearity. The obvious prescription to collect better data is considered, but the discussion is mostly directed at improving interpretation of the existing data. Alternatives to the ordinary least squares estimation method that perform efficiently in the presence of multicollinearity are considered in Chapter 10.

9.2 EFFECTS ON INFERENCE

This first example demonstrates the ambiguity that may result when attempting to identify important predictor variables from among a linearly dependent collection of predictor variables. The context of the example is borrowed from research on equal opportunity in public education as reported by Coleman et al. (1966), Mosteller and Moynihan (1972), and others.

In conjunction with the Civil Rights Act of 1964, the Congress of the United States ordered a survey "concerning the lack of availability of equal educational opportunities for individuals by reason of race, color, religion or national origin in public educational institutions...." Data were collected from a cross-section of school districts throughout the country. In addition to reporting summary statistics on variables such as level of student achievement and school facilities, regression analysis was used to try to establish factors that are the most important determinants of achievement. The data for this example consist of measurements taken in 1965 for 70 schools selected at random. The data consist of variables that measure student achievement, school facilities, and faculty credentials. The objective is to evaluate the effect of school inputs on achievement.

Assume that an acceptable index has been developed to measure those aspects of the school environment that would be expected to affect achievement. The index includes evaluations of the physical plant, teaching materials, special programs, training and motivation of the faculty, and so on. Achievement can be measured by using an index constructed from standardized test scores. There are also other variables that may affect the relationship between school inputs and achievement. Students' performances may be affected by their home environments and the influence of their peer group in the school. These variables must be accounted for in the analysis before the effect of school inputs can be evaluated. We assume that indexes have been constructed for these variables that are satisfactory for our purposes. The data are given in Tables 9.1 and 9.2, and can also be found in the book's Web site.[1]

Adjustment for the two basic variables (achievement and school) can be accomplished by using the regression model

$$\text{ACHV} = \beta_0 + \beta_1 \cdot \text{FAM} + \beta_2 \cdot \text{PEER} + \beta_3 \cdot \text{SCHOOL} + \varepsilon. \qquad (9.1)$$

The contribution of the school variable can be tested using the t-value for β_3. Recall that the t-value for β_3 tests whether SCHOOL is necessary in the equation when FAM and PEER are already included. Effectively, the model above is being compared to

$$\text{ACHV} - \beta_1 \cdot \text{FAM} - \beta_2 \cdot \text{PEER} = \beta_0 + \beta_3 \cdot \text{SCHOOL} + \varepsilon, \qquad (9.2)$$

that is, the contribution of the school variable is being evaluated after adjustment for FAM and PEER. Another view of the adjustment notion is obtained by noting that the left-hand side of (9.1) is an adjusted achievement index where adjustment is accomplished by subtracting the linear contributions of FAM and PEER. The equation is in the form of a regression of the adjusted achievement score on the SCHOOL variable. This representation is used only for the sake of interpretation. The estimated β's are obtained from the original model given in Equation (9.1). The regression results are summarized in Table 9.3 and a plot of the residuals against the predicted values of ACHV appears as Figure 9.1.

Checking first the residual plot we see that there are no glaring indications of misspecification. The point located in the lower left of the graph has a residual value that is about 2.5 standard deviations from the mean of zero and should possibly be looked at more closely. However, when it is deleted from the sample, the regression

[1] http://www.ilr.cornell.edu/~hadi/RABE4

Table 9.1 First 50 Observations of the Equal Educational Opportunity (EEO) Data; Standardized Indexes

Row	ACHV	FAM	PEER	SCHOOL
1	−0.43148	0.60814	0.03509	0.16607
2	0.79969	0.79369	0.47924	0.53356
3	−0.92467	−0.82630	−0.61951	−0.78635
4	−2.19081	−1.25310	−1.21675	−1.04076
5	−2.84818	0.17399	−0.18517	0.14229
6	−0.66233	0.20246	0.12764	0.27311
7	2.63674	0.24184	−0.09022	0.04967
8	2.35847	0.59421	0.21750	0.51876
9	−0.91305	−0.61561	−0.48971	−0.63219
10	0.59445	0.99391	0.62228	0.93368
11	1.21073	1.21721	1.00627	1.17381
12	1.87164	0.41436	0.71103	0.58978
13	−0.10178	0.83782	0.74281	0.72154
14	−2.87949	−0.75512	−0.64411	−0.56986
15	3.92590	−0.37407	−0.13787	−0.21770
16	4.35084	1.40353	1.14085	1.37147
17	1.57922	1.64194	1.29229	1.40269
18	3.95689	−0.31304	−0.07980	−0.21455
19	1.09275	1.28525	1.22441	1.20428
20	−0.62389	−1.51938	−1.27565	−1.36598
21	−0.63654	−0.38224	−0.05353	−0.35560
22	−2.02659	−0.19186	−0.42605	−0.53718
23	−1.46692	1.27649	0.81427	0.91967
24	3.15078	0.52310	0.30720	0.47231
25	−2.18938	−1.59810	−1.01572	−1.48315
26	1.91715	0.77914	0.87771	0.76496
27	−2.71428	−1.04745	−0.77536	−0.91397
28	−6.59852	−1.63217	−1.47709	−1.71347
29	0.65101	0.44328	0.60956	0.32833
30	−0.13772	−0.24972	0.07876	−0.17216
31	−2.43959	−0.33480	−0.39314	−0.37198
32	−3.27802	−0.20680	−0.13936	0.05626
33	−2.48058	−1.99375	−1.69587	−1.87838
34	1.88639	0.66475	0.79670	0.69865
35	5.06459	−0.27977	0.10817	−0.26450
36	1.96335	−0.43990	−0.66022	−0.58490
37	0.26274	−0.05334	−0.02396	−0.16795
38	−2.94593	−2.06699	−1.31832	−1.72082
39	−1.38628	−1.02560	−1.15858	−1.19420
40	−0.20797	0.45847	0.21555	0.31347
41	−1.07820	0.93979	0.63454	0.69907
42	−1.66386	−0.93238	−0.95216	−1.02725
43	0.58117	−0.35988	−0.30693	−0.46232
44	1.37447	−0.00518	0.35985	0.02485
45	−2.82687	−0.18892	−0.07959	0.01704
46	3.86363	0.87271	0.47644	0.57036
47	−2.64141	−2.06993	−1.82915	−2.16738
48	0.05387	0.32143	−0.25961	0.21632
49	0.50763	−1.42382	−0.77620	−1.07473
50	0.64347	−0.07852	−0.21347	−0.11750

Table 9.2 Last 20 Observations of Equal Educational Opportunity (EEO) Data; Standardized Indexes

Row	ACHV	FAM	PEER	SCHOOL
51	2.49414	−0.14925	−0.03192	−0.36598
52	0.61955	0.52666	0.79149	0.71369
53	0.61745	−1.49102	−1.02073	−1.38103
54	−1.00743	−0.94757	−1.28991	−1.24799
55	−0.37469	0.24550	0.83794	0.59596
56	−2.52824	−0.41630	−0.60312	−0.34951
57	0.02372	1.38143	1.54542	1.59429
58	2.51077	1.03806	0.91637	0.97602
59	−4.22716	−0.88639	−0.47652	−0.77693
60	1.96847	1.08655	0.65700	0.89401
61	1.25668	−1.95142	−1.94199	−1.89645
62	−0.16848	2.83384	2.47398	2.79222
63	−0.34158	1.86753	1.55229	1.80057
64	−2.23973	−1.11172	−0.69732	−0.80197
65	3.62654	1.41958	1.11481	1.24558
66	0.97034	0.53940	0.16182	0.33477
67	3.16093	0.22491	0.74800	0.66182
68	−1.90801	1.48244	1.47079	1.54283
69	0.64598	2.05425	1.80369	1.90066
70	−1.75915	1.24058	0.64484	0.87372

Table 9.3 EEO Data: Regression Results

Variable	Coefficient	s.e.	t-test	p-value
Constant	−0.070	0.251	−0.28	0.7810
FAM	1.101	1.411	0.78	0.4378
PEER	2.322	1.481	1.57	0.1218
SCHOOL	−2.281	2.220	−1.03	0.3080
$n = 70$	$R^2 = 0.206$	$R_a^2 = 0.170$	$\hat{\sigma} = 2.07$	$d.f. = 66$

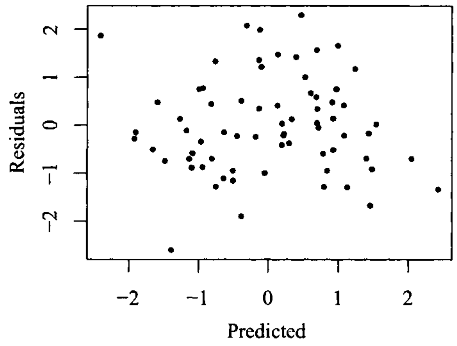

Figure 9.1 Standardized residuals against fitted values of ACHV.

results show almost no change. Therefore, the observation has been retained in the analysis.

From Table 9.3 we see that about 20% of the variation in achievement score is accounted for by the three predictors jointly ($R^2 = 0.206$). The F-value is 5.72 based on 3 and 66 degrees of freedom and is significant at better than the 0.01 level. Therefore, even though the total explained variation is estimated at only 20%, it is accepted that FAM, PEER, and SCHOOL are valid predictor variables. However, the individual t-values are all small. In total, the summary statistics say that the three predictors taken together are important but from the t-values, it follows that any one predictor may be deleted from the model provided the other two are retained.

These results are typical of a situation where extreme multicollinearity is present. The predictor variables are so highly correlated that each one may serve as a proxy for the others in the regression equation without affecting the total explanatory power. The low t-values confirm that any one of the predictor variables may be dropped from the equation. Hence the regression analysis has failed to provide any information for evaluating the importance of school inputs on achievement. The culprit is clearly multicollinearity. The pairwise correlation coefficients of the three predictor variables and the corresponding scatter plots (Figure 9.2), all show strong linear relationships among all pairs of predictor variables. All pairwise correlation coefficients are high. In all scatter plots, all the observations lie close to the straight line through the average values of the corresponding variables.

Multicollinearity in this instance could have been expected. It is the nature of these three variables that each is determined by and helps to determine the others. It is not unreasonable to conclude that there are not three variables but in fact only one. Unfortunately, that conclusion does not help to answer the original question about the effects of school facilities on achievement. There remain two possibilities. First, multicollinearity may be present because the sample data are deficient, but can be improved with additional observations. Second, multicollinearity may be present

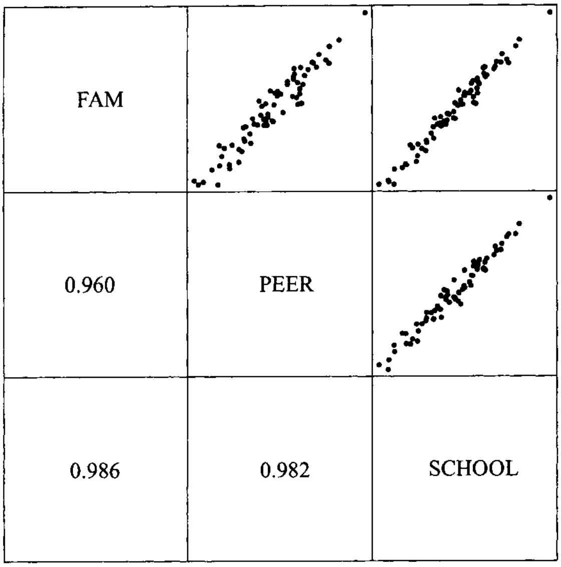

Figure 9.2 Pairwise scatter plots of the three predictor variables FAM, PEER, and SCHOOL; and the corresponding pairwise correlation coefficients.

because the interrelationships among the variables are an inherent characteristic of the process under investigation. Both situations are discussed in the following paragraphs.

In the first case the sample should have been selected to insure that the correlations between the predictor variables were not large. For example, in the scatter plot of FAM versus SCHOOL (the graph in the top right corner in Figure 9.2), there are no schools in the sample with values in the upper left or lower right regions of the graph. Hence there is no information in the sample on achievement when the value of FAM is high and SCHOOL is low, or FAM is low and SCHOOL is high. But it is only with data collected under these two conditions that the individual effects of FAM and SCHOOL on ACHV can be determined. For example, assume that there were some observations in the upper left quadrant of the graph. Then it would at least be possible to compare average ACHV for low and high values of SCHOOL when FAM is held constant.

Since there are three predictor variables in the model, then there are eight distinct combinations of data that should be included in the sample. Using + to represent a value above the average and – to represent a value below the average, the eight possibilities are represented in Table 9.4.

The large correlations that were found in the analysis suggest that only combinations 1 and 8 are represented in the data. If the sample turned out this way by chance, the prescription for resolving the multicollinearity problem is to collect additional data on some of the other combinations. For example, data based on

Table 9.4 Data Combinations for Three Predictor Variables

	Variable		
Combination	FAM	PEER	SCHOOL
1	+	+	+
2	+	+	−
3	+	−	+
4	−	+	+
5	+	−	−
6	−	+	−
7	−	−	+
8	−	−	−

combinations 1 and 2 alone could be used to evaluate the effect of SCHOOL on ACHV holding FAM and PEER at a constant level, both above average. If these were the only combinations represented in the data, the analysis would consist of the simple regression of ACHV against SCHOOL. The results would give only a partial answer, namely, an evaluation of the school-achievement relationship when FAM and PEER are both above average.

The prescription for additional data as a way to resolve multicollinearity is not a panacea. It is often not possible to collect more data because of constraints on budgets, time, and staff. It is always better to be aware of impending data deficiencies beforehand. Whenever possible, the data should be collected according to design. Unfortunately, prior design is not always feasible. In surveys, or observational studies such as the one being discussed, the values of the predictor variables are usually not known until the sampling unit is selected for the sample and some costly and time-consuming measurements are developed. Following this procedure, it is fairly difficult to ensure that a balanced sample will be obtained.

The second reason that multicollinearity may appear is because the relationships among the variables are an inherent characteristic of the process being sampled. If FAM, PEER, and SCHOOL exist in the population only as data combinations 1 and 8 of Table 9.4, it is not possible to estimate the individual effects of these variables on achievement. The only recourse for continued analysis of these effects would be to search for underlying causes that may explain the interrelationships of the predictor variables. Through this process, one may discover other variables that are more basic determinants affecting equal opportunity in education and achievement.

9.3 EFFECTS ON FORECASTING

We shall examine the effects of multicollinearity in forecasting when the forecasts are based on a multiple regression equation. A historical data set with observations indexed by time is used to estimate the regression coefficients. Forecasts of the

Table 9.5 Data on French Economy

YEAR	IMPORT	DOPROD	STOCK	CONSUM
49	15.9	149.3	4.2	108.1
50	16.4	161.2	4.1	114.8
51	19.0	171.5	3.1	123.2
52	19.1	175.5	3.1	126.9
53	18.8	180.8	1.1	132.1
54	20.4	190.7	2.2	137.7
55	22.7	202.1	2.1	146.0
56	26.5	212.4	5.6	154.1
57	28.1	226.1	5.0	162.3
58	27.6	231.9	5.1	164.3
59	26.3	239.0	0.7	167.6
60	31.1	258.0	5.6	176.8
61	33.3	269.8	3.9	186.6
62	37.0	288.4	3.1	199.7
63	43.3	304.5	4.6	213.9
64	49.0	323.4	7.0	223.8
65	50.3	336.8	1.2	232.0
66	56.6	353.9	4.5	242.9

Source: Malinvaud (1968).

response variable are produced by using future values of the predictor variables in the estimated regression equation. The future values of the predictor variables must be known or forecasted from other data and models. We shall not treat the uncertainty in the forecasted predictor variables. In our discussion it is assumed that the future values of the predictor variables are given.

We have chosen an example based on aggregate data concerning import activity in the French economy. The data have been analyzed by Malinvaud (1968). Our discussion follows his presentation. The variables are imports (IMPORT), domestic production (DOPROD), stock formation (STOCK), and domestic consumption (CONSUM), all measured in billions of French francs for the years 1949 through 1966. The data are given in Table 9.5 and can be obtained from the book's Web site. The model being considered is

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon. \quad (9.3)$$

The regression results appear as Table 9.6. The index plot of residuals (Figure 9.3) shows a distinctive pattern, suggesting that the model is not well specified. Even though multicollinearity appears to be present ($R^2 = 0.973$ and all t-values small), it should not be pursued further in this model. Multicollinearity should only be attacked after the model specification is satisfactory. The difficulty with the model is that the European Common Market began operations in 1960, causing changes in import-export relationships. Since our objective in this chapter is to study the effects of multicollinearity, we shall not complicate the model by attempting to

Table 9.6 Import data (1949–1966): Regression Results

ANOVA Table				
Source	Sum of Squares	*d.f.*	Mean Square	F-test
Regression	2576.92	3	858.974	168
Residuals	71.39	14	5.099	

Coefficients Table				
Variable	Coefficient	s.e.	t-test	p-value
Constant	−19.725	4.125	−4.78	0.0003
DOPROD	0.032	0.187	0.17	0.8656
STOCK	0.414	0.322	1.29	0.2195
CONSUM	0.243	0.285	0.85	0.4093
$n = 18$	$R^2 = 0.973$	$R_a^2 = 0.967$	$\hat{\sigma} = 2.258$	$d.f. = 14$

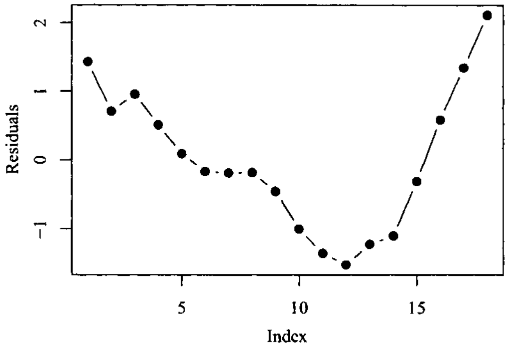

Figure 9.3 Import data (1949–1966): Index plot of the standardized residuals.

capture the behavior after 1959. We shall assume that it is now 1960 and look only at the 11 years 1949–1959. The regression results for those data are summarized in Table 9.7. The residual plot is now satisfactory (Figure 9.4).

The value of $R^2 = 0.99$ is high. However, the coefficient of DOPROD is negative and not statistically significant, which is contrary to prior expectation. We believe that if STOCK and CONSUM were held constant, an increase in DOPROD would cause an increase in IMPORT, probably for raw materials or manufacturing equipment. Multicollinearity is a possibility here and in fact is the case. The simple correlation between CONSUM and DOPROD is 0.997. Upon further investigation it turns out that CONSUM has been about two-thirds of DOPROD throughout the

Table 9.7 Import data (1949–1959): Regression Results

ANOVA Table

Source	Sum of Squares	*d.f.*	Mean Square	F-test
Regression	204.776	3	68.2587	286
Residuals	1.673	7	0.2390	

Coefficients Table

Variable	Coefficient	s.e.	t-test	p-value
Constant	−10.128	1.212	−8.36	< 0.0001
DOPROD	−0.051	0.070	−0.73	0.4883
STOCK	0.587	0.095	6.20	0.0004
CONSUM	0.287	0.102	2.81	0.0263
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.4889$	$d.f. = 7$

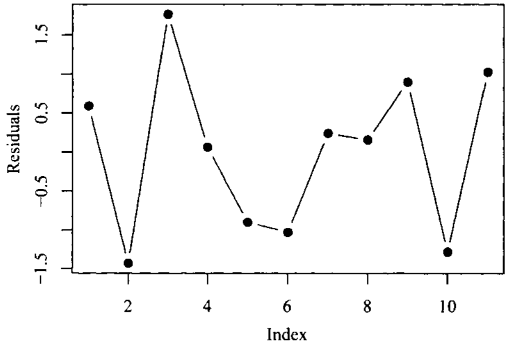

Figure 9.4 Import data (1949–1959): Index plot of the standardized residuals.

11-year period. The estimated relationship between the two quantities is

$$\text{CONSUM} = 6.259 + 0.686 \cdot \text{DOPROD}.$$

Even in the presence of such severe multicollinearity the regression equation may produce some good forecasts. From Table 9.7, the forecasting equation is

$$\begin{aligned}\text{IMPORT} &= -10.13 - 0.051 \cdot \text{DOPROD} + 0.587 \cdot \text{STOCK} \\ &\quad + 0.287 \cdot \text{CONSUM}.\end{aligned}$$

Recall that the fit to the historical data is very good and the residual variation appears to be purely random. To forecast we must be confident that the character and strength of the overall relationship will hold into future periods. This matter of confidence is a problem in all forecasting models whether or not multicollinearity is present. For the purpose of this example we assume that the overall relationship does hold into future periods.[2] Implicit in this assumption is the relationship between DOPROD and CONSUM. The forecast will be accurate as long as the future values of DOPROD, STOCK, and CONSUM have the relationship that CONSUM is approximately equal to $0.7\times$ DOPROD.

For example, let us forecast the change in IMPORT next year corresponding to an increase in DOPROD of 10 units while holding STOCK and CONSUM at their current levels. The resulting forecast is

$$\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.051(10),$$

which means that IMPORT will decrease by -0.51 units. However, if the relationship between DOPROD and CONSUM is kept intact, CONSUM will increase by $10(2/3) = 6.67$ units and the forecasted result is

$$\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.51 + 0.287 \times 6.67 = \text{IMPORT}_{1959} + 1.5.$$

IMPORT actually increases by 1.5 units, a more satisfying result and probably a better forecast. The case where DOPROD increases alone corresponds to a change in the basic structure of the data that were used to estimate the model parameters and cannot be expected to produce meaningful forecasts.

In summary, the two examples demonstrate that multicollinear data can seriously limit the use of regression analysis for inference and forecasting. Extreme care is required when attempting to interpret regression results when multicollinearity is suspected. In Section 9.4 we discuss methods for detecting extreme collinearity among predictor variables.

[2] For the purpose of convenient exposition we ignore the difficulties that arise because of our previous finding that the formation of the European Common Market has altered the relationship since 1960. But we are impelled to advise the reader that changes in structure make forecasting a very delicate endeavor even when the historical fit is excellent.

9.4 DETECTION OF MULTICOLLINEARITY

In the preceding examples some of the ideas for detecting multicollinearity were already introduced. In this section we review those ideas and introduce additional criteria that indicate collinearity. Multicollinearity is associated with unstable estimated regression coefficients. This situation results from the presence of strong linear relationships among the predictor variables. It is not a problem of misspecification. Therefore, the empirical investigation of problems that result from a collinear data set should begin only after the model has been satisfactorily specified. However, there may be some indications of multicollinearity that are encountered during the process of adding, deleting, and transforming variables or data points in search of the good model. Indication of multicollinearity that appear as instability in the estimated coefficients are as follows:

- Large changes in the estimated coefficients when a variable is added or deleted.
- Large changes in the coefficients when a data point is altered or dropped.

Once the residual plots indicate that the model has been satisfactorily specified, multicollinearity may be present if:

- The algebraic signs of the estimated coefficients do not conform to prior expectations; or
- Coefficients of variables that are expected to be important have large standard errors (small t-values).

For the IMPORT data discussed previously, the coefficient of DOPROD was negative and not significant. Both results are contrary to prior expectations. The effects of dropping or adding a variable can be seen in Table 9.8. There we see that the presence or absence of certain variables has a large effect on the other coefficients. For the EEO data (Tables 9.1 and 9.2) the algebraic signs are all correct, but their standard errors are so large that none of the coefficients are statistically significant. It was expected that they would all be important.

The presence of multicollinearity is also indicated by the size of the correlation coefficients that exist among the predictor variables. A large correlation between a pair of predictor variables indicates a strong linear relationship between those two variables. The correlations for the EEO data (Figure 9.2) are large for all pairs of predictor variables. For the IMPORT data, the correlation coefficient between DOPROD and CONSUM is 0.997.

The source of multicollinearity may be more subtle than a simple relationship between two variables. A linear relation can involve many of the predictor variables. It may not be possible to detect such a relationship with a simple correlation coefficient. As an example, we shall look at an analysis of the effects of advertising expenditures (A_t), promotion expenditures (P_t), and sales expense (E_t) on the

Table 9.8 Import Data (1949–1959): Regression Coefficients for All Possible Regressions

	Variable			
Regression	Constant	DOPROD	STOCK	CONSUM
1	−6.558	0.146	–	–
2	19.611	–	0.691	–
3	−8.013	–	–	0.214
4	−8.440	0.145	0.622	–
5	−8.884	−0.109	–	0.372
6	−9.743	–	0.596	0.212
7	−10.128	−0.051	0.587	0.287

aggregate sales of a firm in period t. The data represent a period of 23 years during which the firm was operating under fairly stable conditions. The data are given in Table 9.9 and can be obtained from the book's Web site.

The proposed regression model is

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t, \qquad (9.4)$$

where A_{t-1} and P_{t-1} are the lagged one-year variables. The regression results are given in Table 9.10. The plot of residuals versus fitted values and the index plot of residuals (Figures 9.5 and 9.6), as well as other plots of the residuals versus the predictor variables (not shown), do not suggest any problems of misspecification. Furthermore, the correlation coefficients between the predictor variables are small (Table 9.11). However, if we do a little experimentation to check the stability of the coefficients by dropping the contemporaneous advertising variable A from the model, many things change. The coefficient of P_t drops from 8.37 to 3.70; the coefficients of lagged advertising A_{t-1} and lagged promotions P_{t-1} change signs. But the coefficient of sales expense is stable and R^2 does not change much.

The evidence suggests that there is some type of relationship involving the contemporaneous and lagged values of the advertising and promotions variables. The regression of A_t on P_t, A_{t-1}, and P_{t-1} returns an R^2 of 0.973. The equation takes the form

$$\hat{A}_t = 4.63 - 0.87P_t - 0.86A_{t-1} - 0.95P_{t-1}.$$

Upon further investigation into the operations of the firm, it was discovered that close control was exercised over the expense budget during those 23 years of stability. In particular, there was an approximate rule imposed on the budget that the sum of A_t, A_{t-1}, P_t, and P_{t-1} was to be held to approximately five units over every two-year period. The relationship

$$A_t + P_t + A_{t-1} + P_{t-1} \doteq 5$$

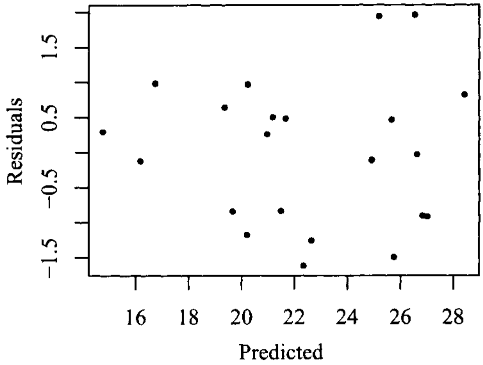

Figure 9.5 Standardized residuals versus fitted values of Sales.

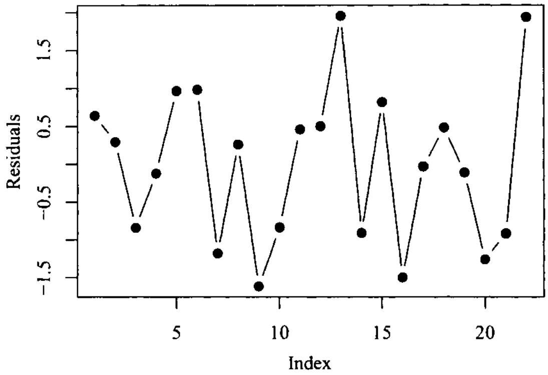

Figure 9.6 Index plot of the standardized residuals.

Table 9.9 Annual Data on Advertising, Promotions, Sales Expenses, and Sales (Millions of Dollars)

Row	S_t	A_t	P_t	E_t	A_{t-1}	P_{t-1}
1	20.11371	1.98786	1.0	0.30	2.01722	0.0
2	15.10439	1.94418	0.0	0.30	1.98786	1.0
3	18.68375	2.19954	0.8	0.35	1.94418	0.0
4	16.05173	2.00107	0.0	0.35	2.19954	0.8
5	21.30101	1.69292	1.3	0.30	2.00107	0.0
6	17.85004	1.74334	0.3	0.32	1.69292	1.3
7	18.87558	2.06907	1.0	0.31	1.74334	0.3
8	21.26599	1.01709	1.0	0.41	2.06907	1.0
9	20.48473	2.01906	0.9	0.45	1.01709	1.0
10	20.54032	1.06139	1.0	0.45	2.01906	0.9
11	26.18441	1.45999	1.5	0.50	1.06139	1.0
12	21.71606	1.87511	0.0	0.60	1.45999	1.5
13	28.69595	2.27109	0.8	0.65	1.87511	0.0
14	25.83720	1.11191	1.0	0.65	2.27109	0.8
15	29.31987	1.77407	1.2	0.65	1.11191	1.0
16	24.19041	0.95878	1.0	0.65	1.77407	1.2
17	26.58966	1.98930	1.0	0.62	0.95878	1.0
18	22.24466	1.97111	0.0	0.60	1.98930	1.0
19	24.79944	2.26603	0.7	0.60	1.97111	0.0
20	21.19105	1.98346	0.1	0.61	2.26603	0.7
21	26.03441	2.10054	1.0	0.60	1.98346	0.1
22	27.39304	1.06815	1.0	0.58	2.10054	1.0

is the cause of the multicollinearity.

A thorough investigation of multicollinearity will involve examining the value of R^2 that results from regressing each of the predictor variables against all the others. The relationship between the predictor variables can be judged by examining a quantity called the *variance inflation factor* (VIF). Let R_j^2 be the square of the multiple correlation coefficient that results when the predictor variable X_j is regressed against all the other predictor variables. Then the variance inflation for X_j is

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p, \tag{9.5}$$

where p is the number of predictor variables. It is clear that if X_j has a strong linear relationship with the other predictor variables, R_j^2 would be close to 1, and VIF_j would be large. Values of variance inflation factors greater than 10 is often taken as a signal that the data have collinearity problems.

In absence of any linear relationship between the predictor variables (i.e., if the predictor variables are orthogonal), R_j^2 would be zero and VIF_j would be one. The deviation of VIF_j value from 1 indicates departure from orthogonality and tendency toward collinearity. The value of VIF_j also measures the amount by

Table 9.10 Regression Results for the Advertising Data

ANOVA Table				
Source	Sum of Squares	$d.f.$	Mean Square	F-test
Regression	307.572	5	61.514	35.3
Residuals	27.879	16	1.742	
Coefficients Table				
Variable	Coefficient	s.e.	t-test	p-value
Constant	−14.194	18.715	−0.76	0.4592
A	5.361	4.028	1.33	0.2019
P	8.372	3.586	2.33	0.0329
E	22.521	2.142	10.51	< 0.0001
A_{t-1}	3.855	3.578	1.08	0.2973
P_{t-1}	4.125	3.895	1.06	0.3053
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 1.320$	$d.f. = 16$

Table 9.11 Pairwise Correlation Coefficients for the Advertising Data

	A_t	P_t	E_t	A_{t-1}	P_{t-1}
A_t	1.000				
P_t	−0.357	1.000			
E_t	−0.129	0.063	1.000		
A_{t-1}	−0.140	−0.316	−0.166	1.000	
P_{t-1}	−0.496	−0.296	0.208	−0.358	1.000

which the variance of the jth regression coefficient is increased due to the linear association of X_j with other predictor variables relative to the variance that would result if X_j were not related to them linearly. This explains the naming of this particular diagnostic.

As R_j^2 tends toward 1, indicating the presence of a linear relationship in the predictor variables, the VIF for $\hat{\beta}_j$ tends to infinity. It is suggested that a VIF in excess of 10 is an indication that multicollinearity may be causing problems in estimation.

The precision of an ordinary least squares (OLS) estimated regression coefficient is measured by its variance, which is proportional to σ^2, the variance of the error term in the regression model. The constant of proportionality is the VIF. Thus, the VIFs may be used to obtain an expression for the expected squared distance of the OLS estimators from their true values. Denoting the square of the distance by D^2, it can be shown that, on average,

$$D^2 = \sigma^2 \sum_{j=1}^{p} \text{VIF}_j.$$

This distance is another measure of precision of the least squares estimators. The smaller the distance, the more accurate are the estimates. If the predictor variables were orthogonal, the VIFs would all be 1 and D^2 would be $p\sigma^2$. It follows that the ratio

$$\frac{\sigma^2 \sum_{i=1}^{p} \text{VIF}_i}{p\sigma^2} = \frac{\sum_{i=1}^{p} \text{VIF}_i}{p} = \overline{\text{VIF}},$$

which shows that the average of the VIFs measures the squared error in the OLS estimators relative to the size of that error if the data were orthogonal. Hence, $\overline{\text{VIF}}$ may also be used as an index of multicollinearity.

Most computer packages now furnish values of VIF_j routinely. Some have built-in messages when high values of VIF_j are observed. In any regression analysis the values of VIF_j should always be examined to avoid the pitfalls resulting from fitting a regression model to collinear data by least squares.

In each of the three examples (EEO, Import, and Advertising) we have seen evidence of collinearity. The VIF_j's and their average values for these data sets are given in Table 9.12. For the EEO data the values of VIF_j range from 30.2 to 83.2, showing that all three variables are strongly intercorrelated and that dropping one of the variables will not eliminate collinearity. The average value of VIF of 50.3 indicates that the squared error in the OLS estimators is 50 times as large as it would be if the predictor variables were orthogonal.

For the Import data, the squared error in the OLS estimators is 313 times as large as it would be if the predictor variables were orthogonal. However, the VIF_j's indicate that domestic production and consumption are strongly correlated but are not correlated with the STOCK variable. A regression equation containing either CONSUM or DOPROD along with STOCK will eliminate collinearity.

Table 9.12 Variance Inflation Factors for Three Data Sets

EEO		Import		Advertising	
Variable	VIF	Variable	VIF	Variable	VIF
FAM	37.6	DOPROD	469.7	A_t	37.4
PEER	30.2	STOCK	1.0	P_t	33.5
SCHOOL	83.2	CONSUM	469.4	E_t	1.1
				A_{t-1}	26.6
				P_{t-1}	44.1
Average	50.3	Average	313.4	Average	28.5

For the Advertising data, VIF_E (for the variable E) is 1.1, indicating that this variable is not correlated with the remaining predictor variables. The VIF_j's for the other four variables are large, ranging from 26.6 to 44.1. This indicates that there is a strong linear relationship among the four variables, a fact that we have already noted. Here the prescription might be to regress sales S_t against E_t and three of the remaining four variables $(A_t, P_t, A_{t-1}, S_{t-1})$ and examine the resulting VIF_j's to see if collinearity has been eliminated.

9.5 CENTERING AND SCALING

The indicators of multicollinearity that have been described so far can all be obtained using standard regression computations. There is another, more unified way to analyze multicollinearity which requires some calculations that are not usually included in standard regression packages. The analysis follows from the fact that every linear regression model can be restated in terms of a set of orthogonal predictor variables. These new variables are obtained as linear combinations of the original predictor variables. They are referred to as the *principal components* of the set of predictor variables (Seber, 1984; Johnson and Wichern, 1992).

To develop the method of principal components, we may first need to *center* and/or *scale* the variables. We have been mainly dealing with regression models of the form

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \tag{9.6}$$

which are models with a constant term β_0. But we have also seen situations where fitting the *no-intercept* model

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \tag{9.7}$$

is necessary (see, e.g., Chapters 3 and 7). When dealing with constant term models, it is convenient to center and scale the variables, but when dealing with a no-intercept model, we need only to scale the variables.

9.5.1 Centering and Scaling in Intercept Models

If we are fitting an intercept model as in (9.6), we need to center and scale the variables. A *centered* variable is obtained by subtracting from each observation the mean of all observations. For example, the centered response variable is $(Y - \bar{y})$ and the centered jth predictor variable is $(X_j - \bar{x}_j)$. The mean of a centered variable is zero. The centered variables can also be scaled. Two types of scaling are usually needed: *unit length scaling* and *standardizing*. Unit length scaling of the response variable Y and the jth predictor variable X_j is obtained as follows:

$$\begin{aligned} \tilde{Z}_y &= \frac{Y-\bar{y}}{L_y}, \\ \tilde{Z}_j &= \frac{X_j - \bar{x}_j}{L_j}, \quad j = 1, \ldots, p, \end{aligned} \tag{9.8}$$

where $\bar{y}$ is the mean of Y, $\bar{x}_j$ is the mean of X_j, and

$$L_y = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \text{ and } L_j = \sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}, \; j = 1, \ldots, p. \tag{9.9}$$

The quantities L_y is referred to as the *length* of the centered variable $Y - \bar{y}$ because it measures the size or the magnitudes of the observations in $Y - \bar{y}$. Similarly, L_j measure the length of the variable $X_j - \bar{x}_j$. The variables $\tilde{Z}_y$ and $\tilde{Z}_j$ in (9.8) have zero means and unit lengths, hence this type of scaling is called unit length scaling. In addition, unit length scaling has the following property:

$$\text{Cor}(X_j, X_k) = \sum_{i=1}^{n} z_{ij} z_{ik}. \tag{9.10}$$

That is, the correlation coefficient between the original variables, X_j and X_k, can be computed easily as the sum of the products of the scaled versions Z_j and Z_k.

The second type of scaling is called standardizing, which is defined by

$$\begin{aligned} \tilde{Y} &= \frac{Y-\bar{y}}{s_y}, \\ \tilde{X}_j &= \frac{X_j - \bar{x}_j}{s_j}, \; j = 1, \ldots, p, \end{aligned} \tag{9.11}$$

where

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}, \text{ and } s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}}, \; j = 1, \ldots, p, \tag{9.12}$$

are standard deviations of the response and jth predictor variable, respectively. The standardized variables $\tilde{Y}$ and $\tilde{X}_j$ in (9.11) have means zero and unit standard deviations.

Since correlations are unaffected by shifting or scaling the data, it is both sufficient and convenient to deal with either the unit length scaled or the standardized versions of the variables. The variances and covariances of a set of p variables, $X_1, \ldots, X_p$, can be neatly displayed as a squared array of numbers called a *matrix.* This matrix is known as the *variance-covariance matrix.* The elements on the diagonal that runs from the upper-left corner to the lower-right corner of the matrix are known as the *diagonal elements.* The elements on the diagonal of a variance-covariance matrix are the variances and the elements off the diagonal are the covariances.[3] The variance-covariance matrix of the three predictor variables in the Import data for the years 1949–1959 is:

$$\begin{array}{l} \\ \text{DOPROD} \\ \text{STOCK} \\ \text{CONSUM} \end{array} \begin{array}{c} \begin{array}{ccc} \text{DOPROD} & \text{STOCK} & \text{CONSUM} \end{array} \\ \begin{pmatrix} 899.971 & 1.279 & 617.326 \\ 1.279 & 2.720 & 1.214 \\ 617.326 & 1.214 & 425.779 \end{pmatrix} \end{array}.$$

Thus, for example, Var(DOPROD) = 899.971, which is in the first diagonal element, and Cov(DOPROD, CONSUM) = 617.326, which is the value is the intersection of the first row and third column (or the third row and first column).

Similarly, the pairwise correlation coefficients can be displayed in matrix known as the *correlation matrix.* The correlation matrix of the three predictor variables in the Import data is:

$$\begin{array}{l} \\ \text{DOPROD} \\ \text{STOCK} \\ \text{CONSUM} \end{array} \begin{array}{c} \begin{array}{ccc} \text{DOPROD} & \text{STOCK} & \text{CONSUM} \end{array} \\ \begin{pmatrix} 1.000 & 0.026 & 0.997 \\ 0.026 & 1.000 & 0.036 \\ 0.997 & 0.036 & 1.000 \end{pmatrix} \end{array}. \qquad (9.13)$$

This is the same as the variance-covariance matrix of the standardized predictor variables. Thus, for example, Cor(DOPROD, CONSUM) = 0.997, which indicates that the two variables are highly correlated. Note that all the diagonal elements of the correlation matrix are equal to one.

Recall that a set of variables is said to be orthogonal if there exists no linear relationships among them. If the standardized predictor variables are orthogonal, their matrix of variances and covariances consists of one for the diagonal elements and zero for the off-diagonal elements.

9.5.2 Scaling in No-Intercept Models

If we are fitting a no-intercept model as in (9.7), we do not center the data because centering has the effect of including a constant term in the model. This can be seen from:

$$Y - \bar{y} = \beta_1(X_1 - \bar{x}_1) + \ldots + \beta_p(X_p - \bar{x}_p) + \varepsilon. \qquad (9.14)$$

[3] Readers not familiar with matrix algebra may benefit from reading the book, *Matrix Algebra As a Tool*, by Hadi (1996).

rearranging terms, we obtain

$$\begin{aligned} Y &= \bar{y} - (\beta_1\bar{x}_1 + \ldots + \beta_p\bar{x}_p) + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \end{aligned} \tag{9.15}$$

where $\beta_0 = \bar{y} - (\beta_1\bar{x}_1 + \ldots + \beta_p\bar{x}_p)$. Although a constant term does not appear in an explicit form in (9.14), it is clearly seen in (9.15). Thus, when we deal with no-intercept models, we need only to scale the data. The scaled variables are defined by:

$$\begin{aligned} \tilde{Z}_y &= \frac{Y}{L_y}, \\ \tilde{Z}_j &= \frac{X_j}{L_j}, \ j = 1, \ldots, p, \end{aligned} \tag{9.16}$$

where

$$L_y = \sqrt{\sum_{i=1}^{n} y_i^2}, \text{ and } L_j = \sqrt{\sum_{i=1}^{n} x_{ij}^2}, \ j = 1, \ldots, p. \tag{9.17}$$

The scaled variables in (9.16) have unit lengths but do not necessarily have means zero. Nor do they satisfy (9.10) unless the original variables have zero means.

We should mention here that centering (when appropriate) and/or scaling can be done without loss of generality because the regression coefficients of the original variables can be recovered from the regression coefficients of the transformed variables. For example, if we fit a regression model to centered data, the obtained regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are the same as the estimates obtained from fitting the model to the original data. The estimate of the constant term when using the centered data will always be zero. The estimate of the constant term for an intercept model can be obtained from:

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_p\bar{x}_p).$$

Scaling, however, will change the values of the estimated regression coefficients. For example, the relationship between the estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_p$, obtained from using the original data and the those obtained using the standardized data is given by

$$\begin{aligned} \hat{\beta}_j &= (s_y/s_j)\hat{\theta}_j, \qquad j = 1, 2, 3, 4, p, \\ \hat{\beta}_0 &= \bar{y} - \sum_{j=1}^{5} \hat{\beta}_j\bar{x}_j, \end{aligned} \tag{9.18}$$

where $\hat{\beta}_j$ and $\hat{\theta}_j$ are the jth estimated regression coefficients obtained when using the original and standardized data, respectively. Similar formulas can be obtained when using unit length scaling instead of standardizing.

We shall make extensive use of the centered and/or scaled variables in the rest of this chapter and in Chapter 10.

9.6 PRINCIPAL COMPONENTS APPROACH

As we mentioned in the previous section, the principal components approach to the detection of multicollinearity is based on the fact that any set of p variables can be transformed to a set of p orthogonal variables. The new orthogonal variables are known as the *principal components* (PCs) and are denoted by $C_1, \ldots, C_p$. Each variable C_j is a linear combination of the variables $\tilde{X}_1, \ldots, \tilde{X}_p$ in (9.11). That is,

$$C_j = v_{1j}\tilde{X}_1 + v_{2j}\tilde{X}_2 + \ldots + v_{pj}\tilde{X}_p, \quad j = 1, 2, \ldots, p. \tag{9.19}$$

The linear combinations are chosen so that the variables $C_1, \ldots, C_p$ are orthogonal.[4] The variance-covariance matrix of the PCs is of the form:

$$\begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_p \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_p \end{array} \\ \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \end{array}.$$

All the off-diagonal elements are zero because the PCs are orthogonal. The value on the jth diagonal element, λ_j is the variance of C_j, the jth PC. The PCs are arranged so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, that is, the first PC has the largest variance and the last PC has the smallest variance. The λ's are called *eigenvalues* of the correlation matrix of the predictor variables $X_1, \ldots, X_p$. The coefficients involved in the creation of C_j in (9.19) can be neatly arranged in a column like

$$\begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{pj} \end{pmatrix},$$

which is known as the *eigenvector* associated with the jth eigenvalue λ_j. If any one of the λ's is exactly equal to zero, there is a perfect linear relationship among the original variables, which is an extreme case of multicollinearity. If one of the λ's is much smaller than the others (and near zero), multicollinearity is present. The number of near zero λ's is equal to the number of different sets of multicollinearity that exist in the data. So, if there is only one near zero λ, there is only one set of multicollinearity; if there are two near zero λ's, there are two sets of different multicollinearity; and so on.

The eigenvalues of the correlation matrix in (9.13) are $\lambda_1 = 1.999$, $\lambda_2 = 0.998$, and $\lambda_3 = 0.003$. The corresponding eigenvectors are

$$\begin{pmatrix} 0.706 \\ 0.044 \\ 0.707 \end{pmatrix}, \begin{pmatrix} -0.036 \\ 0.999 \\ -0.026 \end{pmatrix}, \begin{pmatrix} -0.707 \\ -0.007 \\ 0.707 \end{pmatrix}.$$

[4] A description of this technique employing matrix algebra is given in the Appendix to this chapter.

Table 9.13 The PCs for the Import Data (1949–1959)

Year	C_1	C_2	C_3
49	−2.1258	0.6394	−0.0204
50	−1.6189	0.5561	−0.0709
51	−1.1153	−0.0726	−0.0216
52	−0.8944	−0.0821	0.0110
53	−0.6449	−1.3064	0.0727
54	−0.1907	−0.6591	0.0266
55	0.3593	−0.7438	0.0427
56	0.9726	1.3537	0.0627
57	1.5600	0.9635	0.0233
58	1.7677	1.0146	−0.0453
59	1.9304	−1.6633	−0.0809

Thus, the PCs for the Import data for the years 1949–1959 are:

$$\begin{aligned} C_1 &= 0.706\,\tilde{X}_1 + 0.044\,\tilde{X}_2 + 0.707\,\tilde{X}_3, \\ C_2 &= -0.036\,\tilde{X}_1 + 0.999\,\tilde{X}_2 - 0.026\,\tilde{X}_3, \\ C_3 &= -0.707\,\tilde{X}_1 - 0.007\,\tilde{X}_2 + 0.707\,\tilde{X}_3. \end{aligned} \tag{9.20}$$

These PCs are given in Table 9.13. The variance-covariance matrix of the new variables is

$$\begin{array}{cc} & \begin{array}{ccc} C_1 & C_2 & C_3 \end{array} \\ \begin{array}{c} C_1 \\ C_2 \\ C_3 \end{array} & \begin{pmatrix} 1.999 & 0 & 0 \\ 0 & 0.998 & 0 \\ 0 & 0 & 0.003 \end{pmatrix} \end{array}.$$

The PCs lack simple interpretation since each is, in a sense, a mixture of the original variables. However, these new variables provide a unified approach for obtaining information about multicollinearity and serve as the basis of one of the alternative estimation techniques described in Chapter 10.

For the Import data, the small value of $\lambda_3 = 0.003$ points to multicollinearity. The other data sets considered in this chapter also have informative eigenvalues. For the EEO data, $\lambda_1 = 2.952$, $\lambda_2 = 0.040$, and $\lambda_3 = 0.008$. For the advertising data, $\lambda_1 = 1.701$, $\lambda_2 = 1.288$, $\lambda_3 = 1.145$, $\lambda_4 = 0.859$, and $\lambda_5 = 0.007$. In each case the presence of a small eigenvalue is indicative of multicollinearity.

A measure of the overall multicollinearity of the variables can be obtained by computing the *condition number* of the correlation matrix. The condition number is defined by

$$\kappa = \sqrt{\frac{\text{maximum eigenvalue of the correlation matrix}}{\text{minimum eigenvalue of the correlation matrix}}} = \sqrt{\frac{\lambda_1}{\lambda_p}}\,.$$

The condition number will always be greater than 1. A large condition number indicates evidence of strong collinearity. The harmful effects of collinearity in the data become strong when the values of the condition number exceeds 15 (which means that λ_1 is more than 225 times λ_p). The condition numbers for the three data sets EEO, Import, and advertising data are 19.20, 25.81, and 15.59, respectively. The cutoff value of 15 is not based on any theoretical considerations, but arises from empirical observation. Corrective action should always be taken when the condition number of the correlation matrix exceeds 30.

Another empirical criterion for the presence of multicollinearity is given by the sum of the reciprocals of the eigenvalues, that is,

$$\sum_{j=1}^{p} \frac{1}{\lambda_j} . \tag{9.21}$$

If this sum is greater than, say, five times the number of predictor variables, multicollinearity is present.

One additional piece of information is available through this type of analysis. Since λ_j is the variance of the jth PC, if λ_j is approximately zero, the corresponding PC, C_j, is approximately equal to a constant. It follows that the equation defining the PC gives some idea about the type of relationship among the predictor variables that is causing multicollinearity. For example, in the Import data, $\lambda_3 = 0.003 \doteq 0$. Therefore, C_3 is approximately constant. The constant is the mean value of C_3 which is zero. The PCs all have means of zero since they are linear functions of the standardized variables and each standardized variable has a zero mean. Therefore

$$C_3 = -0.707\,\tilde{X}_1 - 0.007\,\tilde{X}_2 + 0.707\,\tilde{X}_3 \doteq 0.$$

Rearranging the terms yields

$$\tilde{X}_1 \doteq \tilde{X}_3, \tag{9.22}$$

where the coefficient of $\tilde{X}_2$ (−0.007) has been approximated as zero. Equation (9.22) represents the approximate relationship that exists between the standardized versions of CONSUM and DOPROD. This result is consistent with our previous finding based on the high simple correlation coefficient (r = 0.997) between the predictor variables CONSUM and DOPROD. (The reader can confirm this high value of r by examining the scatter plot of CONSUM versus DOPROD.) Since λ_3 is the only small eigenvalue, the analysis of the PCs tells us that the dependence structure among the predictor variables as reflected in the data is no more complex than the simple relationship between CONSUM and DOPROD as given in Equation (9.22).

For the advertising data, the smallest eigenvalue is $\lambda_5 = 0.007$. The corresponding PC is

$$C_5 = 0.514\,\tilde{X}_1 + 0.489\,\tilde{X}_2 - 0.010\,\tilde{X}_3 + 0.428\,\tilde{X}_4 + 0.559\,\tilde{X}_5. \tag{9.23}$$

Setting C_5 to zero and solving for $\tilde{X}_1$ leads to the approximate relationship,

$$\tilde{X}_1 \doteq -0.951\,\tilde{X}_2 - 0.833\,\tilde{X}_4 - 1.087\,\tilde{X}_5, \tag{9.24}$$

where we have taken the coefficient of $\tilde{X}_3$ to be approximately zero. This equation reflects our earlier findings about the relationship between A_t, P_t, A_{t-1}, and P_{t-1}. Furthermore, since $\lambda_4 = 0.859$ and the other λ's are all large, we can be confident that the relationship involving A_t, P_t, A_{t-1}, and P_{t-1} in (9.24) is the only source of multicollinearity in the data.

Throughout this section, investigations concerning the presence of multicollinearity have been based on judging the magnitudes of various indicators, either a correlation coefficient or an eigenvalue. Although we speak in terms of large and small, there is no way to determine these threshold values. The size is relative and is used to give an indication either that everything seems to be in order or that something is amiss. The only reasonable criterion for judging size is to decide whether the ambiguity resulting from the perceived multicollinearity is of material importance in the underlying problem.

We should also caution here that the data analyzed may contain one or few observations that can have an undue influence on the various measures of collinearity (e.g., correlation coefficients, eigenvalues, or the condition number). These observations are called *collinearity-influential observations*. For more details the reader is referred to Hadi (1988).

9.7 IMPOSING CONSTRAINTS

We have noted that multicollinearity is a condition associated with deficient data and not due to misspecification of the model. It is assumed that the form of the model has been carefully structured and that the residuals are acceptable before questions of multicollinearity are considered. Since it is usually not practical and often impossible to improve the data, we shall focus our attention on methods of better interpretation of the given data than would be available from a direct application of least squares. In this section, rather than trying to interpret individual regression coefficients, we shall attempt to identify and estimate informative linear functions of the regression coefficients. Alternative estimating methods for the individual coefficients are treated in Chapter 10.

Before turning to the problem of searching the data for informative linear functions of the regression coefficients, one additional point concerning model specification must be discussed. A subtle step in specifying a relationship that can have a bearing on multicollinearity is acknowledging the presence of theoretical relationships among the regression coefficients. For example in the model for the Import data,

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon, \quad (9.25)$$

one may argue that the marginal effects of DOPROD and CONSUM are equal. That is, on the basis of economic reasoning, and before looking at the data, it is decided the $\beta_1 = \beta_3$ or equivalently, $\beta_1 - \beta_3 = 0$. As described in Section 3.9.3,

Table 9.14 Regression Results of Import Data (1949–1959) with the Constraint $\beta_1 = \beta_3$

Variable	Coefficient	s.e.	t-test	p-value
Constant	−9.007	1.245	−7.23	< 0.0001
STOCK	0.612	0.109	5.60	0.0005
NEWVAR	0.086	0.004	24.30	< 0.0001
$n = 11$	$R^2 = 0.987$	$R_a^2 = 0.984$	$\hat{\sigma} = 0.5693$	$d.f. = 8$

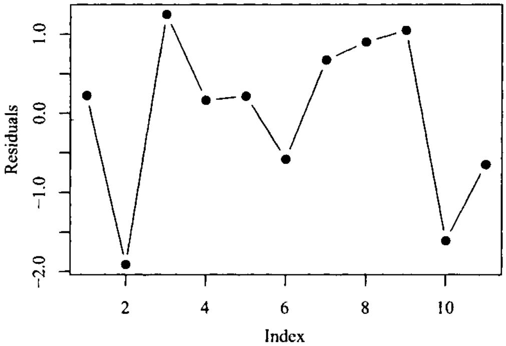

Figure 9.7 Index plot of the standardized residuals. Import data (1949–1959) with the constraint $\beta_1 = \beta_3$.

the model in (9.25) becomes

$$\begin{aligned}\text{IMPORT} &= \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_1 \cdot \text{CONSUM} + \varepsilon, \\ &= \beta_0 + \beta_2 \cdot \text{STOCK} + \beta_1\ (\text{DOPROD} + \text{CONSUM}) + \varepsilon.\end{aligned}$$

Thus, the common value of β_1 and β_3 is estimated by regressing IMPORT on STOCK and a new variable constructed as NEWVAR = DOPROD + CONSUM. The new variable has significance only as a technical manipulation to extract an estimate of the common value of β_1 and β_3. The results of the regression appear in Table 9.14. The correlation between the two predictor variables, STOCK and NEWVAR, is 0.0299 and the eigenvalues are $\lambda_1 = 1.030$ and $\lambda_2 = 0.970$. There is no longer any indication of multicollinearity. The residual plots against time and the fitted values indicate that there are no other problems of specification (Figures 9.7 and 9.8, respectively). The estimated model is

$$\begin{aligned}\text{IMPORT} = \ & -9.007 + 0.086 \cdot \text{DOPROD} + 0.612 \cdot \text{STOCK} \\ & + 0.086 \cdot \text{CONSUM}.\end{aligned}$$

Note that following the methods outlined in Section 3.9.3, it is also possible to test the constraint, $\beta_1 = \beta_3$, as a hypothesis. Even though the argument for $\beta_1 = \beta_3$

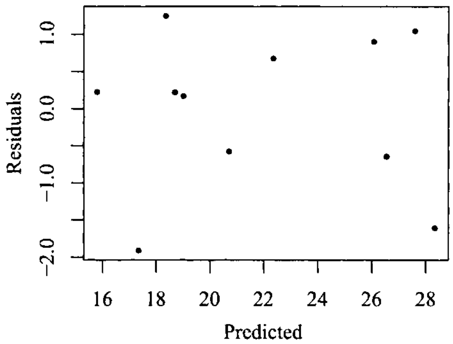

Figure 9.8 Standardized residuals against fitted values of Import data (1949–1959) with the constraint $\beta_1 = \beta_3$.

may have been imposed on the basis of existing theory, it is still interesting to evaluate the effect of the constraint on the explanatory power of the full model. The values of R^2 for the full and restricted models are 0.992 and 0.987, respectively. The F-ratio for testing $H_0(\beta_1 = \beta_3)$ is 3.36 with 1 and 8 degrees of freedom. Both results suggest that the constraint is consistent with the data.

The constraint that $\beta_1 = \beta_3$ is, of course, only one example of the many types of constraints that may be used when specifying a regression model. The general class of possibilities is found in the set of linear constraints described in Chapter 3. Constraints are usually justified on the basis of underlying theory. They may often resolve what appears to be a problem of multicollinearity. In addition, any particular constraint may be viewed as a testable hypothesis and judged by the methods described in Chapter 3.

9.8 SEARCHING FOR LINEAR FUNCTIONS OF THE β'S

We assume that the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

has been carefully specified so that the regression coefficients appearing are of primary interest for policy analysis and decision making. We have seen that the presence of multicollinearity may prevent individual β's from being accurately estimated. However, as demonstrated below, it is always possible to estimate some linear functions of the β's accurately (Silvey, 1969). The obvious questions are: Which linear functions can be estimated, and of those that can be estimated, which are of interest in the analysis? In this section we use the data to help identify those linear functions that can be accurately estimated and, at the same time, have some value in the analysis.

Table 9.15 Regression Results When Fitting Model (9.26) to the Import Data (1949–1959)

Variable	Coefficient	s.e.	t-test	p-value
Constant	−8.440	1.435	−5.88	0.0004
DOPROD	0.145	0.007	20.70	< 0.0001
STOCK	0.622	0.128	4.87	0.0012
$n = 11$	$R^2 = 0.983$	$R_a^2 = 0.978$	$\hat{\sigma} = 0.667$	$d.f. = 8$

First we shall demonstrate in an indirect way that there are always linear functions of the β's that can be accurately estimated.[5] Consider once again the Import data. We have argued that there is a historical relationship between CONSUM and DOPROD that is approximated as CONSUM = (2/3) DOPROD. Replacing CONSUM in the original model,

$$\text{IMPORT} = \beta_0 + (\beta_1 + \frac{2}{3}\beta_3) \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \varepsilon. \qquad (9.26)$$

Equivalently stated, by dropping CONSUM from the equation we are able to obtain accurate estimates of $\beta_1 + (2/3)\beta_3$ and β_2. Multicollinearity is no longer present. The correlation between DOPROD and STOCK is 0.026. The results are given in Table 9.15. R^2 is almost unchanged and the residual plots (not shown) are satisfactory. In this case we have used information in addition to the data to argue that the coefficient of DOPROD in the regression of IMPORT on DOPROD and STOCK is the linear combination $\beta_1 + (2/3)\beta_3$. Also, we have demonstrated that this linear function can be estimated accurately even though multicollinearity is present in the data. Whether or not it is useful to know the value of $\beta_1 + (2/3)\beta_3$, of course, is another question. At least it is important to know that the estimate of the coefficient of DOPROD in this regression is not measuring the pure marginal effect of DOPROD, but includes part of the effect of CONSUM.

The above example demonstrates in an indirect way that there are always linear functions of the β's that can be accurately estimated. However, there is a constructive approach for identifying the linear combinations of the β's that can be accurately estimated. We shall use the advertising data introduced in Section 9.4 to demonstrate the method. The concepts are less intuitive than those found in the other sections of the chapter. We have attempted to keep things simple. A formal development of this problem is given in the Appendix to this chapter.

We begin with the linear transformation introduced in Section 9.6 that takes the standardized predictor variables into a new orthogonal set of variables. The standardized versions of the five predictor variables are denoted by $\tilde{X}_1, \ldots, \tilde{X}_5$. The standardized response variable, sales, is denoted by $\tilde{Y}$. The transformation

[5] Refer to the Appendix to this chapter for further treatment of this problem.

that takes $X_1, \ldots, X_5$ into the new set of orthogonal variables $C_1, \ldots, C_5$ is

$$\begin{aligned}
C_1 &= 0.532\tilde{X}_1 - 0.232\tilde{X}_2 - 0.389\tilde{X}_3 + 0.395\tilde{X}_4 - 0.595\tilde{X}_5, \\
C_2 &= -0.024\tilde{X}_1 + 0.825\tilde{X}_2 - 0.022\tilde{X}_3 - 0.260\tilde{X}_4 - 0.501\tilde{X}_5, \\
C_3 &= -0.668\tilde{X}_1 + 0.158\tilde{X}_2 - 0.217\tilde{X}_3 + 0.692\tilde{X}_4 - 0.057\tilde{X}_5, \\
C_4 &= 0.074\tilde{X}_1 - 0.037\tilde{X}_2 + 0.895\tilde{X}_3 + 0.338\tilde{X}_4 - 0.279\tilde{X}_5, \\
C_5 &= -0.514\tilde{X}_1 - 0.489\tilde{X}_2 + 0.010\tilde{X}_3 - 0.428\tilde{X}_4 - 0.559\tilde{X}_5.
\end{aligned} \tag{9.27}$$

The coefficients in the equation defining C_1 are the components of the eigenvector corresponding to the largest eigenvalue of the correlation matrix of the predictor variables. Similarly, the coefficients defining C_2 through C_5 are components of the eigenvectors corresponding to the remaining eigenvalues in order by size. The variables C_1, ..., C_5 are the PCs associated with the standardized versions of the predictors variables, as described in the preceding Section 9.6.

The regression model stated, as given in (9.4) in terms of the original variables is

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t. \tag{9.28}$$

In terms of standardized variables, the equation is written as

$$\tilde{Y} = \theta_1 \tilde{A}_t + \theta_2 \tilde{P}_t + \theta_3 \tilde{E}_t + \theta_4 \tilde{A}_{t-1} + \theta_5 \tilde{P}_{t-1} + \varepsilon', \tag{9.29}$$

where $\tilde{A}_t$ denotes the standardized version of the variable A_t. The regression coefficients in Equation (9.29) are often referred to as the *beta coefficients*. They represent marginal effects of the predictor variables in standard deviation units. For example, θ_1 measures the change in standardized units of sales (S) corresponding to an increase of one standard deviation unit in advertising (A).

Let $\hat{\beta}_j$ be the least squares estimate of β_j when model (9.28) is fit to the data. Similarly, let $\hat{\theta}_j$ be the least squares estimate of θ_j obtained from fitting model (9.29). Then $\hat{\beta}_j$ and $\hat{\theta}_j$ are related by

$$\begin{aligned}
\hat{\beta}_j &= (s_y/s_j)\hat{\theta}_j, \qquad j = 1, 2, 3, 4, 5, \\
\hat{\beta}_0 &= \bar{y} - \sum_{j=1}^{5} \hat{\beta}_j \bar{x}_j,
\end{aligned} \tag{9.30}$$

where $\bar{y}$ is the mean of Y and s_y and s_j are standard deviations of the response and jth predictor variable, respectively.

Equation (9.29) has an equivalent form, given as

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \varepsilon'. \tag{9.31}$$

The equivalence of Equations (9.29) and (9.31) results from the relationship between the $\tilde{X}$'s and C's in Equations (9.27) and the relationship between the α's and θ's and their estimated values, $\hat{\alpha}$'s and $\hat{\theta}$'s, given as

$$\begin{aligned}
\hat{\theta}_1 &= 0.532\hat{\alpha}_1 - 0.024\hat{\alpha}_2 - 0.668\hat{\alpha}_3 + 0.074\hat{\alpha}_4 - 0.514\hat{\alpha}_5, \\
\hat{\theta}_2 &= -0.232\hat{\alpha}_1 + 0.825\hat{\alpha}_2 + 0.158\hat{\alpha}_3 - 0.037\hat{\alpha}_4 - 0.489\hat{\alpha}_5, \\
\hat{\theta}_3 &= -0.389\hat{\alpha}_1 - 0.022\hat{\alpha}_2 - 0.217\hat{\alpha}_3 + 0.895\hat{\alpha}_4 + 0.010\hat{\alpha}_5, \\
\hat{\theta}_4 &= 0.395\hat{\alpha}_1 - 0.260\hat{\alpha}_2 + 0.692\hat{\alpha}_3 + 0.338\hat{\alpha}_4 - 0.428\hat{\alpha}_5, \\
\hat{\theta}_5 &= -0.595\hat{\alpha}_1 - 0.501\hat{\alpha}_2 - 0.057\hat{\alpha}_3 - 0.279\hat{\alpha}_4 - 0.559\hat{\alpha}_5.
\end{aligned} \tag{9.32}$$

Note that the transformation involves the same weights that are used to define Equation (9.27). The advantage of the transformed model is that the PCs are orthogonal. The precision of the estimated regression coefficients as measured by the variance of the $\hat{\alpha}$'s is easily evaluated. The estimated variance of $\hat{\alpha}_j$ is $\hat{\sigma}^2/\lambda_j$. It is inversely proportional to the ith eigenvalue. All but $\hat{\alpha}_5$ may be accurately estimated since only λ_5 is small. (Recall that $\lambda_1 = 1.701, \lambda_2 = 1.288, \lambda_3 = 1.145, \lambda_4 = 0.859$, and $\lambda_5 = 0.007$.)

Our interest in the $\hat{\alpha}$'s is only as a vehicle for analyzing the $\hat{\theta}$'s. From the representation of Equation (9.32) it is a simple matter to compute and analyze the variances and, in turn, the standard errors of the $\hat{\theta}$'s. The variance of $\hat{\theta}_j$ is

$$Var(\hat{\theta}_j) = \sum_{i=1}^{p} v_{ij}^2 Var(\hat{\alpha}_i), \; j = 1, \ldots, p, \tag{9.33}$$

where v_{ij} is the coefficient of $\hat{\alpha}_i$ in the j Equation in (9.32). Since the estimated variance of $\hat{\alpha}_i = \hat{\sigma}^2/\lambda_i$, where $\hat{\sigma}^2$ is the residual mean square, (9.33) becomes

$$Var(\hat{\theta}_j) = \hat{\sigma}^2 \sum_{i=1}^{p} \frac{v_{ij}^2}{\lambda_i}\,. \tag{9.34}$$

For example, the estimated variance of $\hat{\theta}_1$ is

$$\hat{\sigma}^2 \left[\frac{(0.532)^2}{\lambda_1} + \frac{(-0.024)^2}{\lambda_2} + \frac{(-0.668)^2}{\lambda_3} + \frac{(0.074)^2}{\lambda_4} + \frac{(-0.514)^2}{\lambda_5} \right]. \tag{9.35}$$

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_5$ and only λ_5 is small, $(\lambda_5 = 0.007)$. Therefore, it is only the last term in the expression for the variance that is large and could destroy the precision of $\hat{\theta}_1$. Since expressions for the variances of the other $\hat{\theta}_j$'s are similar to Equation (9.35), a requirement for small variance is equivalent to the requirement that the coefficient of $1/\lambda_5$ be small. Scanning the equations that define the transformation from $\{\hat{\alpha}_i\}$ to $\{\hat{\theta}_j\}$, we see that $\hat{\theta}_3$ is the most precise estimate since the coefficient of $1/\lambda_5$ in the variance expression for $\hat{\theta}_3$ is $(-0.01)^2 = 0.0001$.

Expanding this type of analysis, it may be possible to identify meaningful linear functions of the θ's that can be more accurately estimated than individual θ's. For example, we may be more interested in estimating $\theta_1 - \theta_2$ than θ_1 and θ_2 separately. In the sales model, $\theta_1 - \theta_2$ measures the increment to sales that corresponds to increasing the current year's advertising budget, X_1, by one unit and simultaneously reducing the current year's promotions budget, X_2, by one unit. In other words, $\theta_1 - \theta_2$ represents the effect of a shift in the use of resources in the current year. The estimate of $\theta_1 - \theta_2$ is $\hat{\theta}_1 - \hat{\theta}_2$. The variance of this estimate is obtained simply by subtracting the equation for $\hat{\theta}_2$ and $\hat{\theta}_1$ in (9.32) and using the resulting coefficients of the $\hat{\alpha}$'s as before. That is,

$$\hat{\theta}_1 - \hat{\theta}_2 = 0.764\hat{\alpha}_1 - 0.849\hat{\alpha}_2 - 0.826\hat{\alpha}_3 + 0.111\hat{\alpha}_4 - 0.025\hat{\alpha}_5,$$

from which we obtain the estimated variance of $(\hat{\theta}_1 - \hat{\theta}_2)$ as

$$\begin{aligned}(0.764)^2\ Var(\hat{\alpha}_1) + (-0.849)^2\ Var(\hat{\alpha}_2) + (-0.826)^2\ Var(\hat{\alpha}_3) \\ + (0.111)^2\ Var(\hat{\alpha}_4) + (-0.025)^2\ Var(\hat{\alpha}_1),\end{aligned} \tag{9.36}$$

or, equivalently, as

$$\hat{\sigma}^2\left[\frac{(0.764)^2}{\lambda_1} + \frac{(-0.849)^2}{\lambda_2} + \frac{(-0.826)^2}{\lambda_3} + \frac{(0.111)^2}{\lambda_4} + \frac{(-0.025)^2}{\lambda_5}\right]. \tag{9.37}$$

The small coefficient of $1/\lambda_5$ makes it possible to estimate $\theta_1 - \theta_2$ accurately. Generalizing this procedure we see that any linear function of the θ's that results in a small coefficient for $1/\lambda_5$ in the variance expression can be estimated with precision.

9.9 COMPUTATIONS USING PRINCIPAL COMPONENTS

The computations required for this analysis involve something in addition to a standard least squares computer program. The raw data must be processed through a principal components subroutine that operates on the correlation matrix of the predictor variables in order to compute the eigenvalues and the transformation weights found in Equations (9.32). Most regression packages produce the estimated beta coefficients as part of the standard output.

For the advertising data, the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_5$ can be computed in two equivalent ways. They can be obtained directly from a regression of the standardized variables as represented in Equation (9.29). The results of this regression are given in Table 9.16. Alternatively, we can fit the model in (9.31) by the least squares regression of the standardized response variable on the the five PCs and obtain the estimates $\hat{\alpha}_1, \ldots, \hat{\alpha}_5$. The results of this regression are shown in Table 9.17. Then, we use (9.32) to obtain $\hat{\theta}_1, \ldots, \hat{\theta}_5$. For example,

$$\begin{aligned}\hat{\theta}_1 &= (0.532)(-0.346019) + (-0.024)(0.417889) + (-0.668)(-0.151328) \\ &\quad + (0.074)(0.659946) + (-0.514)(-1.22026) = 0.5830.\end{aligned}$$

Using the coefficients in (9.32), the standard error of $\hat{\theta}_1, \ldots, \hat{\theta}_5$ can be computed. For example the estimated variance of $\hat{\theta}_1$ is

$$(0.532 \times \text{s.e.}(\hat{\alpha}_1))^2 + (-0.024 \times \text{s.e.}(\hat{\alpha}_2))^2 + (-0.668 \times \text{s.e.}(\hat{\alpha}_3))^2$$

Table 9.16 Regression Results Obtained From Fitting the Model in (9.29)

Variable	Coefficient	s.e.	t-test	p-value
$\tilde{X}_1$	0.583	0.438	1.33	0.2019
$\tilde{X}_2$	0.973	0.417	2.33	0.0329
$\tilde{X}_3$	0.786	0.075	10.50	< 0.0001
$\tilde{X}_4$	0.395	0.367	1.08	0.2973
$\tilde{X}_5$	0.503	0.476	1.06	0.3053
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 0.3303$	$d.f. = 16$

Table 9.17 Regression Results Obtained From Fitting the Model in (9.31)

Variable	Coefficient	s.e.	t-test	p-value
C_1	−0.346	0.053	−6.55	< 0.0001
C_2	0.418	0.064	6.58	< 0.0001
C_3	−0.151	0.067	−2.25	0.0391
C_4	0.660	0.078	8.46	< 0.0001
C_5	−1.220	0.846	−1.44	0.1683
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 0.3303$	$d.f. = 16$

$$+ (0.074 \times \text{s.e.}(\hat{\alpha}_4))^2 + (-0.514 \times \text{s.e.}(\hat{\alpha}_5))^2 = (0.532 \times 0.0529)^2$$
$$+ (-0.024 \times 0.0635)^2 + (-0.668 \times 0.0674)^2 + (0.074 \times 0.0780)^2$$
$$(-0.514 \times 0.8456)^2 = 0.1918,$$

which means that the standard error of $\hat{\theta}_1$ is

$$\text{s.e.}(\hat{\theta}_1) = \sqrt{0.1918} = 0.438.$$

It should be noted that the t-values for testing β_j and θ_j equal to zero are identical. The beta coefficient, θ_j is a scaled version of β_j. When constructing t-values as either $\hat{\beta}_j/\text{s.e.}(\hat{\beta}_j)$, or $\hat{\theta}_j/\text{s.e.}(\hat{\theta}_j)$, the scale factor is canceled.

The estimate of $\theta_1 - \theta_2$ is 0.583 −0.973 = −0.390. The variance of $\hat{\theta}_1 - \hat{\theta}_2$ can be computed from Equation (9.36) as 0.008. A 95% confidence interval for $\theta_1 - \theta_2$ is $-0.390 \pm 2.12\sqrt{0.008}$ or −0.58 to −0.20. That is, the effect of shifting one unit of expenditure from promotions to advertising in the current year is a loss of between 0.20 and 0.58 standardized sales unit.

There are other linear functions that may also be accurately estimated. Any function that produces a small coefficient for $1/\lambda_5$ in the variance expression is a possibility. For example, Equations (9.31) suggest that all differences involving $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_4$, and $\hat{\theta}_5$ can be considered. However, some of the differences are meaningful in the problem, whereas others are not. For example, the difference $(\theta_1 - \theta_2)$

is meaningful, as described previously. It represents a shift in current expenditures from promotions to advertising. The difference $\theta_1 - \theta_4$ is not particularly meaningful. It represents a shift from current advertising expenditure to a previous year's advertising expenditure. A shift of resources backward in time is impossible. Even though $\theta_1 - \theta_4$ could be accurately estimated, it is not of interest in the analysis of sales.

In general, when the weights in Equation (9.32) are displayed and the corresponding values of the eigenvalues are known, it is always possible to scan the weights and identify those linear functions of the original regression coefficients that can be accurately estimated. Of those linear functions that can be accurately estimated, only some will be of interest for the problem being studied.

To summarize, where multicollinearity is indicated and it is not possible to supplement the data, it may still be possible to estimate some regression coefficients and some linear functions accurately. To investigate which coefficients and linear functions can be estimated, we recommend the analysis (transformation to principal components) that has just been described. This method of analysis will not overcome multicollinearity if it is present. There will still be regression coefficients and functions of regression coefficients that cannot be estimated. But the recommended analysis will indicate those functions that are estimable and indicate the structural dependencies that exist among the predictor variables.

9.10 BIBLIOGRAPHIC NOTES

The principal components techniques used in this chapter are derived in most books on multivariate statistical analysis. It should be noted that principal components analysis involves only the predictor variables. The analysis is aimed at characterizing and identifying dependencies (if they exist) among the predictor variables. For a comprehensive discussion of principal components, the reader is referred to Johnson and Wichern (1992) or Seber (1984). Several statistical software packages are now commercially available to carry out the analysis described in this chapter.

EXERCISES

9.1 In the analysis of the Advertising data in Section 9.4 it is suggested that the regression of sales S_t against E_t and three of the remaining four variables $(A_t, P_t, A_{t-1}, S_{t-1})$ may resolve the collinearity problem. Run the four suggested regressions and, for each of them, examine the resulting VIF$_j$'s to see if collinearity has been eliminated.

9.2 Gasoline Consumption: To study the factors that determine the gasoline consumption of cars, data were collected on 30 models of cars. Besides the gasoline consumption (Y), measured in miles per gallon for each car, 11 other measurements representing physical and mechanical characteristics are given. The source of the data in Table 9.19 is *Motor Trend* magazine for the year

1975. Definitions of variables are given in Table 9.18. We wish to determine whether the data set is collinear.

(a) Compute the correlation matrix of the predictor variables $X_1, \ldots, X_{11}$ and the corresponding pairwise scatter plots. Identify any evidence of collinearity.

(b) Compute the eigenvalues, eigenvectors, and the condition number of the correlation matrix. Is multicollinearity present in the data?

(c) Identify the variables involved in multicollinearity by examining the eigenvectors corresponding to small eigenvalues.

(d) Regress Y on the 11 predictor variables and compute the VIF for each of the predictors. Which predictors are affected by the presence of collinearity?

9.3 Refer to the Presidential Election Data in Table 5.17 and consider fitting a model relating V to all the variables (including a time trend representing year of election) plus as many interaction terms involving two or three variables as you possibly can.

(a) What is the maximum number of terms (coefficients) in a linear regression model that you can fit to these data? [*Hint*: Consider the number of observations in the data.]

(b) Examine the predictor variables in the above model for the presence of multicollinearity. (Compute the correlation matrix, the condition number, and the VIFs.)

(c) Identify the subsets of variables involved in collinearity. Attempt to solve the multicollinearity problem by deleting some of the variables involved in multicollinearity.

(d) Fit a model relating V to the set of predictors you found to be free from multicollinearity.

Appendix: Principal Components

In this appendix we present the principal components approach to the detection of multicollinearity using matrix notation.

A. The Model

The regression model can be expressed as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations on the response variable, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ is an $n \times p$ matrix of n observations on p predictor variables, $\boldsymbol{\theta}$ is a $p \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. It is assumed

Table 9.18 Variables for the Gasoline Consumption Data in Table 9.19

Variable	Definition
Y	Miles/gallon
X_1	Displacement (cubic inches)
X_2	Horsepower (feet/pound)
X_3	Torque (feet/pound)
X_4	Compression ratio
X_5	Rear axle ratio
X_6	Carburetor (barrels)
X_7	Number of transmission speeds
X_8	Overall length (inches)
X_9	Width (inches)
X_{10}	Weight (pounds)
X_{11}	Type of transmission (1 = automatic; 0 = manual)

Table 9.19 Gasoline Consumption and Automotive Variables.

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
18.9	350.0	165	260	8.00	2.56	4	3	200.3	69.9	3910	1
17.0	350.0	170	275	8.50	2.56	4	3	199.6	72.9	3860	1
20.0	250.0	105	185	8.25	2.73	1	3	196.7	72.2	3510	1
18.3	351.0	143	255	8.00	3.00	2	3	199.9	74.0	3890	1
20.1	225.0	95	170	8.40	2.76	1	3	194.1	71.8	3365	0
11.2	440.0	215	330	8.20	2.88	4	3	184.5	69.0	4215	1
22.1	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3020	1
21.5	262.0	110	200	8.50	2.56	2	3	179.3	65.4	3180	1
34.7	89.7	70	81	8.20	3.90	2	4	155.7	64.0	1905	0
30.4	96.9	75	83	9.00	4.30	2	5	165.2	65.0	2320	0
16.5	350.0	155	250	8.50	3.08	4	3	195.4	74.4	3885	1
36.5	85.3	80	83	8.50	3.89	2	4	160.6	62.2	2009	0
21.5	171.0	109	146	8.20	3.22	2	4	170.4	66.9	2655	0
19.7	258.0	110	195	8.00	3.08	1	3	171.5	77.0	3375	1
20.3	140.0	83	109	8.40	3.40	2	4	168.8	69.4	2700	0
17.8	302.0	129	220	8.00	3.00	2	3	199.9	74.0	3890	1
14.4	500.0	190	360	8.50	2.73	4	3	224.1	79.8	5290	1
14.9	440.0	215	330	8.20	2.71	4	3	231.0	79.7	5185	1
17.8	350.0	155	250	8.50	3.08	4	3	196.7	72.2	3910	1
16.4	318.0	145	255	8.50	2.45	2	3	197.6	71.0	3660	1
23.5	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3050	1
21.5	360.0	180	290	8.40	2.45	2	3	214.2	76.3	4250	1
31.9	96.9	75	83	9.00	4.30	2	5	165.2	61.8	2275	0
13.3	460.0	223	366	8.00	3.00	4	3	228.0	79.8	5430	1
23.9	133.6	96	120	8.40	3.91	2	5	171.5	63.4	2535	0
19.7	318.0	140	255	8.50	2.71	2	3	215.3	76.3	4370	1
13.9	351.0	148	243	8.00	3.25	2	3	215.5	78.5	4540	1
13.3	351.0	148	243	8.00	3.26	2	3	216.1	78.5	4715	1
13.8	360.0	195	295	8.25	3.15	4	3	209.3	77.4	4215	1
16.5	350.0	165	255	8.50	2.73	4	3	185.2	69.0	3660	1

that $E(\varepsilon) = \mathbf{0}$, $E(\varepsilon\varepsilon^T) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of order n. It is also assumed, without loss of generality, that $\mathbf{Y}$ and $\mathbf{Z}$ have been centered and scaled so that $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{Z}^T\mathbf{Y}$ are matrices of correlation coefficients.

There exist square matrices, $\mathbf{\Lambda}$ and $\mathbf{V}$ satisfying[6]

$$\mathbf{V}^T(\mathbf{Z}^T\mathbf{Z})\mathbf{V} = \mathbf{\Lambda} \text{ and } \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}. \qquad \text{(A.2)}$$

The matrix $\mathbf{\Lambda}$ is diagonal with the ordered eigenvalues of $\mathbf{Z}^T\mathbf{Z}$ on the diagonal. These eigenvalues are denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The columns of $\mathbf{V}$ are the normalized eigenvectors corresponding to $\lambda_1, \ldots, \lambda_p$. Since $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, the regression model in (A.1) can be restated in terms of the PCs as

$$\mathbf{Y} = \mathbf{Z}\mathbf{V}\mathbf{V}^T\boldsymbol{\theta} + \varepsilon = \mathbf{C}\boldsymbol{\alpha} + \varepsilon, \qquad \text{(A.3)}$$

where

$$\mathbf{C} = \mathbf{Z}\mathbf{V}; \text{ and } \boldsymbol{\alpha} = \mathbf{V}^T\boldsymbol{\theta}. \qquad \text{(A.4)}$$

The matrix $\mathbf{C}$ contains p columns $\mathbf{C}_1, \ldots, \mathbf{C}_p$, each of which is a linear functions of the predictor variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$. The columns of $\mathbf{C}$ are orthogonal and are referred to as principal components (PCs) of the predictor variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$. The columns of $\mathbf{C}$ satisfy $\mathbf{C}_j^T\mathbf{C}_j = \lambda_j$ and $\mathbf{C}_i^T\mathbf{C}_j = 0$ for $i \neq j$.

The PCs and the eigenvalues may be used to detect and analyze collinearity in the predictor variables. The restatement of the regression model given in Equation (A.3) is a reparameterization of Equation (A.1) in terms of orthogonal predictor variables. The λ's may be viewed as sample variances of the PCs. If $\lambda_i = 0$, all observations on the ith PC are also zero. Since the jth PC is a linear function of $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$, when $\lambda_j = 0$ an exact linear dependence exists among the predictor variables. It follows that when λ_j is small (approximately equal to zero) there is an approximate linear relationship among the predictor variables. That is, a small eigenvalue is an indicator of multicollinearity. In addition, from Equation (A.4) we have

$$\mathbf{C}_j = \sum_{i=1}^{p} v_{ij} Z_i,$$

which identifies the exact form of the linear relationship that is causing the multicollinearity.

B. Precision of Linear Functions of $\hat{\theta}$

Denoting $\hat{\alpha}$ and $\hat{\boldsymbol{\theta}}$ as the least squares estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, respectively, it can be shown that $\hat{\alpha} = \mathbf{V}^T\hat{\boldsymbol{\theta}}$, and conversely, $\hat{\boldsymbol{\theta}} = \mathbf{V}\hat{\alpha}$. With $\hat{\alpha} = (\mathbf{C}^T\mathbf{C})^{-1}$ $\mathbf{C}^T\mathbf{Y}$, it follows that the variance-covariance matrix of $\hat{\alpha}$ is $\mathbf{V}(\hat{\alpha}) = \mathbf{\Lambda}^{-1}\sigma^2$, and the corresponding matrix for $\hat{\boldsymbol{\theta}}$ is $\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\sigma^2$. Let $\mathbf{L}$ be an arbitrary $p \times 1$ vector of constants. The linear function $\delta = \mathbf{L}^T\boldsymbol{\theta}$ has least squares estimator

[6] See, for example, Strang (1988) or Hadi (1996).

$\hat{\delta} = \mathbf{L}^T\hat{\boldsymbol{\theta}}$ and variance

$$Var(\hat{\delta}) = \mathbf{L}^T\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T\mathbf{L}\sigma^2. \tag{A.5}$$

Let $\mathbf{V}_j$ be the jth column of $\mathbf{V}$. Then $\mathbf{L}$ can be represented as

$$\mathbf{L} = \sum_{j=1}^{p} r_j\mathbf{V}_j$$

for appropriately chosen constants $r_1, \ldots, r_p$. Then (A.5) becomes $Var(\hat{\delta}) = \mathbf{R}^T\boldsymbol{\Lambda}^{-1}\mathbf{R}\sigma^2$ or, equivalently,

$$Var(\hat{\delta}) = \left(\sum_{j=1}^{p} \frac{r_j^2}{\lambda_j}\right)\sigma^2, \tag{A.6}$$

where $\boldsymbol{\Lambda}^{-1}$ is the inverse of $\boldsymbol{\Lambda}$.

To summarize, the variance of $\hat{\delta}$ is a linear combination of the reciprocals of the eigenvalues. It follows that $\hat{\delta}$ will have good precision either if none of the eigenvalues are near zero or if r_j^2 is at most the same magnitude as λ_j when λ_j is small. Furthermore, it is always possible to select a vector, $\mathbf{L}$, and thereby a linear function of $\hat{\boldsymbol{\theta}}$, so that the effect of one or few small eigenvalues is eliminated and $\mathbf{L}^T\hat{\boldsymbol{\theta}}$ has a small variance. Refer to Silvey (1969) for a more complete development of these concepts.