Need for More Powerful Diagnostics for Multicollinearity

As we have seen, multicollinearity among the predictor variables can have important consequences for interpreting and using a fitted regression model. The diagnostic tool considered here for identifying multicollinearity—namely, the pairwise coefficients of simple correlation between the predictor variables—is frequently helpful. Often, however, serious multicollinearity exists without being disclosed by the pairwise correlation coefficients. In Chapter 10, we present a more powerful tool for identifying the existence of serious multicollinearity. Some remedial measures for lessening the effects of multicollinearity will be considered in Chapter 11.

Cited Reference

7.1. Kennedy, W. J., Jr., and J. E. Gentle. *Statistical Computing*. New York: Marcel Dekker, 1980.

Problems

7.1. State the number of degrees of freedom that are associated with each of the following extra sums of squares: (1) $SSR(X_1|X_2)$; (2) $SSR(X_2|X_1, X_3)$; (3) $SSR(X_1, X_2|X_3, X_4)$; (4) $SSR(X_1, X_2, X_3|X_4, X_5)$.

7.2. Explain in what sense the regression sum of squares $SSR(X_1)$ is an extra sum of squares.

7.3. Refer to **Brand preference** Problem 6.5.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1 and with X_2, given X_1.

b. Test whether X_2 can be dropped from the regression model given that X_1 is retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.4. Refer to **Grocery retailer** Problem 6.9.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1; with X_3, given X_1; and with X_2, given X_1 and X_3.

b. Test whether X_2 can be dropped from the regression model given that X_1 and X_3 are retained. Use the F^* test statistic and $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. Does $SSR(X_1) + SSR(X_2|X_1)$ equal $SSR(X_2) + SSR(X_1|X_2)$ here? Must this always be the case?

*7.5. Refer to **Patient satisfaction** Problem 6.15.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

b. Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. Use the F^* test statistic and level of significance .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.6. Refer to **Patient satisfaction** Problem 6.15. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 is retained. Use $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

7.7. Refer to **Commercial properties** Problem 6.18.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_4; with X_1, given X_4; with X_2, given X_1 and X_4; and with X_3, given X_1, X_2 and X_4.

b. Test whether X_3 can be dropped from the regression model given that X_1, X_2 and X_4 are retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

7.8. Refer to **Commercial properties** Problems 6.18 and 7.7. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 and X_4 are retained; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.9. Refer to **Patient satisfaction** Problem 6.15. Test whether $\beta_1 = -1.0$ and $\beta_2 = 0$; use $\alpha = .025$. State the alternatives, full and reduced models, decision rule, and conclusion.

7.10. Refer to **Commercial properties** Problem 6.18. Test whether $\beta_1 = -.1$ and $\beta_2 = .4$; use $\alpha = .01$. State the alternatives, full and reduced models, decision rule, and conclusion.

7.11. Refer to the work crew productivity example in Table 7.6.

a. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2. Explain what each coefficient measures and interpret your results.

b. Are any of the results obtained in part (a) special because the two predictor variables are uncorrelated?

7.12. Refer to **Brand preference** Problem 6.5. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2. Explain what each coefficient measures and interpret your results.

*7.13. Refer to **Grocery retailer** Problem 6.9. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, $R^2_{Y2|13}$, and R^2. Explain what each coefficient measures and interpret your results.

*7.14. Refer to **Patient satisfaction** Problem 6.15.

a. Calculate R^2_{Y1}, $R^2_{Y1|2}$, and $R^2_{Y1|23}$. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_2? When adjusted for both X_2 and X_3?

b. Make a similar analysis to that in part (a) for the degree of marginal linear association between Y and X_2. Are your findings similar to those in part (a) for Y and X_1?

7.15. Refer to **Commercial properties** Problems 6.18 and 7.7. Calculate R^2_{Y4}, R^2_{Y1}, $R^2_{Y1|4}$, R^2_{14}, $R^2_{Y2|14}$, $R^2_{Y3|124}$, and R^2. Explain what each coefficient measures and interpret your results. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_4?

7.16. Refer to **Brand preference** Problem 6.5.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Interpret the standardized regression coefficient b_1^*.

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.5b.

*7.17. Refer to **Grocery retailer** Problem 6.9.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Calculate the coefficients of determination between all pairs of predictor variables. Is it meaningful here to consider the standardized regression coefficients to reflect the effect of one predictor variable when the others are held constant?

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.10a.

*7.18. Refer to **Patient satisfaction** Problem 6.15.

b. Compare the estimated regression coefficient for total cases shipped obtained in part (a) with the corresponding coefficient obtained in Problem 6.10a. What do you find?

c. Does $SSR(X_1)$ equal $SSR(X_1|X_2)$ here? If not, is the difference substantial?

d. Refer to the correlation matrix obtained in Problem 6.9c. What bearing does this have on your findings in parts (b) and (c)?

*7.26. Refer to **Patient satisfaction** Problem 6.15.

a. Fit first-order linear regression model (6.1) for relating patient satisfaction (Y) to patient's age (X_1) and severity of illness (X_2). State the fitted regression function.

b. Compare the estimated regression coefficients for patient's age and severity of illness obtained in part (a) with the corresponding coefficients obtained in Problem 6.15c. What do you find?

c. Does $SSR(X_1)$ equal $SSR(X_1|X_3)$ here? Does $SSR(X_2)$ equal $SSR(X_2|X_3)$?

d. Refer to the correlation matrix obtained in Problem 6.15b. What bearing does it have on your findings in parts (b) and (c)?

7.27. Refer to **Commercial properties** Problem 6.18.

a. Fit first-order linear regression model (6.1) for relating rental rates (Y) to property age (X_1) and size (X_4). State the fitted regression function.

b. Compare the estimated regression coefficients for property age and size with the corresponding coefficients obtained in Problem 6.18c. What do you find?

c. Does $SSR(X_4)$ equal $SSR(X_4|X_3)$ here? Does $SSR(X_1)$ equal $SSR(X_1|X_3)$?

d. Refer to the correlation matrix obtained in Problem 6.18b. What bearing does this have on your findings in parts (b) and (c)?

Exercises

7.28. a. Define each of the following extra sums of squares: (1) $SSR(X_5|X_1)$; (2) $SSR(X_3, X_4|X_1)$; (3) $SSR(X_4|X_1, X_2, X_3)$.

b. For a multiple regression model with five X variables, what is the relevant extra sum of squares for testing whether or not $\beta_5 = 0$? whether or not $\beta_2 = \beta_4 = 0$?

7.29. Show that:

a. $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3)$.

b. $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3)$.

7.30. Refer to **Brand preference** Problem 6.5.

a. Regress Y on X_2 using simple linear regression model (2.1) and obtain the residuals.

b. Regress X_1 on X_2 using simple linear regression model (2.1) and obtain the residuals.

c. Calculate the coefficient of simple correlation between the two sets of residuals and show that it equals $r_{Y1|2}$.

7.31. The following regression model is being considered in a water resources study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \varepsilon_i$$

State the reduced models for testing whether or not: (1) $\beta_3 = \beta_4 = 0$, (2) $\beta_3 = 0$, (3) $\beta_1 = \beta_2 = 5$, (4) $\beta_4 = 7$.

7.32. The following regression model is being considered in a market research study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$$

b. Plot a set of contour curves for the response surface. How is the interaction effect of X_1 and X_2 on Y apparent from this graph?

8.10. Consider the response function $E\{Y\} = 14 + 7X_1 + 5X_2 - 4X_1X_2$.

a. Prepare a conditional effects plot of the response function against X_2 when $X_1 = 1$ and when $X_1 = 4$. How does the graph indicate that the effects of X_1 and X_2 on Y are not additive? What is the nature of the interaction effect?

b. Plot a set of contour curves for the response surface. How does the graph indicate that the effects of X_1 and X_2 on Y are not additive?

8.11. Refer to **Brand preference** Problem 6.5.

a. Fit regression model (8.22).

b. Test whether or not the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

8.12. A student who used a regression model that included indicator variables was upset when receiving only the following output on the multiple regression printout: XTRANSPOSE X SINGULAR. What is a likely source of the difficulty?

8.13. Refer to regression model (8.33). Portray graphically the response curves for this model if $\beta_0 = 25.3$, $\beta_1 = .20$, and $\beta_2 = -12.1$.

8.14. In a regression study of factors affecting learning time for a certain task (measured in minutes), gender of learner was included as a predictor variable (X_2) that was coded $X_2 = 1$ if male and 0 if female. It was found that $b_2 = 22.3$ and $s\{b_2\} = 3.8$. An observer questioned whether the coding scheme for gender is fair because it results in a positive coefficient, leading to longer learning times for males than females. Comment.

8.15. Refer to **Copier maintenance** Problem 1.20. The users of the copiers are either training institutions that use a small model, or business firms that use a large, commercial model. An analyst at Tri-City wishes to fit a regression model including both number of copiers serviced (X_1) and type of copier (X_2) as predictor variables and estimate the effect of copier model (S—small, L—large) on number of minutes spent on the service call. Records show that the models serviced in the 45 calls were:

i:	1	2	3	...	43	44	45
X_{i2}:	S	L	L	...	L	L	L

Assume that regression model (8.33) is appropriate, and let $X_2 = 1$ if small model and 0 if large, commercial model.

a. Explain the meaning of all regression coefficients in the model.

b. Fit the regression model and state the estimated regression function.

c. Estimate the effect of copier model on mean service time with a 95 percent confidence interval. Interpret your interval estimate.

d. Why would the analyst wish to include X_1, number of copiers, in the regression model when interest is in estimating the effect of type of copier model on service time?

e. Obtain the residuals and plot them against X_1X_2. Is there any indication that an interaction term in the regression model would be helpful?

8.16. Refer to **Grade point average** Problem 1.19. An assistant to the director of admissions conjectured that the predictive power of the model could be improved by adding information on whether the student had chosen a major field of concentration at the time the application was submitted. Assume that regression model (8.33) is appropriate, where X_1 is entrance test score and $X_2 = 1$ if student had indicated a major field of concentration at the time of application and 0 if the major field was undecided. Data for X_2 were as follows:

i:	1	2	3	...	118	119	120
X_{i2}:	0	1	0	...	1	1	0

a. Explain how each regression coefficient in model (8.33) is interpreted here.

b. Fit the regression model and state the estimated regression function.

c. Test whether the X_2 variable can be dropped from the regression model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

d. Obtain the residuals for regression model (8.33) and plot them against X_1X_2. Is there any evidence in your plot that it would be helpful to include an interaction term in the model?

8.17. Refer to regression models (8.33) and (8.49). Would the conclusion that $\beta_2 = 0$ have the same implication for each of these models? Explain.

8.18. Refer to regression model (8.49). Portray graphically the response curves for this model if $\beta_0 = 25$, $\beta_1 = .30$, $\beta_2 = -12.5$, and $\beta_3 = .05$. Describe the nature of the interaction effect.

*8.19. Refer to **Copier maintenance** Problems 1.20 and 8.15.

a. Fit regression model (8.49) and state the estimated regression function.

b. Test whether the interaction term can be dropped from the model; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of the test? If the interaction term cannot be dropped from the model, describe the nature of the interaction effect.

8.20. Refer to **Grade point average** Problems 1.19 and 8.16.

a. Fit regression model (8.49) and state the estimated regression function.

b. Test whether the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. If the interaction term cannot be dropped from the model, describe the nature of the interaction effect.

8.21. In a regression analysis of on-the-job head injuries of warehouse laborers caused by falling objects, Y is a measure of severity of the injury, X_1 is an index reflecting both the weight of the object and the distance it fell, and X_2 and X_3 are indicator variables for nature of head protection worn at the time of the accident, coded as follows:

Type of Protection	X_2	X_3
Hard hat	1	0
Bump cap	0	1
None	0	0

The response function to be used in the study is $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

a. Develop the response function for each type of protection category.

b. For each of the following questions, specify the alternatives H_0 and H_a for the appropriate test: (1) With X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection? (2) With X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

8.22. Refer to tool wear regression model (8.36). Suppose the indicator variables had been defined as follows: $X_2 = 1$ if tool model M2 and 0 otherwise, $X_3 = 1$ if tool model M3 and 0 otherwise, $X_4 = 1$ if tool model M4 and 0 otherwise. Indicate the meaning of each of the following: (1) β_3, (2) $\beta_4 - \beta_3$, (3) β_1.