# ILL-CONDITIONING AND COLLINEARITY

From Graybill FA and Iyer H K
***Regression Analysis: Concepts and Applications***
Duxbury Press Belmont Ca 1994

EXAMPLE 5.5.2 (p394-397)

Suppose we want to develop a function for predicting the weight Y using age (X1), and length (X2) for babies with ages ranging from 1 month to 12 months, and that a sample of size 12 was selected by first preselecting the ages and then randomly choosing one baby from each preselected age group. The length and weight of each chosen baby are recorded along with age. The data are displayed in Table 5.5.2.

| **Observation Number** | **Weight Y (pounds)** | **Age X1 (months)** | **Length X2 (inches)** |
|---|---|---|---|
| 1 | 9.2 | 1 | 20.4 |
| 2 | 9.8 | 2 | 20.9 |
| 3 | 9.1 | 3 | 22.1 |
| 4 | 9.6 | 4 | 21.7 |
| 5 | 11.7 | 5 | 22.9 |
| 6 | 10.7 | 6 | 24.2 |
| 7 | 12.7 | 7 | 24.9 |
| 8 | 13.0 | 8 | 26.1 |
| 9 | 13.4 | 9 | 26.9 |
| 10 | 14.7 | 10 | 27.6 |
| 11 | 14.4 | 11 | 28.1 |
| 12 | 15.2 | 12 | 29.2 |

Suppose that assumptions (A) for regression are satisfied with

$$\mu_y(X1,X2) = \beta_0 + \beta_1 X1 + \beta_2 X2 \qquad (5.5.1)$$

We estimate the parameters, $\beta_i$ using the formula (4.4.8) and carrying out all calculations exactly (by hand). The results are

$$X^TX = \begin{matrix} 12 & 78 & 295 \\ 78 & 650 & 2035.7 \\ 295 & 2035.7 & 7351.16 \end{matrix}$$

$$X^T\mathbf{y} = \begin{matrix} 143.5 \\ 1018.5 \\ 3598.99 \end{matrix}$$

$$(X^TX)^{-1} = \begin{matrix} 63417951 / 236068 & 1357051 / 118034 & -824135 / 59017 \\ 1357051 / 118034 & 29723 / 59017 & -35460 / 59017 \\ -824135 / 59017 & -35460 / 59017 & 42900 / 59017 \end{matrix}$$

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} = \begin{matrix} 5743609 / 2360680 \\ 421987 / 1180340 \\ 34577 / 118034 \end{matrix} = \begin{matrix} 2.433031584 \\ 0.357513089 \\ 0.292941017 \end{matrix}$$

(final result rounded to 9 decimals)

Now suppose that X2 and Y values are measured slightly inaccurately, resulting in the data in Table 5.5.3. Note that the data values in Table 5.5.3 differ from the corresponding values in Table 5.5.2 by at most plus or minus 0.1.

| **Observation Number** | **Weight Y (pounds)** | **Age X1 (months)** | **Length X2 (inches)** |
|---|---|---|---|
| 1 | 9.3 | 1 | 20.5 |
| 2 | 9.7 | 2 | 20.8 |
| 3 | 9.2 | 3 | 22.2 |
| 4 | 9.5 | 4 | 21.6 |
| 5 | 11.8 | 5 | 23.0 |
| 6 | 10.6 | 6 | 24.1 |
| 7 | 12.8 | 7 | 25.0 |
| 8 | 13.9 | 7 | 26.0 |
| 9 | 13.5 | 9 | 27.0 |
| 10 | 14.6 | 10 | 27.5 |
| 11 | 14.5 | 11 | 28.2 |
| 12 | 15.1 | 12 | 29.1 |

We again estimate the parameters $\beta_i$ using formula (4.4.8) and carrying out all calculations exactly (by hand). The results are

$$X^TX = \begin{matrix} 12 & 78 & 295 \\ 78 & 650 & 2035.1 \\ 295 & 2035.1 & 7350.40 \end{matrix}$$

$$X^T\mathbf{y} = \begin{matrix} 143.5 \\ 1017.9 \\ 3598.42 \end{matrix}$$

$$(X^TX)^{-1} = \begin{matrix} 63612799 / 275428 & 1351165 / 137714 & -825305 / 68857 \\ 1351165 / 137714 & 29495 / 68857 & -35280 / 68857 \\ -825305 / 68857 & -35280 / 68857 & 42900 / 68857 \end{matrix}$$

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} = \begin{matrix} -377089 / 2754280 \\ 335833 / 1377140 \\ 58877 / 137714 \end{matrix} = \begin{matrix} -0.13691019 \\ 0.243862643 \\ 0.427530970 \end{matrix}$$

(final result rounded to 9 decimals)

We see that small perturbations (changes or errors) in the sample values have resulted in substantial changes in the estimated parameter values, even though the calculations are exact.

The matrix X may be ill-conditioned because of one or both of the following reasons:

1 One or more columns of X consist of elements all of which are *very nearly* equal to zero.

2 One or more columns of X are *very nearly* obtainable as linear combinations of the remaining columns. In this case we say that **multicollinearity** exists among the columns of X. This is what happens in Example 5.5.2. You may verify that in Example 5.5.2, X1 and X2 are nearly linearly related and that in fact

$$X2 = 19.2106 + 0.82657\ X1,$$

causing the X matrix to be ill-conditioned.

Multicollinearity among the columns of X can occur due to one or more of the following reasons:

a In the population, one or more of the predictor variables X1, . . ., Xk is nearly an exact linear combination of some or all of the remaining predictor variables. For instance, variable Xj may be very nearly an exact linear combination of the remaining predictors so that we have

$$X_j \sim c_0 + c_1X_1 + \bullet\bullet\bullet + c_{j-1}\,X_{j-1} + c_{j+1}\,X_{j+1} + + c_kX_k$$

for every observation. In this situation we say that **multicollinearity exists** among the predictor variables in the population. If sample data are obtained by simple random sampling, then the sample values of the predictor variables also tend to exhibit a relation such as *(5.5.2)*, resulting in multicollinearity among the columns of the X matrix, making it ill-conditioned. When the predictor variables exhibit multicollinearity in the population, then even if the data are obtained by sampling with preselected X values, we are unable to avoid an ill-conditioned X matrix because a relation such as *(5.5.2)* holds for every set of values ($X_1$, . . ., $X_k$) that occurs in the population.

Consider, for instance, a study in which the predictor variables X1, X2, X3, X4, and X5 are heights at ages 4, 5, 6, 7, and 8, respectively, of a population of children, and the response variable Y is height at age 9. In all likelihood, the height at age 8 can be predicted very well using the heights at ages 4, 5, 6, and 7 in a linear prediction function. This means that Xs is very nearly a linear function of X1, X2, X3, and X4 in the population, and no matter how the sample is selected, the sample values of X5 will also be very nearly a linear function of the sample values of X1, X2, X3, and X4.

b Data were obtained by sampling with preselected X values, but practical constraints such as cost, infeasibility of obtaining samples of the response variable at certain combinations of the predictors, etc., may have resulted in a choice of preselected values for the predictors leading to an ill-conditioned X matrix.

c The design of the study is bad. Here investigators could have selected values of the predictor variables in such a way that the X matrix would not be ill-conditioned, but they failed to take advantage of this opportunity.

Presence of *multicollinearity* among the columns of the X matrix has the following implications:

a Computations are very sensitive to rounding, and even if several significant digits are retained during various steps of the calculations, they often yield incorrect values for estimates of various parameters. This can perhaps be overcome by using double precision or multiple precision calculations.

b The results are highly sensitive to errors in the sample data. Even seemingly negligible errors in the measurements can lead to results that have no resemblance to the results that would be obtained if there were no errors in the data. Because practically all measurements are subject to errors, the resulting statistics cannot be taken seriously when the columns of the X matrix exhibit multicollinearity. The standard errors of the parameter estimates may reflect this situation by taking on values that are extremely large relative to the magnitude of the estimates.

c Based on the sample at hand, it is not possible to separate the influences of each of the predictors on the response. This is again related to the fact that the estimated regression coefficients tend to have large standard errors relative to their magnitudes. Whereas we may be able to find good prediction functions, we have to choose arbitrarily from among several sets of nearly equally good prediction functions. Knowledge related to the field of application can often guide us in making a rational selection.