Morris, Max D.; Vardeman, Stephen B.

Working Paper

A sufficient condition related to mistaken intuition about adjusted sums-of-squares in linear regression

Technical Report // Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, No. 2004,04

Provided in cooperation with:
Technische Universität Dortmund

A Sufficient Condition Related to Mistaken Intuition about "Adjusted" Sums-of-Squares in Linear Regression

Max D. Morris and Stephen B. Vardeman*
Departments of Statistics and Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA

January 5, 2004

Abstract

We consider a misconception common among students of statistics involving "adjusted" and "unadjusted" sums-of-squares. While the presence of misconception has been noted before (e.g. Hamilton (1986)), we argue that it may be related to the language we use in describing the meaning of sums-of-squares. For linear regression with two independent variables, we then present a sufficient condition for $SSR(X_1 \mid X_2) > SSR(X_1)$ in terms of the signs of the sample correlations between pairs of predictor and response variables, and note how this sufficient condition may also be related to misconceptions held by some students of statistics.

Introduction

Students of statistics are often struck by the specific technical definitions we assign to words like *bias, sufficient,* and *expected,* which have related but less precise meanings in other contexts. Such terms may help students quickly establish an understanding of, and even intuition for, some basic statistical ideas because their common definitions are usefully suggestive of their statistical meanings. But the use of terms from common language can mislead as well, if we expect too much from such parallels. For example, it is common language in describing statistical regression or analysis of variance to speak of one variable's effect on the response *after adjusting for* another variable, or in explaining the variation *remaining after accounting for effects of* another variable. The colloquial meanings of such phrases might suggest that we expect

* The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") through the University of Dortmund is gratefully acknowledged by the second author.

to see something of reduced magnitude relative to the corresponding quantity before the "adjustment," as with the annual increase in value of a savings account after adjusting for inflation.

This impression is often enforced by the examples we offer in textbooks and the classroom. For example, in their discussion of the "extra" sum-of-squares, Neter, Kutner, Nachtsheim, and Wasserman (1996, section 7.1) describe a study in which amount of body fat (Y) is related to triceps skinfold thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3) in a human physiology study. Based on a sample of 20 subjects, the sum-of-squares for regression of Y on X_1 alone is shown to be $SSR(X_1) = 143.12$, while the extra sum-of-squares associated with X_1 after adjusting for X_2 is $SSR(X_1 | X_2) = 33.17$. The authors carefully (and correctly) say that the second value represents *additional or extra reduction in the error sum-of-squares associated with* X_1*, given that* X_2 *is already included in the model*. However, depending on the words used to further describe this idea, a student may erroneously conclude that $SSR(X_1 | X_2)$ should never be more than $SSR(X_1)$.

Hamilton (1987) pointed out that while most examples in regression textbooks used in the 1980's followed this pattern, adjusted sums-of-squares need not be smaller than their unadjusted counterparts, and offered some geometric insights related to this phenomenon. In a more recent textbook, Mendenhall and Sincich (2003, pp 173-175) present a regression example in which the price of antique grandfather clocks at auction (Y in dollars) is related to the age of the clock (X_1 in years) and the number of bidders present (X_2). In the data set of 32 observations, $SSR(X_1) = 2,555,225$, while $SSR(X_1 | X_2) = 3,533,400$, demonstrating the opposite of what some students might expect. Data from this example are reproduced here in Table 1 for convenience.

Table 1: Data from Grandfather Clock Example of Mendenhall and Sincich

X_1	X_2	Y	X_1	X_2	Y
127	13	1235	170	14	2131
115	12	1080	182	8	1550
127	7	845	162	11	1884
150	9	1522	184	10	2041
156	6	1047	143	6	845
182	11	1979	159	9	1483
156	12	1822	108	14	1055
132	10	1253	175	8	1545
137	9	1297	108	6	729
113	9	946	179	9	1792
137	15	1713	111	15	1175
117	11	1024	187	8	1593
137	8	1147	111	7	785
153	6	1092	115	7	744
117	13	1152	194	5	1356
126	10	1336	168	7	1262

A Sufficient Condition for $SSR(X_1 \mid X_2) > SSR(X_1)$ and Related Intuition

It is instructive to think about the sort of data structures that can lead to adjusted sums-of-squares that are larger than their unadjusted counterparts. For our purposes, consider a linear regression problem with two predictors (one degree of freedom each), corresponding to the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon .$$

Without loss of generality, suppose the response and predictor values have been centered, and the predictors further scaled so that, in vector notation

$$\begin{aligned} \mathbf{Y}'\mathbf{1} &= 0 \\ \mathbf{X}_1'\mathbf{1} &= 0 \\ \mathbf{X}_2'\mathbf{1} &= 0 \\ \mathbf{X}_1'\mathbf{X}_1 &= 1 \\ \mathbf{X}_2'\mathbf{X}_2 &= 1 \end{aligned}$$

and the model may be written without the intercept. For notational convenience, denote the sample correlation coefficients between pairs of variables as:

$$c_{12} = \mathbf{X}_1'\mathbf{X}_2, c_{01} = \mathbf{Y}'\mathbf{X}_1, \text{ and } c_{02} = \mathbf{Y}'\mathbf{X}_2 .$$

We are interested in conditions that lead to

$$SSR\left(X_1 \mid X_2\right) - SSR\left(X_1\right) > 0$$

or

$$\mathbf{Y}'\left(\left(\mathbf{X}_1\mathbf{X}_2\right)\begin{pmatrix} 1 & c_{12} \\ c_{12} & 1 \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} - \mathbf{X}_2\mathbf{X}_2' - \mathbf{X}_1\mathbf{X}_1'\right)\mathbf{Y} > 0 .$$

Noting that we may write the inverse matrix as

$$\begin{pmatrix} 1 & c_{12} \\ c_{12} & 1 \end{pmatrix}^{-1} = \frac{1}{\left(1 - c_{12}^2\right)}\begin{pmatrix} 1 & -c_{12} \\ -c_{12} & 1 \end{pmatrix}$$

the condition can be rewritten as

$$\frac{c_{12}}{\left(1 - c_{12}^2\right)}\left(c_{01}^2 c_{12} + c_{02}^2 c_{12} - 2c_{01}c_{02}\right) > 0 .$$

Because the denominator on the left side is positive, this is equivalent to:

$$\operatorname{sign}\left(c_{12}\right)\left(c_{01}^2 c_{12} + c_{02}^2 c_{12} - 2c_{01}c_{02}\right) > 0$$

that is

$$\left(c_{01}^2 + c_{02}^2\right)c_{12}\operatorname{sign}\left(c_{12}\right) > 2c_{01}c_{02}\operatorname{sign}\left(c_{12}\right).$$

Since $c_{12}\operatorname{sign}\left(c_{12}\right) > 0$ if the two predictors are not orthogonal, this is equivalent to

$$c_{01}^2 + c_{02}^2 > 2c_{01}c_{02} / c_{12} .$$

The above inequality does not have an obvious statistical interpretation, but does provide an interesting sufficient condition for $SSR\left(X_1 \mid X_2\right) - SSR\left(X_1\right) > 0$, because it is satisfied when

$$c_{01}c_{02}c_{12} < 0 .$$

That is, the condition is satisfied by any arrangement of data in which an odd number of correlations between pairs of $Y, X_1,$ and X_2 are negative – the cases listed in Table 2.

Table 2: Signs of correlations satisfying the condition for $SSR(X_1 \mid X_2) > SSR(X_1)$

c_{01}	c_{02}	c_{12}
−	−	−
+	+	−
−	+	+
+	−	+

These situations may also be counterintuitive to beginning statistics students, who may mistakenly interpret the regression equation, excluding the error term, as a linear relationship between any two variables *with the third variable fixed*. For example, for points on the plane described by

$$Y = X_1 - X_2$$

X_1 and X_2 are inversely related given a fixed value of Y. Here one of the three relationships between pairs of variables, given the third, is direct (Y and X_1), while the other two are inverse. In general, either one or all three such relationships must be direct, depending on the signs of the coefficients – these are the patterns absent from Table 2. It may not necessarily be obvious to a beginning student of statistics that, for data modeled as

$$Y = X_1 - X_2 + \varepsilon,$$

c_{12} may be negative, positive, or zero, and that which is the case is not apparent from the coefficients of the fitted model. The example cited above from Mendenhall and Sincich (2003) may help to clarify this; recall that:

$$\begin{aligned} Y &= \text{sale price of the clock (dollars)} \\ X_1 &= \text{age of the clock (years), and} \\ X_2 &= \text{number of bidders participating} \quad . \end{aligned}$$

Here c_{01}, c_{02}, and c_{12} are 0.730, 0.395, and –0.254, respectively, and conform to the pattern displayed in the 2nd line of Table 2. Intuition for the signs of the first two correlations is clear; we might speculate that the correlation between predictors is negative because fewer bidders can afford more expensive clocks.

A Simple Example

A very simple example may help reinforce the idea that adjustment for X_2 may enhance the apparent linear relationship between Y and X_1. The data in Table 3, panel 1, result in equal row means, so that $SSR(X_1) = 0$. But after adjusting for column averages in panel 2, the new row means differ and $SSR(X_1 | X_2) > 0$. The data are also presented graphically in the two panels of Figure 1. In the left panel, the plane denotes the fitted regression of Y on X_1 alone; it must be flat left-to-right because X_2 is not included in the model, but then must also be flat front-to-back (e.g. $SSR(X_1) = 0$) since any other angle would increase the sum of squared distances between data and plane. The plane in the right panel denotes the fitted regression of Y on both X_1 and X_2 together – a perfect fit in this contrived case. Here, because the plane can be tilted left-to-right (e.g. X_2 is included), the best (perfect) fit is achieved when it is also tilted front-to-back (e.g. $SSR(X_1 | X_2) > 0$). Note that in this case, correlations between X_1 and X_2, and Y and X_2, are negative; the correlation between Y and X_1 is exactly zero, but it is clear that small perturbations in the data leading to either positive or negative values would continue to result in $SSR\left(X_1 | X_2\right) > SSR\left(X_1\right)$.

Table 3: Example in which $SSR\left(X_1 | X_2\right) > SSR\left(X_1\right)$

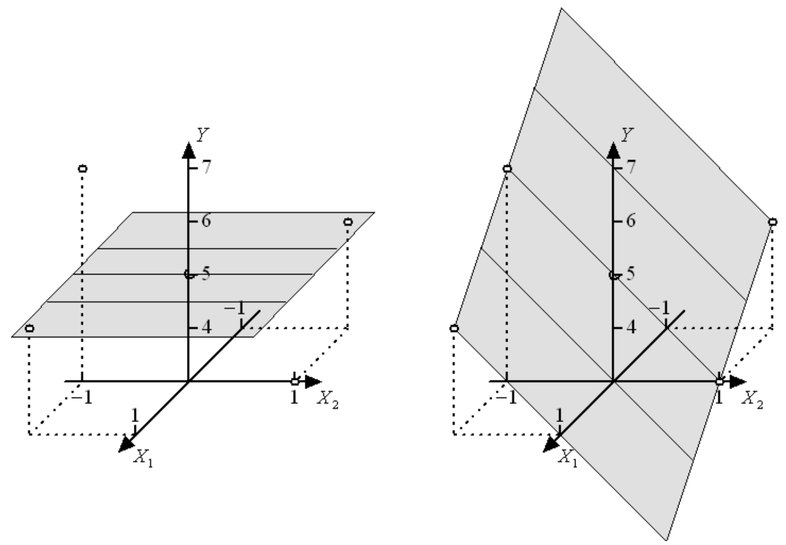
Panel 1: Raw Data

X_1 \ X_2	-1	1	
-1	--	5	5
0	7	3	5
1	5	--	5
	6	4	

Panel 2: Data "Adjusted" for X_2

X_1 \ X_2	-1	1	
-1	--	1	1
0	1	-1	0
1	-1	--	-1
	0	0	

Figure 1: The Case of Table 3



An Adjusted Sum-of-Squares by Any Other Name …

As noted above, our contention is that some misunderstandings about the relationship between unadjusted and adjusted sums-of-squares associated with a variable may begin with intuition accompanying the word *adjusted*. If we are right, it may help to substitute phrases such as "in the presence of other variables" or "ignoring other variables." It may help students to stress that the explanatory value of X_1 may be reduced in the presence of X_2 if the two predictors carry similar information, or it may be increased if the unique predictive value of X_1 is more apparent after the effects associated only with X_2 have been "filtered." Hamilton (1986) points to an example in Kendall and Stewart (1973) in which they call X_2 a "masking variable" when its omission from the model reduces the apparent importance of X_1. Regardless of the words used, we agree with Hamilton that this issue (still) deserves more attention in our classrooms and textbooks. It is important that students be shown examples in which the adjusted sums-of-squares are both larger and smaller than their unadjusted counterparts, and that the fact that both cases are possible be made clear.

Acknowledgement

The authors thank Bob Stephenson for bringing the example from Mendenhall and Sincich to their attention.

References

Hamilton, D. (1987). "Sometimes $R^2 > r_{yx_1}^2 + r_{yx_2}^2$: Correlated Variables Are Not Always Redundant," *The American Statistician*, 41, 129-132.

Kendall, M.G., and Stuart, A. (1973). *The Advanced Theory of Statistics* (Vol. 2, 3rd ed.), New York: Hafner Publishing.

Mendenhall, W., and Sincich, T.L. (1996). *A Second Course in Statistics: Regression Analysis* (5th Edition), Prentice Hall.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models, Fourth Edition*, Chicago: Irwin.