FRIEDMAN, L., AND WALL, M. (2005), "GRAPHICAL VIEWS OF SUPPRESSION AND MULTICOLLINEARITY IN MULTIPLE LINEAR REGRESSION," ***THE AMERICAN STATISTICIAN*****, 59, 127–136: COMMENT BY CHRISTENSEN AND REPLY**

I wish to comment on some geometric aspects of Friedman and Wall (2005). The article has an interesting, and I think unusual, way of looking at suppression and collinearity. In a model $y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, \ldots, n$ we can think of working in n dimensional space with four key vectors, $Y = (y_1, \ldots, y_n)'$, X_1 and X_2 with $X_j = (x_{1j}, \ldots, x_{nj})'$, and $J = (1, \ldots, 1)'$. The key geometrical idea in fitting the model is that the predicted values from the least squares fit are the values in the vector $\hat{Y}$ which is the perpendicular projection of Y onto the space spanned by J, X_1, and X_2. The inner product between two vectors, say Y and X_1, is simply $Y'X_1$. The length of a vector is $\|Y\| \equiv \sqrt{Y'Y}$. Moreover, if θ is the angle between the two vectors, $Y'X_1 = \|Y\| \, \|X_1\| \cos(\theta)$, so the inner product is 0 when θ is 90 degrees, it is positive when θ is less than 90 degrees, and it is negative when θ is greater than 90.

Friedman and Wall are interested in correlations. The sample correlation between, say, y and x_1 is denoted r_{y1} and it is simply the inner product between Y and X_1 after adjusting both vectors so that they (a) have length one and (b) are orthogonal to J. Making everything orthogonal to J simply reduces the dimension of the space from n to $n-1$, so henceforth we will assume that Y, X_1, and X_2 have already been orthogonalized to J and we are working in the $n-1$ dimensional space. In particular, making a vector orthogonal to J turns it into a mean adjusted vector, for example, Y becomes $(y_1 - \bar{y}_\cdot, \ldots, y_n - \bar{y}_\cdot)'$. Another way to think about the correlation is that it is the cosine of the angle between the mean adjusted vectors. To find the partial correlation between y and x_1, say $r_{y1\cdot 2}$, referred to as a standardized regression coefficient in Friedman and Wall, simply find the correlation between Y and X_1 after orthogonalizing them both to X_2. Because most of us have trouble visualizing spaces with more than three dimensions, I will focus on three-dimensional concepts.

Figure 1 gives a simple geometric illustration of classical suppression in two dimensions. The vectors in question are

$$Y = \begin{bmatrix} 0 \\ 1 \\ \epsilon \end{bmatrix}, \quad X_1 = \begin{bmatrix} 1 \\ \delta \\ 0 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{Y} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

(Formally, these are three-dimensional representations of four-dimensional vectors that have been mean adjusted.) R^2 is the squared length of $\hat{Y}$ divided by the squared length of Y. The correlations r_{y1}, r_{y2}, and r_{12}, are just the inner products between the corresponding vectors after adjusting them to have length 1. To get these three-dimensional vectors into a two-dimensional figure, note that the inner product between Y and X_j is identical to the inner product between $\hat{Y}$ and X_j and that none of $\hat{Y}$, X_1, and X_2 have a nonzero component in the third direction. Therefore, in Figure 1 we plot only the first two directions of each vector. Note also that $r_{y1} > r_{y2} = 0$ because $\delta > 0$. In the figure, r_{y1} is close to 0 and r_{12} is close to 1. r_{y1} gets larger and r_{12} gets closer to 0 as δ gets larger and r_{y1} becomes negative if δ is negative. (The same things would happen as δ changes if $X_1 = (\delta, -1, 0)'$ except that $r_{12} < 0$.)

One way to examine collinearity is to let δ, and thus X_1, change. Various models that have the same space spanned by X_1 and X_2 are equivalent in many ways. In particular, they have the same value of $\hat{Y}$ and R^2. As $\delta \neq 0$ changes in Figure 1, the space spanned by X_1 and X_2 remains the same so $\hat{Y}$ and R^2 remain the same. Moreover, we could even change X_2 to be $X_2 = (1, \eta, 0)'$ and nothing important would change in the figure as long as $\eta \neq \delta$. With η small but nonzero, we have r_{y2} small but nonzero with the same sign as η.

The odd thing about Figure 1 is that when ϵ is small, so that Y and $\hat{Y}$ are very nearly the same, we have R^2 very close to 1 but r_{y1} and r_{y2} both very close to 0. That is because, even though X_1 and X_2 are nearly the same vector, the space spanned by the pair of them is much larger and includes vectors very close to Y, namely $\hat{Y}$. I think the real issue is whether you can rely on X_1 and X_2

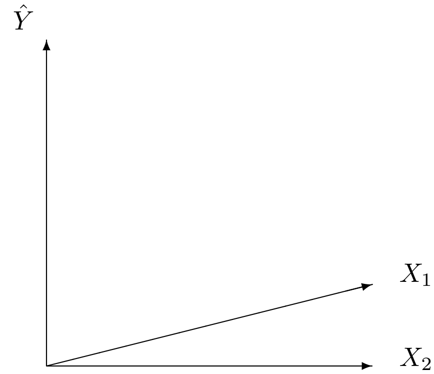

Figure 1: Vectors illustrating classical suppression.

to produce reasonable predictions under the kind of collinearity displayed in Figure 1. Christensen (2002, Sec. 14.4) discussed what can happen if X_1 and X_2 contain small errors.

Rather than keeping the space spanned by X_1 and X_2 fixed, Friedman and Wall look at what happens to r_{12} and R^2 when you keep r_{y1} and r_{y2} fixed. As they did, I assume $r_{y1} > r_{y2} \geq 0$. The set of all X_2 vectors that have r_{y2} fixed, will be a cone centered around the line determined by multiples of Y. For $r_{y2} = 0$ the cone degenerates to the plane through the origin that is orthogonal to Y. Similarly, having r_{y1} fixed determines another cone, closer to the center line than the first.

Imagine taking the vector Y and either a vector X_1 or X_2, say X_1, from its cone. The X_2 vectors that maximize (minimize) r_{12} are the vectors in the X_2 cone that are closest to (farthest from) X_1 in terms of the angle between them. It is easy to see how to find them geometrically. Everything is symmetric about Y so use the two vectors (lines) Y and X_1 to determine a plane. This plane intersects the X_2 cone in two lines. The line that is closer to X_1 contains the vectors in the X_2 cone that maximize r_{12}. The line that is farther from X_1, minimizes r_{12}. (In more than three dimensions, the lines become (hyper)planes.) Conversely, if we take X_1 along with an X_2 that maximizes or minimizes r_{12}, the plane determined by them will contain Y, so $R^2 = 1$.

You can also see how to minimize R^2 geometrically. An alternative way of thinking about the previous paragraph is that to find X_2 that minimizes r_{12} and maximizes R^2, think of starting at the tip of the X_1 vector, traveling straight into the Y line, and continuing on to the X_2 cone. To minimize R^2, start at X_1, travel to the Y line but turn 90 degrees before going on to hit the X_2 cone. More formally, find the (hyper)plane orthogonal to Y and X_1, expand that to a (hyper)plane that also includes Y, and take X_2 to be a vector in the intersection of this expanded hyperplane and the X_2 cone.

Although I am still open to being convinced to the contrary, at the moment I think that to examine the effect of collinearity, the paradigm based on varying X_1 and X_2 (ignoring Y) while keeping fixed the space spanned by X_1 and X_2 is preferable to the paradigm based on varying X_1 and X_2 while fixing r_{y1} and r_{y2} (and implicitly Y). Both are, in some sense, artificial because in reality nothing (or everything) is fixed.

My personal preference in evaluating collinearity is to examine how the space spanned by the mean adjusted versions of X_1 and X_2 changes when they are subject to small errors. In Figure 1, the X_1, X_2 space is the plane determined by $(1, 0, 0)'$ and $(0, 1, 0)'$. When subjected to small errors, the X_1, X_2 space would be a plane determined by one vector that will be very close to (1,0,0) and another vector that could be virtually anything depending on the exact nature of the errors. When the predictors are subject to small errors, the only reliable direction for fitting is a single predictor variable similar to either X_1 or X_2. In Figure 1, if we use only one predictor vector that is similar to X_1 and X_2, we will get an R^2 that is similar to r_{y1}^2 and r_{y2}^2, both of which are near zero. The fact that the two-predictor variable model has a very high R^2 is because the "small errors" have happened to determine a two-dimensional plane that is very close to the Y vector, even though neither X_1 nor X_2 is close to Y. The issue is whether we can believe in our good luck that this happened. Is the good fit of the two variable regression model just a chance occurrence based on the exact nature of the small errors that occurred in X_1 and X_2 or could it be real and reproducible? Short of collecting more data, I know of no way to make such a determination. Nonetheless, it should come as no surprise if the good fit turns out not to be reproducible. More generally, such problems are handled very well by principal component regression in which directions corresponding to small eigenvalues are directions that may be unreliable due to small errors in the predictor variables.

Principal component regression can also be used to provide—perhaps more reliable—estimates of the regression coefficients on the original variables, see, for example, Christensen (1996, sec. 15.6). Personally, I tend to focus on the predictive aspects of linear models and avoid the difficult task of interpreting regression coefficients. In large part I do this because I think it is a short road from interpreting regression coefficients to making the mistake of treating the fitted model as a description of some causal relationship, rather than as simply a description of the data that were collected. Moreover, any predictive ability of the model depends on collecting future data in a manner similar to that used with the analyzed data.

Ronald Christensen

University of New Mexico

REFERENCES

Christensen, R. (1996), *Analysis of Variance, Design, and Regression: Applied Statistical Methods*, Boca Raton, FL: Chapman and Hall/CRC.

—— (2002), *Plane Answers to Complex Questions: The Theory of Linear Models*, New York: Springer.