Submitted to *the Journal of Statistics Education* on May 02, 2013

Visualizating how collinearity affects fitted regression coefficients

James A. Hanley, Zhihui (Amy) Liu, and Ryan Kyle

Department of Epidemiology, Biostatistics and Occupational Health
McGill University
Montreal, Quebec, H3A 1A2, Canada

Email: James.Hanley@McGill.CA

Abstract

The effects of collinearity on the behavior and reliability of the coefficients estimated from a multiple linear regression are an important and challenging topic in regression courses. Textbooks, authors, and teachers have used a variety of methods – algebraic and graphical – to explain these effects. To complement existing efforts, we employ a variety of static tools (tabular display, by algebra, graphical display), and provide an applet created in Excel which allows teachers to illustrate the statistical behaviors associated with collinearity in a dynamic and interactive way.

Key Words: animation; simulation; instability; knife-edge; tightrope walking; Excel

1 Introduction

Textbooks, authors, and teachers use a variety of methods to describe the effects of collinearity on the behavior of the coefficients estimated from a multiple linear regression model. Their aim is to give an intuitive understanding as to why, for example, when two regressor variables are positively correlated, the estimates of the corresponding regression coefficients are negatively correlated, or why the standard errors can be larger than those obtained from two simple linear regressions.

Some take the algebraic approach, while some prefer a geometrical thus more visual, approach. Previously, those who used the latter had to rely on static diagrams, such as those in Swindel (1974), Hocking and Pendleton (1983), and Neter (1996, p289). Although collinearity was not his primary focus, Franklin (1992) used his final dataset (of 4 observations) and the corresponding 3-D figure to produce seemingly contradictory findings when there is a high degree of collinearity.

The features available in `Excel` and in `R` allow teachers to use animation to illustrate the instability and other statistical behaviors associated with collinearity. We use a simple example to show how this can be done. Even if the data are 'generated', and the dataset is too small to yield very precise estimates, we believe it is important that variables have real names – not

just the Y, X_1, X_2 often used in textbooks – and that the research context is genuine.

2 Example

Each of two researchers, interested on the effect of working in a noisy workplace on hearing loss, has a budget to measure hearing loss in only $n = 9$ workers who have been exposed to a noisy work environment for different numbers of years. They use two different sampling schemes. One randomly selects 3 workers aged 45, another 3 aged 55, and another 3 aged 65, in the hope of obtaining a sample with a sufficiently wide spread in the numbers of years worked in a noisy environment. However, because many workers began working at around the same age, the range of these $ages$=(45, 45, 45, 55, 55, 55, 65, 65, 65), should approximate the corresponding range of the numbers of years worked. We call this the "unbalanced" design, and show examples of the joint distribution of these two variables, both measured in years, in the righthand column of Figure 1. The other researcher selects 3 workers from each of the 3 age groups – 1 who has worked 10, another 1 who has worked 20, and 1 who has worked 30 years. That is, age=(45, 45, 45, 55, 55, 55, 65, 65, 65) and $work$=(10, 20, 30, 10, 20, 30, 10, 20, 30). We call this the "balanced" design, and show examples in the lefthand column of Figure 1. The mean age and the mean numbers of years worked are the

same in both designs; the variance in the years worked is similar in both, while the variance in age is identical. Whereas the main concern of Hocking and Pendleton (1983) was prediction, the focus of these researchers will be on isolating the amount of hearing loss 'due to' the time spent working in a noisy workplace. They measure it as decibels per year, i.e., as a regression 'slope'.

Since n is small, the possible estimates depend on the 'luck of the draw'. In practice, a researcher would never know from the sample selected whether the estimate it produced was an over- or under-estimate. In this didactic piece, we use our privileged position to obtain estimates from several simulated samples. We first report the results in a table, before considering some other heuristics.

2.1 Estimates, presented as numbers

We begin with 8 samples that might have arisen from the balanced design. The estimates from these are shown in the leftmost half of Table 1. For each sample, three sets of estimates are reported. Since it is known that hearing loss is a function of age, even for persons who are never exposed to occupational noise, many analysts would use a multiple linear regression involving both *age* and *work* as covariates. The pair of fitted coefficients from this analysis is shown in the first of the three columns. Other analysts might

reason that since the investigator had arranged that the age distribution was the same in those with 10, 20, and 30 years of work, age does not 'confound' the work-hearing loss relationship. Thus, they might consider it more appropriate to report the coefficient from a *simple* linear regression involving only *work*, shown in the second column. Although not the focus of this study, the coefficient from a simple linear regression involving just *age* is shown in the third column for didactic purposes.

From the Table, one can see that no matter which of the 8 possible samples was selected from the balanced $work \times age$ grid, the coefficient for *work* in the multiple regression indicates that those with longer exposure to noisy work have greater hearing loss: the estimated effect is reasonably consistent across the possible samples, and ranges from approximately 0.2 to 0.4 units of hearing loss per year of work. The values of the *age* coefficient are slightly larger, but have a similar spread.

Incidentally, the balanced samples show that, no matter whether *age* is or is not included in the model, one obtains the same estimate for the effect of *work*. What is not well appreciated is that while including it does not make the comparison 'fairer', doing so – in a standard multiple regression – does make it 'sharper' (Hanley, 1883).

We turn now to 8 samples that might have arisen from the unbalanced

design. The corresponding estimates are shown in the rightmost half of Table 1. For this sampling scheme, all analysts would agree that a naive simple regression analysis tends to over-estimate the effect of work, since a comparison among those with approximately 10, 20 and 30 years of occupational exposure is also a comparison between the younger and the older workers – a classic case where age confounds the relationship between exposure and outcome. Thus, they would all fit a multiple linear regression involving both *age* and *work*. The coefficients from this model are shown in the first of the three columns on the right half of the table. The coefficients from a simple linear regression involving *work* alone, and *age* alone, are shown for didactic purposes.

It is readily apparent that the coefficients for *work* in the multiple regression model are far more variable in the imbalanced than in the balanced samples. Some unbalanced samples yielded very large *work* coefficients, while others yielded very small coefficients, even negative ones. The pattern of the eight pairs of numbers in the table tells us that in the multiple regression, if the *work* coefficient from a sample is larger than average, then the *age* coefficient from the same sample tends to be lower than average, and vice versa.

The coefficient for *work* (or *age*) from a simple linear regression is close to the sum of the two coefficients estimated simultaneously from a multi-

ple regression, given that *work* or *age* are positively correlated and are on the same scale. This is not surprising, since, in effect, there is *only one* explanatory variable – experience; it reflects the cumulation of hearing loss caused by both work and non-work exposure. With the exposure variation limited to this one dimension (experience), the task of reliably isolating the separate effects of work and non-work experience from such a small dataset becomes virtually impossible. The same collinearity-curse applies in many dietary studies. For example, if, as one of our colleagues put it, "I like to have sausages with my eggs", then even if Spence et al (2012) had collected data on both, and (as the letters to the journal asked about) on the 'carbs' and other food items that often accompany them, they would have been hard-pressed to reliably isolate their separate effects.

2.2 Estimates: algebraic 'heuristics'

For those who understand best by 'doing the algebra', the unstable behavior in the unbalanced case becomes obvious from the mathematical link between the *work* and *age* variables (in our unbalanced examples, $r_{age,work} = 0.94$). We simulated the relationship between hearing loss and $\{work,\ age\}$ as

$$hearing\ loss \mid age,\ work\ \sim\ N(\mu = \beta_{work} \times work + \beta_{age} \times (age - 25),\ \sigma),$$

where $\beta_{work} = 0.3$, $\beta_{age} = 0.4$, and $\sigma = 2$. In the extreme case, where all subjects started work at age 25, so that $r_{age,work} = 1$, then the expected

hearing loss can be written as either $\{\beta_{work} + \beta_{age}\} \times (age - 25)$ or $\{\beta_{work} + \beta_{age}\} \times work$. The less *age* and *work* are positively correlated, the smaller will be the negative correlation between the estimates β_{work} and β_{age}.

2.3 Estimates: displayed graphically

Figure 1 shows the fitted multiple regressions from four samples of each design, corresponding to the first four rows in Table 1. Each fitted regression can be depicted as a plane, whose gradient in the 'West-East' direction represents the coefficient for *work* and that in the 'South-North' directions represents the *age* coefficient. One quickly notices that the estimates from the four balanced samples are reasonably stable, whereas those from the imbalanced ones are unstable.

The reason becomes clear if one imagines the fitted plane as a *tightrope walker*. If the plane/walker is supported (by data points) at all four corners, its general orientation is not greatly affected by the placement of one point, whereas if is only supported by a long but narrow base/tightrope in the Southwest-Northeast direction, it is quite unstable and likely to be capsized by the slightest individual perturbance at the Southeast or Northwest corner. Hocking and Pendleton (1983) did not give a name to the plane, but to the support for the plane, likening the observed responses in their Figure 1 to

"pickets along a not-so-straight fence row". The task of fitting of the multiple regression equation was thus like "balancing a plane on these pickets".

Despite the narrow base, however, the overall South-West to North-East response gradient can be reliably estimated. This phenomenon is also evident from the last two columns of the table: with data from the imbalanced design, the coefficient from each simple regression is close to the sum of the simultaneously estimated coefficients from a multiple regression.

3 The Excel Applet: animation

Rather than use a table and static figures, we prefer to illustrate the "tightrope walking" in realtime, i.e. interactively and dynamically. We made an applet using an `Excel` spreadsheet. Figures 2 and 3 show the applet, with a switch (0/1) to toggle between the balanced(1) and unbalanced(0) designs. By repeatedly pressing (or holding down) the F9 key, or the equivalent key-combination for 'manual re-calculation' in the MacOS version, the user can observe the sampling distribution of $\{\hat{\beta}_{work},\ \hat{\beta}_{age}\}$, the fitted plane, and the coefficients $\hat{\beta}^*_{work}$ and $\hat{\beta}^*_{age}$ from the two simple regressions.

Both the `Excel` spreadsheet and `R` code, which can be easily modified to suit other examples, are available from the corresponding author's website.

4 Discussion

Some students are more the 'algebra type', so will respond to the 'same-data-different-estimates' story and the accompanying algebra on page 288 of Neter's text. Others, more visual, will prefer the two planes shown on page 289 of the same text.

Some teachers have described collinearity using images such as 'data resting on a knife-edge', or a small (air)plane that crashed and came to rest precariously on a sharp ridge of a mountain, or tightrope walking as we mentioned. Others have tried to be more proximal, and used a large and unwieldy sheet of paper, and imaginary data supports jutting up from the classroom floor, to illustrate the benefits of a wide support for the fitted regression plane. Many are too young to remember Hocking and Pendleton's "picket fence characterization of multi-collinearity". The older of the present authors likens the statistical behaviour to that of a hammock; it reminds the younger authors of the instability encountered when stepping into a canoe or grabbing onto the edge of a life raft.

It is not the purpose of this note to replace existing images and props. Rather, it is to add one more prop, easily built with widely available software, where one can include randomness, and thus impart a better sense of sampling variation in two dimensions. The flexibility and speed of `Excel` or

R allow one to easily animate sampling variation in many other statistical data-analysis contexts.

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, and le Fonds Québécois de la recherche sur la nature et les technologies.

References

Franklin L. A. (1992), "Graphical Insight into Multiple Regression Concepts," *The American Statistician*, 46: 284-288.

Hanley J.A. (1983), "Appropriate uses of multivariate analysis," *Annual Review Public Health*, 4: 155-80.

Hocking R. R. and Pendleton O. J. (1983), "The regression dilemma", *Communications in Statistics - Theory and Methods*, 12:5, 497-527.

Neter J, Kutner M. H., Nachtsheim, C. J., and Wasserman W. (1996), *Applied Linear Statistical Models* (4th ed.) Chicago: Irwin.

Spence J.D., Jenkins D.J., Davignon J. (2012), "Egg yolk consumption and carotid plaque," *Atherosclerosis*, 224(2): 469-73.

Swindel B. F. (1974), "Instability of Regression Coefficients Illustrated," *The American Statistician*, 28: 63-65.

Table 1: Coefficients (units of hearing loss/year) from multiple $\{\hat{\beta}_{work}\ ,\ \hat{\beta}_{age}\}$ and separate simple – $\hat{\beta}^*_{work}$ and $\hat{\beta}^*_{age}$ – linear regression models applied to hearing loss data gathered using balanced and unbalanced designs.

	Balanced				Unbalanced		
sample	$\{\hat{\beta}_{work}\ ,\ \hat{\beta}_{age}\}$	$\hat{\beta}^*_{work}$	$\hat{\beta}^*_{age}$		$\{\hat{\beta}_{work}\ ,\ \hat{\beta}_{age}\}$	$\hat{\beta}^*_{work}$	$\hat{\beta}^*_{age}$
1	0.25 , 0.48	0.25	0.48		0.05 , 0.64	0.66	0.69
2	0.29 , 0.42	0.29	0.42		-0.47 , 1.26	0.74	0.79
3	0.34 , 0.20	0.34	0.20		0.16 , 0.40	0.55	0.56
4	0.20 , 0.48	0.20	0.48		0.68 , 0.20	0.87	0.88
5	0.24 , 0.42	0.24	0.42		0.71 , 0.07	0.78	0.78
6	0.30 , 0.57	0.30	0.57		-0.03 , 0.66	0.60	0.62
7	0.36 , 0.46	0.36	0.46		-0.50 , 1.03	0.49	0.53
8	0.38 , 0.38	0.38	0.38		0.57 , 0.07	0.65	0.65

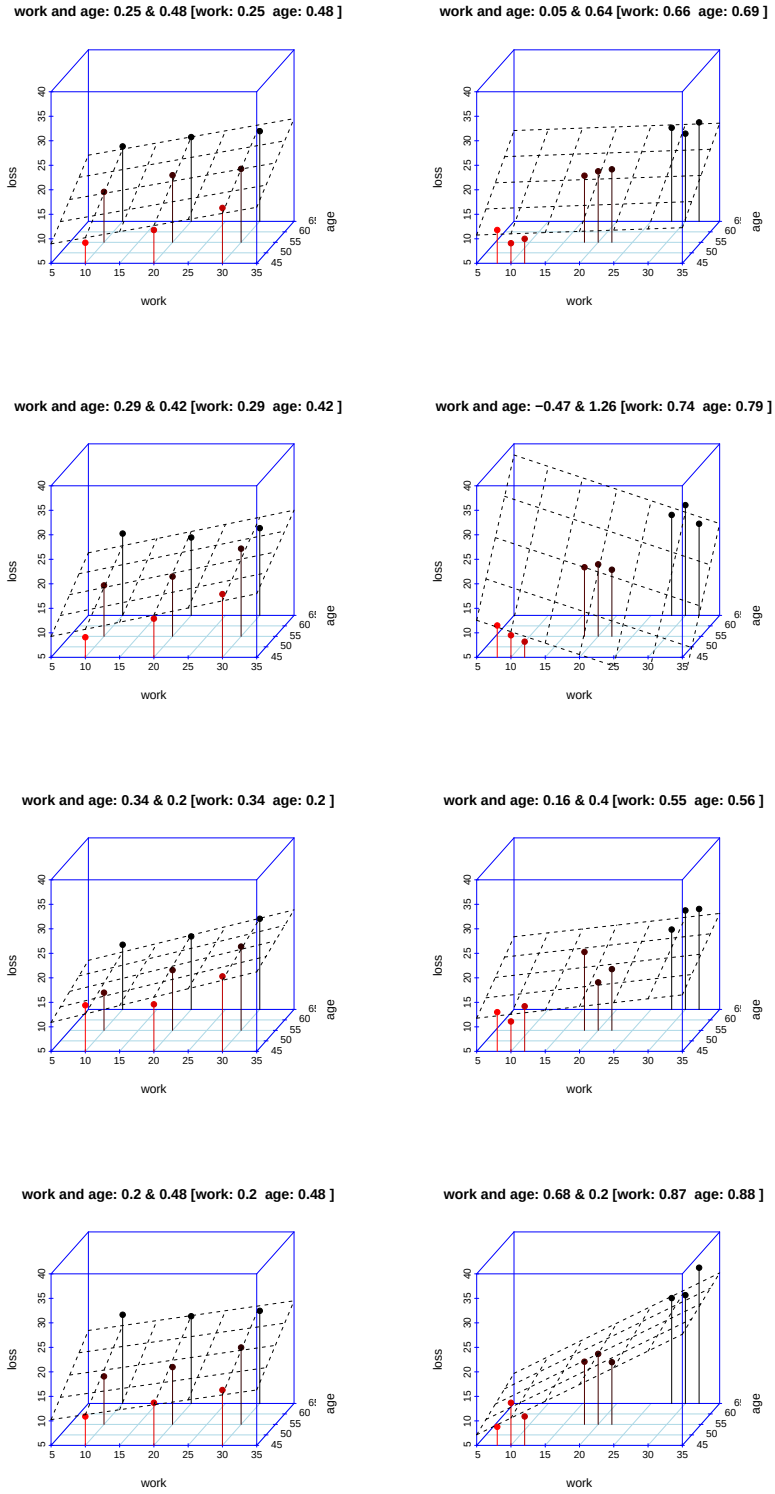

Figure 1: Fitted response surfaces, along with estimates $\{\hat{\beta}_{work}$, $\hat{\beta}_{age}\}$ [and $\hat{\beta}^*_{work}$ and $\hat{\beta}^*_{age}$], from samples with (left) balanced and (right) unbalanced designs. Figure generated by `R`.

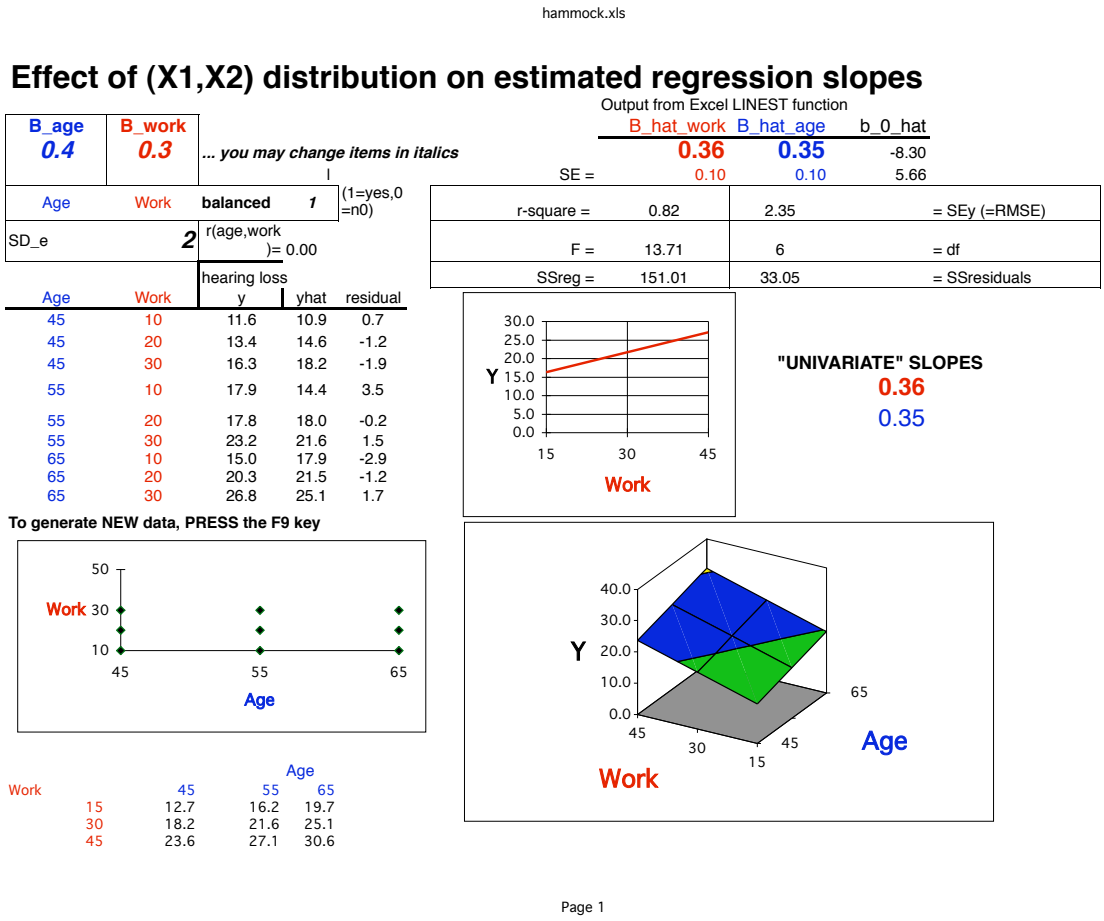

Figure 2: Estimates, $\{\hat{\beta}_{work}$, $\hat{\beta}_{age}\}$ [and $\hat{\beta}^*_{work}$ and $\hat{\beta}^*_{age}$], from a balanced design [`Screenshot from Excel; 2-color plane (response surface) is merely for visual effect.`]

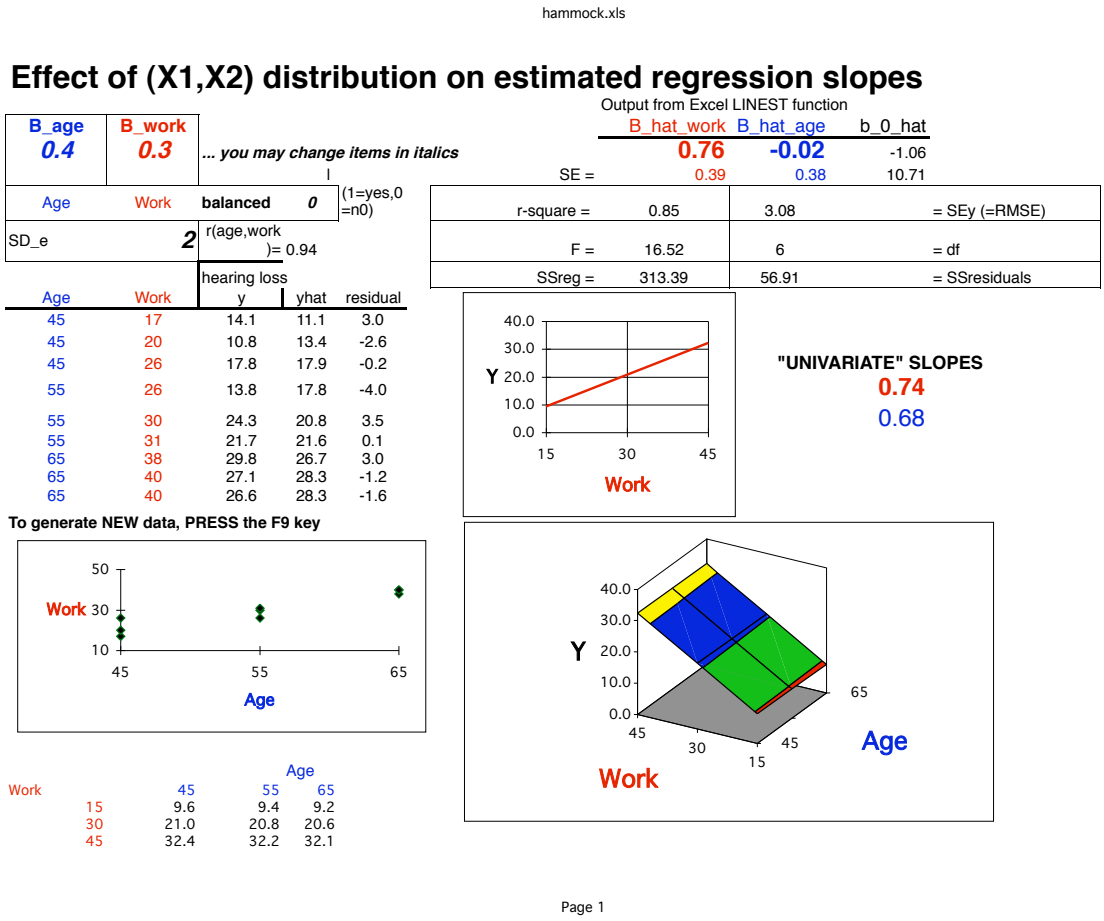

Figure 3: Estimates from an unbalanced design [`Screenshot from Excel. Multi-color plane (response surface) is merely for visual effect.`]