*"How collinearity affects fitted regression coefficients: visualization using a statistical hammock"*

**NOTES, 2025.07.13**

I began using this visualization in the 1990s, when I taught a multiple regression course in our department's summer program. In 2008, I wrote it up and submitted it to The American Statistician, but – as you can read below – the reviewers were underwhelmed. One did point me to to a similar 'prop' (a picket fence) that I had missed.

in 2009, I send it to the Journal of Statistics Education. It got the 'we like the basic idea but we have lots of concerns' reaction, along with a 'we hope you will revise and resubmit'.

I didn't get back to it until 2013, when two students joined me in submitting it to the American Journal of Epidemiology (In my emails from 2013, I see that one of them suggested I should have taken up that invite from JSE, but I was already dealing with JSE on another topic). I can't find reviews from AJE, but I do find evidence that I was subsequently preparing a submission elsewhere, so it looks like AJE was not interested.

I dropped it for a while, but took it up again in 2016 when responding* to an article that promoted a "two subjects per variable" rule of thumb – a rule that I thought was overly-simplistic and dangerous. I thought it was important to "distinguish two of the major uses of regression models that imply very different sample size considerations, neither served well by the rule. The first is etiological research, which contrasts mean Y levels at differing 'exposure' (X) values and thus tends to focus on a single regression coefficient, possibly adjusted for confounders. The second research genre guides clinical practice. It addresses Y levels for individuals with different covariate patterns or 'profiles.' It focuses on the profile-specific (mean) Y levels themselves, estimating them via linear compounds of regression coefficients and covariates."

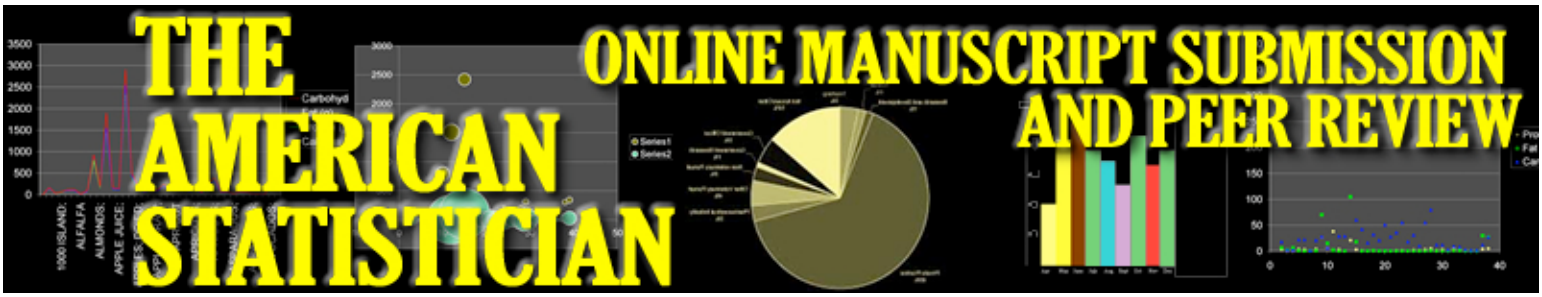
I made extensive use of the 'hammock' in section 3, dealing with etiological research, where it nicely illustrates the sample size cost of adjusting for confounding.

To me, the feeling of motion-sickness induced by watching realizations in the **simple spreadsheet example** was more effective that any mathematical statistics explanation. The spreadsheet (along with a simple `R` implementation) can be found at `https://jhanley.biostat.mcgill.ca/software/`

Sincerely,

**James Hanley**
webpage: https://jhanley.biostat.mcgill.ca | email: `james.hanley@mcgill.ca`

* Hanley JA. Simple and multiple linear regression: sample size considerations. J Clin Epidemiol. 2016:




HOME **Saturday, May 17, 2008**

TO ENSURE PROPER FUNCTIONALITY OF THIS SITE, BOTH JAVASCRIPT AND COOKIES MUST BE ENABLED.

## Detailed Status Information

| | |
|---|---|
| **Manuscript #** | MS08-102 |
| **Current Revision #** | 0 |
| **Submission Date** | 2008-05-17 18:27:29 |
| **Current Stage** | Initial QC Started |
| **Title** | How collinearity affects fitted regression coefficients: visualization using a statistical "hammock" |
| **Running Title** | statistical "hammock" |
| **Manuscript Type** | Teacher's Corner |
| **Special Section** | N/A |
| **Manuscript Comment** | For Teachers Corner.. I hope you are able to share the R code as a text file, and the Excel file as an Excel file, with reviewers. If reviewers would like originals of these 2 files, please contact me. |
| **Corresponding Author** | Dr. James Hanley (McGill University) |
| **Contributing Author** | N/A |
| **Abstract** | The effects of colinearity on the behavior and reliability of the coefficients estimated from a multiple linear regression are an important and challenging topic in multiple regression courses. Textbooks, authors, and teachers have used a variety of methods -- algebraic and graphical -- to explain these effects. The random-number and graphics features now available in Excel and in R allow teachers and students to use animation to visualize the statistical behaviors associated with colinearity. We use a simple example to show how easily this can be done. |
| **Associate Editor** | Not Assigned |
| **Keywords** | collinearity, knife-edge, support, animation, instability |
| **Manuscript Topic** | Experimental Design, Graphical Methods, Regression: Linear |

| | |
|---|---|
| **Copyright Release Date** | Not Received |

| Stage | Start Date |
|---|---|
| Editor Assigned | 2008-05-17 19:24:14 |
| Author Approved Converted Files | 2008-05-17 19:24:13 |
| Preliminary Manuscript Data Submitted | 2008-05-17 18:27:29 |

For assistance, please contact the editorial coordinator at american.statistician@gmail.com

Submitted to *The American Statistician*, May 17, 2008

# How collinearity affects fitted regression coefficients: visualization using a statistical "hammock"

James A. Hanley

James A. Hanley is Professor, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, H3A 1A2, Canada (email: James.Hanley@McGill.CA). 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.

**Abstract**

The effects of colinearity on the behavior and reliability of the coefficients estimated from a multiple linear regression are an important and challenging topic in multiple regression courses. Textbooks, authors, and teachers have used a variety of methods – algebraic and graphical – to explain these effects. The random-number and graphics features now available in `Excel` and in `R` allow teachers and students to use animation to visualize the statistical behaviors associated with colinearity. We use a simple example to show how easily this can be done.

KEY WORDS: collinearity; knife-edge; support; animation; instability.

# 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 the two simple linear regressions.

Some take the algebraic approach, while some prefer a geometrical, and thus more visual, approach. Previously, those who used the latter had to rely on static diagrams, such as those in Swindel(1974) 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 high degree of collinearity.

The features now available in spreadsheets and in `R` allow teachers and students to use animation to visualize the instability and other statistical behaviors associated with collinearity. We use a simple example to show how easily this can be done.

# 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 $n = 9$ 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 wide spread in the numbers of years worked in a noisy environment. The other also uses these three ages as the source, but uses work records to randomly select from *each* of these 3 sources 1 who has worked 10, another 1 who has worked 20, and 1 who has worked 30 years in this environment. The distributions of the two samples with respect to *age* and *work*, both measured in years, are shown in the Figures. The mean age and the mean numbers of years worked are the same in both the "unbalanced" and "balanced" designs; the variance in the years worked is very similar in both, while the variance in age is identical.

## 2.1 Estimates from two designs: tabular display

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 and under-estimate, i.e., whether

the 'single shot' fell above or below the target. In this didactic piece, we use our privileged position to produce estimates from several 'what might have been' samples. We will continue the imagery of statistical shots at a target: one can think of a statistical estimate, such as the fitted regression coefficient(s) derived from a sample, as an arrow shot at a target which is visible just briefly before the arrow is released; one can see where the arrow struck, but not where the target was.

We begin with 10 samples that might have been selected by the "*balanced*" researcher. The estimates from these are shown in the leftmost half of Table 1. For each sample, three sets of analyses/estimates are reported: First, since it is known that hearing loss is a function of age, even in those who were never exposed to occupational noise, many analysts would use as their estimate the regression coefficient for the years of work variable ('*work*') in a *multiple* linear regression involving *work* and *age*. 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 did not 'confound' the work-hearing loss relationship. Thus they would use as the appropriate estimate the coefficient from a *simple* linear regression involving only *work*, shown in the second column. Although not the focus of the study, the coefficient from a simple linear regression involving just *age* is

shown for didactic purposes.

From the Table, one can see that no matter which of the 10 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.

From the analysis of the balanced sample, the investigators who argue for not including *age* in the model can say that 'they told us so': they obtain the same estimates for *work* from the simple linear regression as those who estimated the coefficient for *work* from a multiple regression model that included *age*.

We turn now to 10 samples that might have been selected by the researcher who selected a representative but '*unbalanced*' sample. 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 of those with approximately 10, 20 and 30 years of occupational exposure is also a comparison of those who are younger and those who are older— a classic

case where age 'confounds' the true relationship between the exposure and the 'health measurement' of interest. 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 regressions involving *work* alone, and *age* alone, are shown for didactic purposes.

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, as much as 1 unit of hearing loss per year of age, while others yielded very small coefficients, even some that were negative.

Our privileged position allows us to see something that we could not know in practice with a single 'shot': the pattern of the ten pairs of numbers in the table, just like the positions of the ten arrows, tell us that in the multiple regression, if the *work* coefficient from a sample is larger than average, the *age* coefficient from the same sample tends to be lower than average, and vice versa.

The last two columns also show us what *can* be estimated reliably from the imbalanced design: the coefficient for each variable alone seems to be close to the sum of the coefficients for *age* and *work* (estimated simultaneously

from the balanced, or even the imbalanced, design). This is not all that surprising, since, in effect, there is *only one* variable, 'experience;' it reflects the cumulation of hearing loss caused by both work and non-work exposures. With the exposure variation limited to this one 'experience' dimension, the task of reliably isolating the separate effects of work and non-work experience from such a small dataset becomes virtually impossible.

## 2.2 Estimates from two designs: heuristics, by algebra

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

$$loss \mid age\ work\ \sim\ N(\mu = \beta_{work} \times work + \beta_{age} \times (age - 25),\ \sigma).^{1}$$

In the extreme case, where say all subjects started work at age 18, so that $r_{age,work} = 1$, then, apart from some constants, the expected hearing loss can be written as either $\{\beta_{work} + \beta_{age}\} \times age$, or $\{\beta_{work} + \beta_{age}\} \times work$. Clearly, any other pair of values $\{\beta_{work} - \Delta,\ \beta_{age} + \Delta\}$ will also give this same relationship. The less *age* and *work* are linked, the smaller will be the (negative) correlation between the estimates of $\beta_{work}$ and $\beta_{age}$.

[1] As shown in Figures 2 and 3, we used $\beta_{work} = 0.3$, $\beta_{age} = 0.4$, and $\sigma = 2$.

## 2.3 Estimates from two designs: graphical display

Figure 1 shows the fitted multiple regressions from four samples from each sampling scheme. 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 coefficient for *age*. 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 considers the fitted plane as a *statistical hammock*[2]. If the hammock is anchored (has supporting data) at all four corners, its general orientation is not greatly affected by the placement of any one individual, whereas if is only supported by a long but narrow base, it is quite unstable and likely to be capsized by the slightest individual perturbance.

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 for *age* and *work*.

[2]If one allows a small statistical 'licence' to make one end higher than the other.

# 3 THE STATISTICAL HAMMOCK, ANIMATED

Rather than use a table and figures showing what students might suspect are selected examples, it would be preferable to illustrate these in class in real-time, i.e., dynamically. Fortunately, this is very easy to do using an `Excel` spreadsheet or a simple function in `R`. Figures 2 and 3 show the Excel sheet, with a switch to toggle between the balanced and unbalanced designs. By repeatedly pressing (or holding down) the 'recalculate' keys, 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. The `Excel` and `R` files, which can easily be modified to suit other examples, are available from the author's website.

# 4 DISCUSSION

Some students are more the 'algebra type,' and so will respond to the 'same data, different estimates' story, and 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.

The author – and, I expect, several other 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. Maybe, like I, authors 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.

It is not the purpose of this note to replace these 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 2-dimensions. The flexibility and speed of Excel or R can, of course, also be used to animate sampling variation in many other statistical contexts.

# 5 REFERENCES


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

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

- 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 | 0.26 , 0.34 | 0.26 | 0.34 | | 0.27 , 0.31 | 0.57 | 0.58 |
| 2 | 0.33 , 0.32 | 0.33 | 0.32 | | 0.19 , 0.57 | 0.74 | 0.76 |
| 3 | 0.24 , 0.54 | 0.24 | 0.54 | | 0.32 , 0.26 | 0.58 | 0.59 |
| 4 | 0.32 , 0.31 | 0.32 | 0.31 | | 0.52 , 0.08 | 0.60 | 0.60 |
| 5 | 0.24 , 0.42 | 0.24 | 0.42 | | 0.71 , 0.07 | 0.78 | 0.78 |
| 6 | 0.20 , 0.48 | 0.20 | 0.48 | | 0.35 , 0.38 | 0.71 | 0.73 |
| 7 | 0.30 , 0.57 | 0.30 | 0.57 | | -0.03 , 0.66 | 0.60 | 0.62 |
| 8 | 0.29 , 0.44 | 0.29 | 0.44 | | 0.79, -0.07 | 0.72 | 0.72 |
| 9 | 0.36 , 0.46 | 0.36 | 0.46 | | -0.50 , 1.03 | 0.49 | 0.53 |
| 10 | 0.38 , 0.38 | 0.38 | 0.38 | | 0.57 , 0.07 | 0.65 | 0.65 |

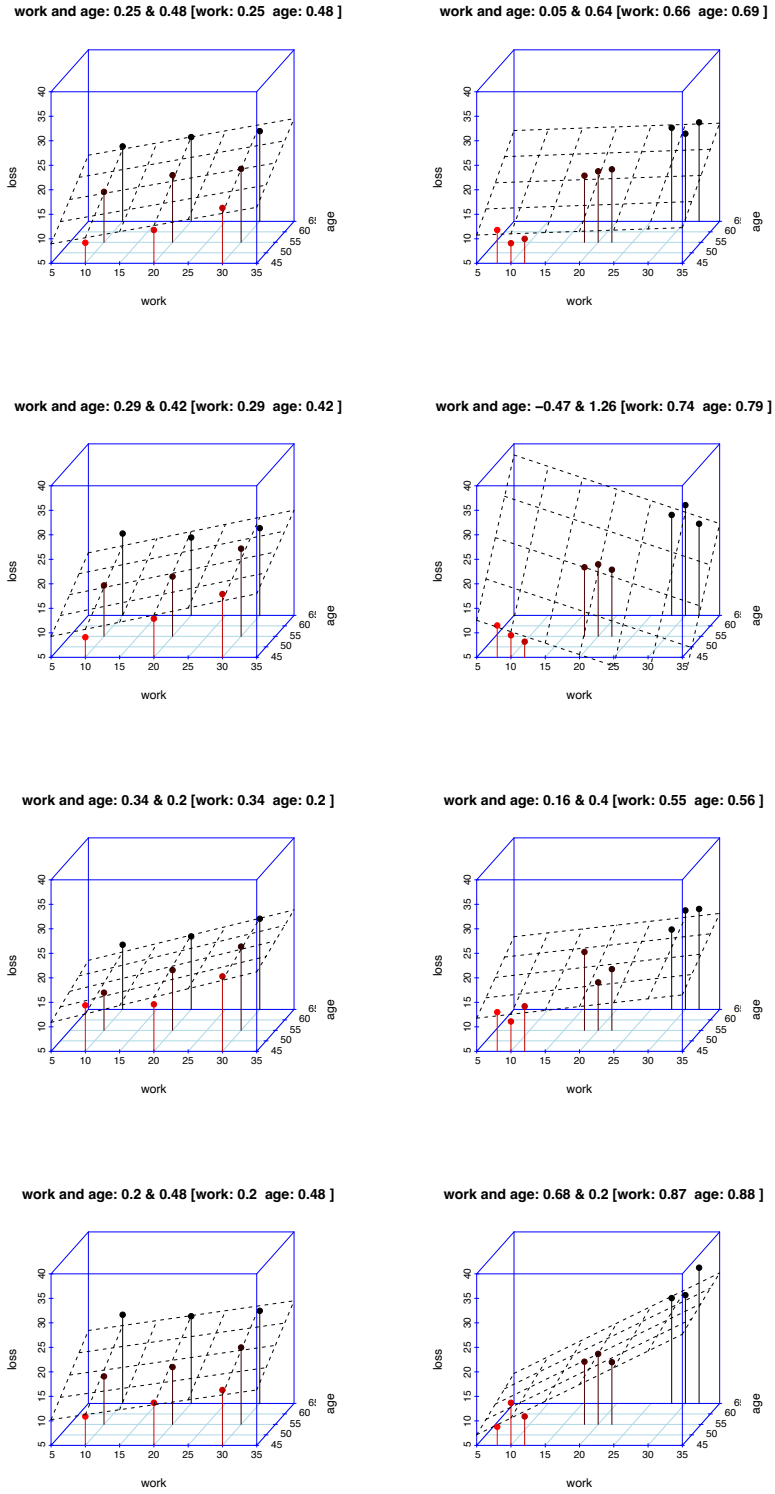





Figure 1: Estimates from samples with Balanced and Unbalanced designs [R]

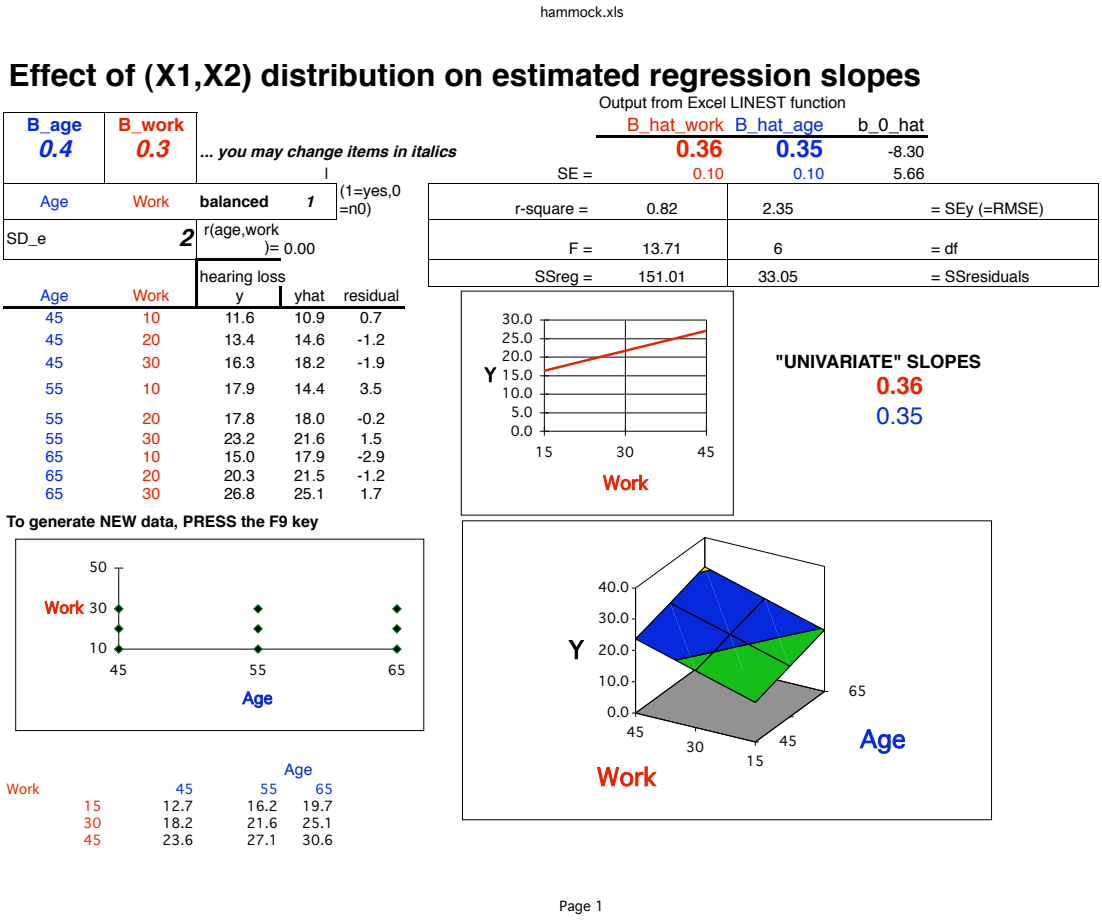



Figure 2: Estimates from a sample: Balanced design [`Excel`]

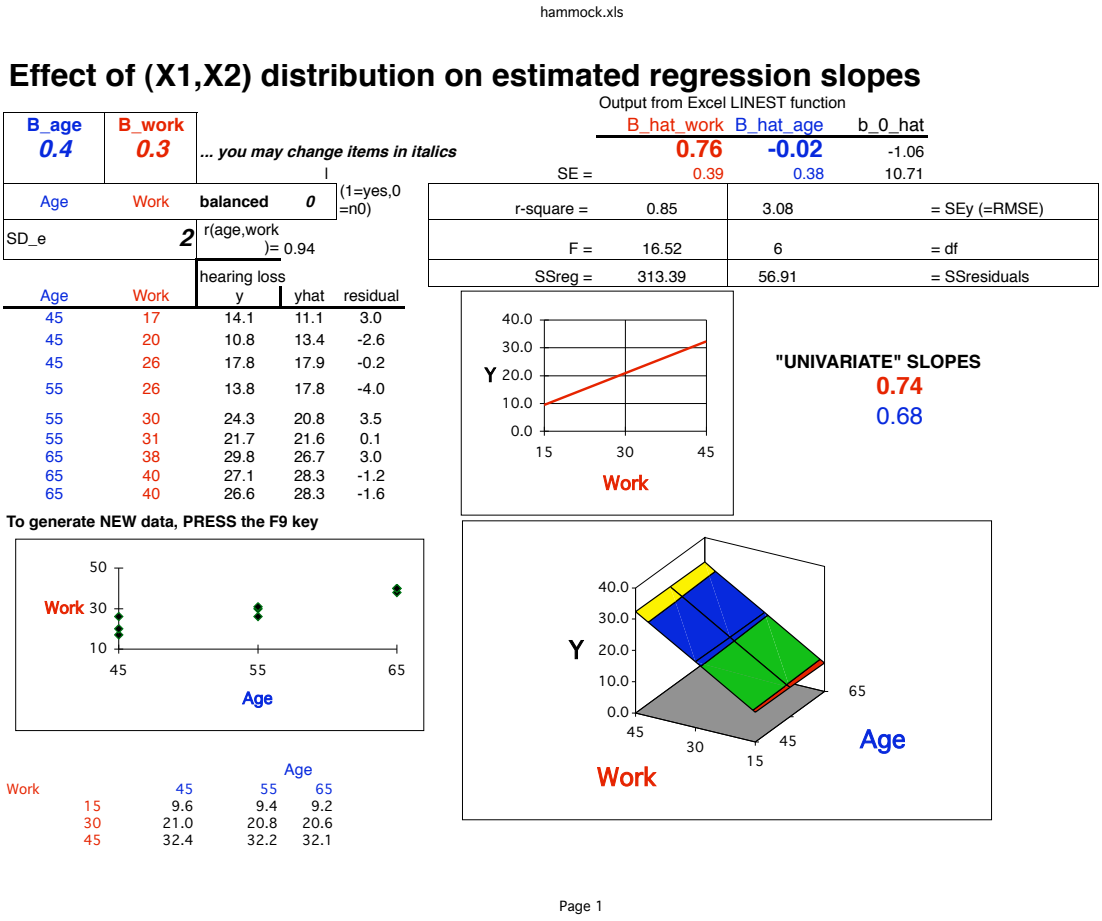



Figure 3: Estimates from a sample: Unbalanced design [`Excel`]

Dear Dr. Hanley,

Thank you for submitting the above manuscript "How collinearity affects fitted regression coefficients: visualization using a statistical "hammock"" by James Hanley for possible publication in The American Statistician (TAS). TAS receives many papers and can publish only a fraction. Based on my reading and on that of an Associate Editor (AE), I have concluded that the paper is not appropriate for TAS.

In addition to the comments provided in the attached pdf file, the AE writes to me personally,

==========AE comments====
I don't think that this paper is worthy of publication in TAS without substantial additional material. The current paper is basically a simple simulation that could be done in class to demonstrate the effects of collinearity on regression coefficients.

You might suggest another journal as an alternative destination for the paper. However, I believe that my suggested changes would be necessary even for consideration elsewhere. There just isn't a lot of interesting stuff in this paper. I have included more specific comments in my comments to the author.

===End AE comments=====

My editorial comments are as follows:

The "Hammock" paper promotes an interactive 3-D visual display to understand how the sampling variation of the estimated regression coefficients is affected by multicollinearity.

As far as the paper goes, it is fine, although it seems a little too simplistic for a TAS paper. While it is true we aim for a broad audience, the main benefit of the paper seems to be to promote a couple of hand-made graphical tools, and the contribution is therefore of limited value.

The term "hammock" itself, while cute, seems somewhat misleading in that the term describes a curved rather than planar surface.

The example is actually very good, but the paper seems shallow otherwise. The references are skimpy; in particular, Hocking's famous "picket fence" visual is not even mentioned.

Rather than string this paper along, I have decided to reject it, in order to encourage a fresh, more scholarly approach.

Whatever you decide to do with the paper, I do warmly wish you the best in finding it a proper publication venue. Apologies for unfortunate news during the holidays, which I hope are otherwise happy. I am trying to clear my desk for the incoming editor, who will take over Jan 1.

Sincerely,

Peter Westfall
Editor, The American Statistician

**MS08-102 Response to Author**

**General comments**

- You should consider who the audience is for this paper. The tone of the paper shifts between between "talking to students" and "talking to instructors". You need to talk more directly to instructors and give tips about how to run this simulation in a way that provides the most effective instruction. You do this much more effectively in the latter part of the paper, but it needs to be done throughout.
- The paper provides simulations from individual samples without adequate summary measures to describe what the reader is seeing. For example, in Table 1 you show only the coefficients for the 10 sample regressions in both the balanced and unbalanced designs. You need to include standard errors of the coefficients, t-statistics, the mean of all the sample coefficients, the standard deviation of all sample coefficient. You also need to tell what the parameter value is.
- You need to provide simulations for more examples of unbalanced designs. Another interesting question is how does the level of collinearity affect the regression coefficients. The paper would be more publishable if the simulations investigated this question.
- A big problem with collinearity is the inflation of the standard errors of the coefficients. This is not demonstrated effectively in the paper. The lack of summary measures for the simulations makes it easy to miss this point. You should focus more on the inflation of the standard errors—especially in the algebraic section where there is no discussion of this problem.
- Use Figure and Table numbers throughout the text of the paper. These are missing in this version.
- I believe that *collinearity* is the more traditional spelling of the term, not *colinearity*.

**Page specific comments**

**p. 2**

- Include a table listing the design points in the simulation and indicate which are balanced and which are not.
- In line 1 of Section 2.1, replace "depend on the luck of the draw" with "are highly variable". All estimates depend on the luck of the draw regardless of sample size, those with small *n* are just more variable.

**p. 3**

- The example of the shots at a target is a good metaphor. You should specifically state what the bulls-eye is "the true parameter value $\beta$".

**p. 5**

- Show a graph of the ten estimate pairs, and compute the correlation.

**p. 6**

- In Section 2.2, show

$$E(y) = \beta_0 + \beta_w x_w + \beta_a x_a$$

$$\text{If } x_w = x_a,$$

$$E(y) = \beta_0 + (\beta_w + \beta_a) x_a$$

$$= \beta_0 + \beta_w^* x_a$$

- You state "The less *age* and *work* are linked, the smaller will be the negative) correlation between the estimates of $\beta_{work}$ and $\beta_{age}$." You should include a formula showing this.

**p. 7**

- I'm not sure that the "hammock" analog is the best one for students. Hammocks sag; a plane cannot sag. The hammock analog might inadvertently imply that the plane can be deformed into a nonlinear shape. In the absence of transformed data or an interaction, this is not true.

**p. 8**

- To understand what is going on with $\hat{\beta}_w + \hat{\beta}_a$, $\hat{\beta}_w^*$, and $\hat{\beta}_a^*$ in Table 1, you need more information in the table on all of these quantities. For example, what are the true $\beta$ values for each? Show a histogram of the $\hat{\beta}_w^*$'s and $\hat{\beta}_a^*$'s. What are the standard errors of these estimates? What are their *t*-values? What is the average value of the $\hat{\beta}_w^*$'s (or $\hat{\beta}_a^*$'s) for all 10 simulations.

**p. 9**

- The discussion of the simulation notes the increased variability of the regression coefficients in the unbalanced case—not enough, but it s at least mentioned. However, the algebraic approach talks only about the fact that $\hat{\beta}_w^* \approx \hat{\beta}_w + \hat{\beta}_a$. It doesn't discuss the increased variability of the estimates at all. I think this is a major oversight, since the increased variability of the estimates is a key result of collinearity.

# *Journal of Statistics Education*

Pre-Review Author Disclosure Form

The corresponding author must complete and sign a copy of this form and submit it to the Editor before a paper will be sent out for review. If the paper is accepted for publication, all authors will be asked to sign a second, similar form before the paper is scheduled for publication. Please note that a conflict or apparent conflict of interest will not necessarily affect decisions on acceptance or publication of a paper, but full disclosure is required.

**How collinearity affects fitted regression coefficients: visualization using a statistical "hammock"**

Name of the corresponding author (please print or type) **James Hanley**

Please check all that apply and include attachments, if necessary.

☐ √ To the best of my knowledge, all authors listed on this paper have participated materially in the research and work presented in this paper and we give our approval of this work.

☐ **N/A** To the best of my knowledge, all parties who have participated materially in the research and work presented in this paper are listed as co-authors. If this box is not checked, you must attach a statement giving the names of any individuals who have participated materially in the research and work and the reasons that they are not listed as co-authors.

☐ √ To the best of my knowledge, the work reported in this paper has not appeared previously in a refereed or copyrighted publication nor is it currently under review with another journal. Also, it will not be submitted elsewhere before notifying the editor that the paper is being withdrawn.

☐ **N/A** As the corresponding author for this paper, I will keep my co-authors informed about communications with the editorial office concerning review and publications matters.

☐ **N/A** I certify that, to the best of my knowledge, all external sources of financial and/or material support for the research in this paper (i.e., sources other than from the employer(s) of the author(s)) have been clearly stated in the acknowledgements section of the paper. (By ASA policy, acknowledgements are to be blinded for referees, but available to the editors).

Check one of the following:

☐ √ The authors of this paper have no affiliations or financial involvements (e.g., stock ownership; employment; consultancies; past or expected expert testimony; past, pending, or anticipated financial aid or patent applications) that could potentially constitute a conflict of interest or an apparent conflict of interest with the research or work reported in this paper.

or

☐ One or more of the authors have affiliations or financial involvements that could potentially constitute a conflict of interest or an apparent conflict of interest with the research or work reported in this paper. I certify that all such affiliations and/or financial involvements have been completely disclosed in the attached statement(s).

**James A Hanley**

Your name (please print or type)

July 22, 2009

Your signature Date Signed

Submitted to *Journal of Statistics Education*, July 22, 2009

# How collinearity affects fitted regression coefficients: visualization using a statistical "hammock"

James A. Hanley

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

email: James.Hanley@McGill.CA

## Abstract

The effects of colinearity on the behavior and reliability of the coefficients estimated from a multiple linear regression are an important and challenging topic in multiple regression courses. Textbooks, authors, and teachers have used a variety of methods – algebraic and graphical – to explain these effects. The random-number and graphics features now available in `Excel` and in `R` allow teachers and students to use animation to visualize the statistical behaviors associated with colinearity. We use a simple example to show how easily this can be done.

KEY WORDS: collinearity; knife-edge; support; animation; instability.

# 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 the two simple linear regressions.

Some take the algebraic approach, while some prefer a geometrical, and 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 high degree of collinearity.

The features now available in spreadsheets and in `R` allow teachers and students to use animation to visualize the instability and other statistical behaviors associated with collinearity. We use a simple example to show how easily this can be done. Even if the data are 'generated', we believe it is important that variables have "real" names – not just the $Y, X_1, X_2$ often

used in articles and books – 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 $n = 9$ 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 wide spread in the numbers of years worked in a noisy environment. The other also uses these three ages as the source, but uses work records to randomly select from *each* of these 3 sources 1 who has worked 10, another 1 who has worked 20, and 1 who has worked 30 years in this environment. The distributions of the two samples with respect to *age* and *work*, both measured in years, are shown in the Figures. The mean age and the mean numbers of years worked are the same in both the "unbalanced" and "balanced" designs; the variance in the years worked is very similar in both, while the variance in age is identical.

Whereas the main concern of Hocking and Pendleton (1983) was prediction, our focus will be on isolating the effect of working in a noisy workplace.

## 2.1 Estimates from two designs: tabular display

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 and under-estimate, i.e., whether the 'single shot' fell above or below the target. In this didactic piece, we use our privileged position to produce estimates from several 'what might have been' samples. We will continue the imagery of statistical shots at a target: one can think of a statistical estimate, such as the fitted regression coefficient(s) derived from a sample, as an arrow shot at a target which is visible just briefly before the arrow is released; one can see where the arrow struck, but not where the target was.

We begin with 10 samples that might have been selected by the "*balanced*" researcher. The estimates from these are shown in the leftmost half of Table 1. For each sample, three sets of analyses/estimates are reported: First, since it is known that hearing loss is a function of age, even in those who were never exposed to occupational noise, many analysts would use as their estimate the regression coefficient for the years of work variable ('*work*') in a *multiple* linear regression involving *work* and *age*. 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 did not 'confound' the work-hearing loss relationship. Thus they would use as the appropriate estimate the coefficient from a *simple* linear regression involving only *work*, shown in the second column. Although not the focus of the study, the coefficient from a simple linear regression involving just *age* is shown for didactic purposes.

From the Table, one can see that no matter which of the 10 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.

From the analysis of the balanced sample, the investigators who argue for not including *age* in the model can say that 'they told us so': they obtain the same estimates for *work* from the simple linear regression as those who estimated the coefficient for *work* from a multiple regression model that included *age*.

We turn now to 10 samples that might have been selected by the researcher who selected a representative but '*unbalanced*' sample. The cor-

responding 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 of those with approximately 10, 20 and 30 years of occupational exposure is also a comparison of those who are younger and those who are older— a classic case where age 'confounds' the true relationship between the exposure and the 'health measurement' of interest. 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 regressions involving *work* alone, and *age* alone, are shown for didactic purposes.

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, as much as 1 unit of hearing loss per year of age, while others yielded very small coefficients, even some that were negative.

Our privileged position allows us to see something that we could not know in practice with a single 'shot': the pattern of the ten pairs of numbers in the table, just like the positions of the ten arrows, tell us that in the multiple regression, if the *work* coefficient from a sample is larger than average, the *age* coefficient from the same sample tends to be lower than average, and

vice versa.

The last two columns also show us what *can* be estimated reliably from the imbalanced design: the coefficient for each variable alone seems to be close to the sum of the coefficients for *age* and *work* (estimated simultaneously from the balanced, or even the imbalanced, design). This is not all that surprising, since, in effect, there is *only one* variable, 'experience;' it reflects the cumulation of hearing loss caused by both work and non-work exposures. With the exposure variation limited to this one 'experience' dimension, the task of reliably isolating the separate effects of work and non-work experience from such a small dataset becomes virtually impossible.

## 2.2 Estimates from two designs: heuristics, by algebra

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

$$loss \mid age\ work\ \sim\ N(\mu = \beta_{work} \times work + \beta_{age} \times (age - 25),\ \sigma).^{1}$$

[1] As shown in Figures 2 and 3, we used $\beta_{work} = 0.3$, $\beta_{age} = 0.4$, and $\sigma = 2$.

In the extreme case, where say all subjects started work at age 18, so that $r_{age,work} = 1$, then, apart from some constants, the expected hearing loss can be written as either $\{\beta_{work} + \beta_{age}\} \times age$, or $\{\beta_{work} + \beta_{age}\} \times work$. Clearly, any other pair of values $\{\beta_{work} - \Delta,\ \beta_{age} + \Delta\}$ will also give this same relationship. The less $age$ and $work$ are linked, the smaller will be the (negative) correlation between the estimates of $\beta_{work}$ and $\beta_{age}$.

## 2.3 Estimates from two designs: graphical display

Figure 1 shows the fitted multiple regressions from four samples from each sampling scheme. 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 coefficient for $age$. 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 considers the fitted plane as a *statistical hammock*[2] If the hammock is anchored (has supporting data) at all four corners, its general orientation is not greatly affected by the placement of any one individual, whereas if is only supported by a long but narrow base, it is quite unstable and likely to be capsized by the slightest individual perturbance. Hocking and Pendleton (1983) didn't give a name to the plane, but did to

[2]If one allows a small statistical 'licence' to make one end higher than the other.

the support for the plane, likening the observed $Y$s ("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 for *age* and *work*.

# 3 THE STATISTICAL HAMMOCK, ANIMATED

Rather than use a table and figures showing what students might suspect are selected examples, it would be preferable to illustrate these in class in real-time, i.e., dynamically. Fortunately, this is very easy to do using an `Excel` spreadsheet or a simple function in `R`. Figures 2 and 3 show the Excel sheet, with a switch to toggle between the balanced and unbalanced designs. By repeatedly pressing (or holding down) the 'recalculate' keys, 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. The `Excel` and `R`

files, which can easily be modified to suit other examples, are available from the author's website.

# 4 DISCUSSION

Some students are more the 'algebra type,' and so will respond to the 'same data, different estimates' story, and 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.

The author – and, I expect, several other 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. Maybe, like I, authors 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."

It is not the purpose of this note to replace these 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 2-dimensions. The flexibility and speed of Excel or R

can, of course, also be used to animate sampling variation in many other statistical contexts.

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.

# 5 REFERENCES


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

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

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

- 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 | 0.26 , 0.34 | 0.26 | 0.34 | | 0.27 , 0.31 | 0.57 | 0.58 |
| 2 | 0.33 , 0.32 | 0.33 | 0.32 | | 0.19 , 0.57 | 0.74 | 0.76 |
| 3 | 0.24 , 0.54 | 0.24 | 0.54 | | 0.32 , 0.26 | 0.58 | 0.59 |
| 4 | 0.32 , 0.31 | 0.32 | 0.31 | | 0.52 , 0.08 | 0.60 | 0.60 |
| 5 | 0.24 , 0.42 | 0.24 | 0.42 | | 0.71 , 0.07 | 0.78 | 0.78 |
| 6 | 0.20 , 0.48 | 0.20 | 0.48 | | 0.35 , 0.38 | 0.71 | 0.73 |
| 7 | 0.30 , 0.57 | 0.30 | 0.57 | | -0.03 , 0.66 | 0.60 | 0.62 |
| 8 | 0.29 , 0.44 | 0.29 | 0.44 | | 0.79, -0.07 | 0.72 | 0.72 |
| 9 | 0.36 , 0.46 | 0.36 | 0.46 | | -0.50 , 1.03 | 0.49 | 0.53 |
| 10 | 0.38 , 0.38 | 0.38 | 0.38 | | 0.57 , 0.07 | 0.65 | 0.65 |

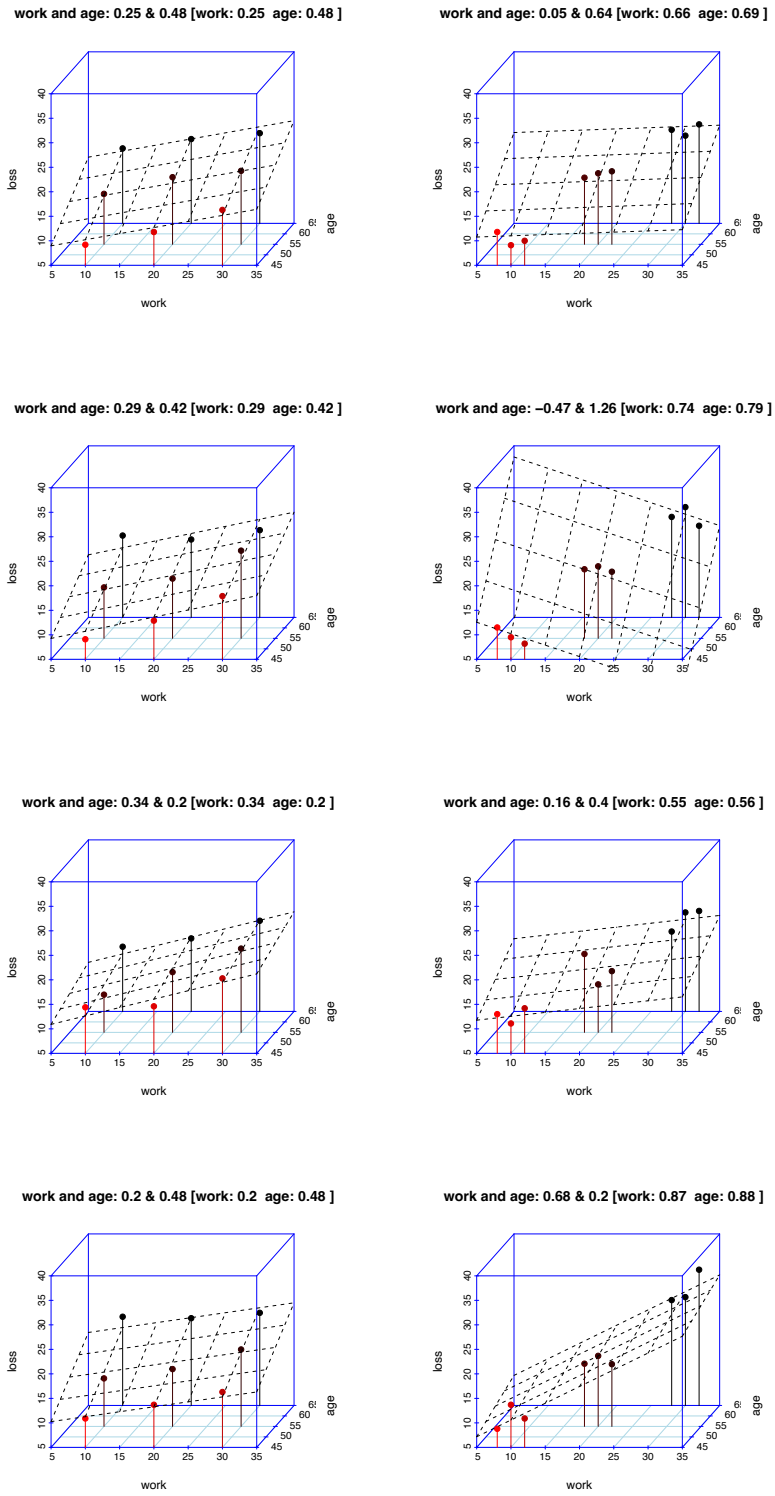





Figure 1: Estimates from samples with Balanced and Unbalanced designs [R]

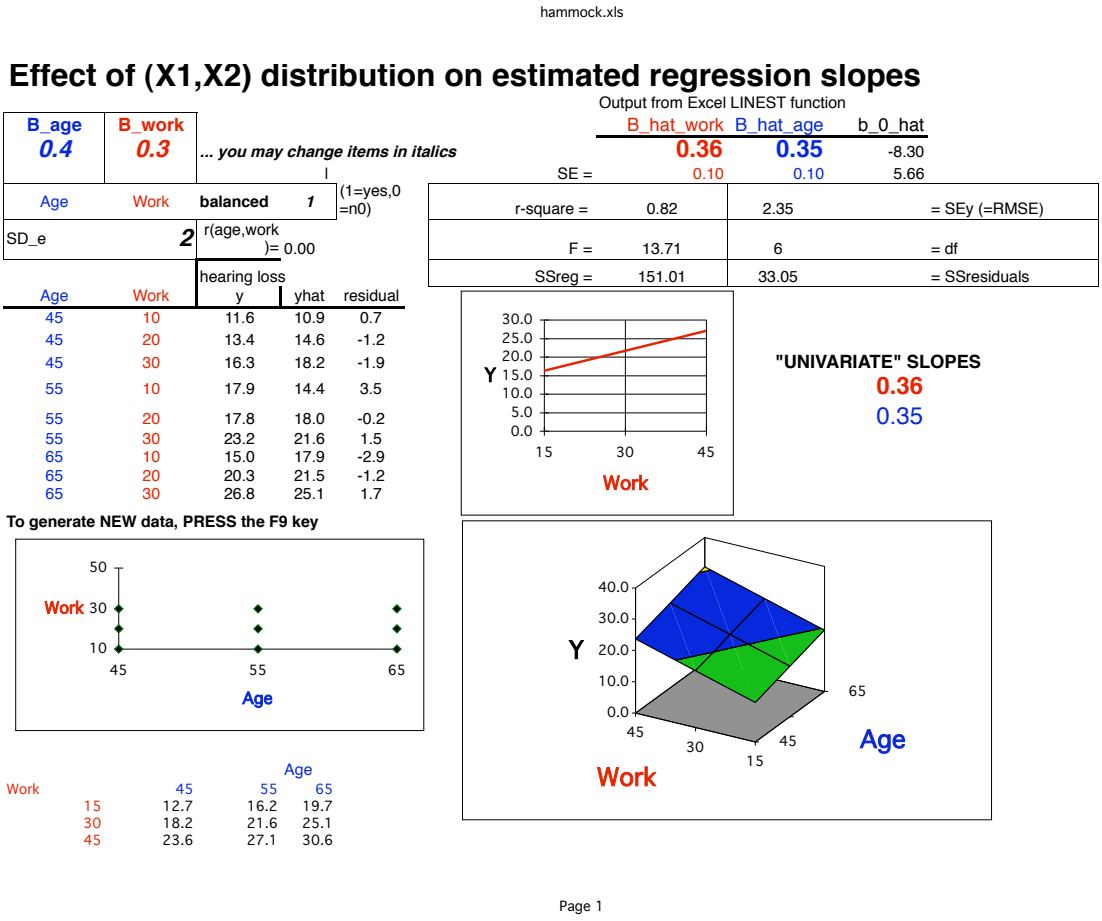



Figure 2: Estimates from a sample: Balanced design [`Excel`]

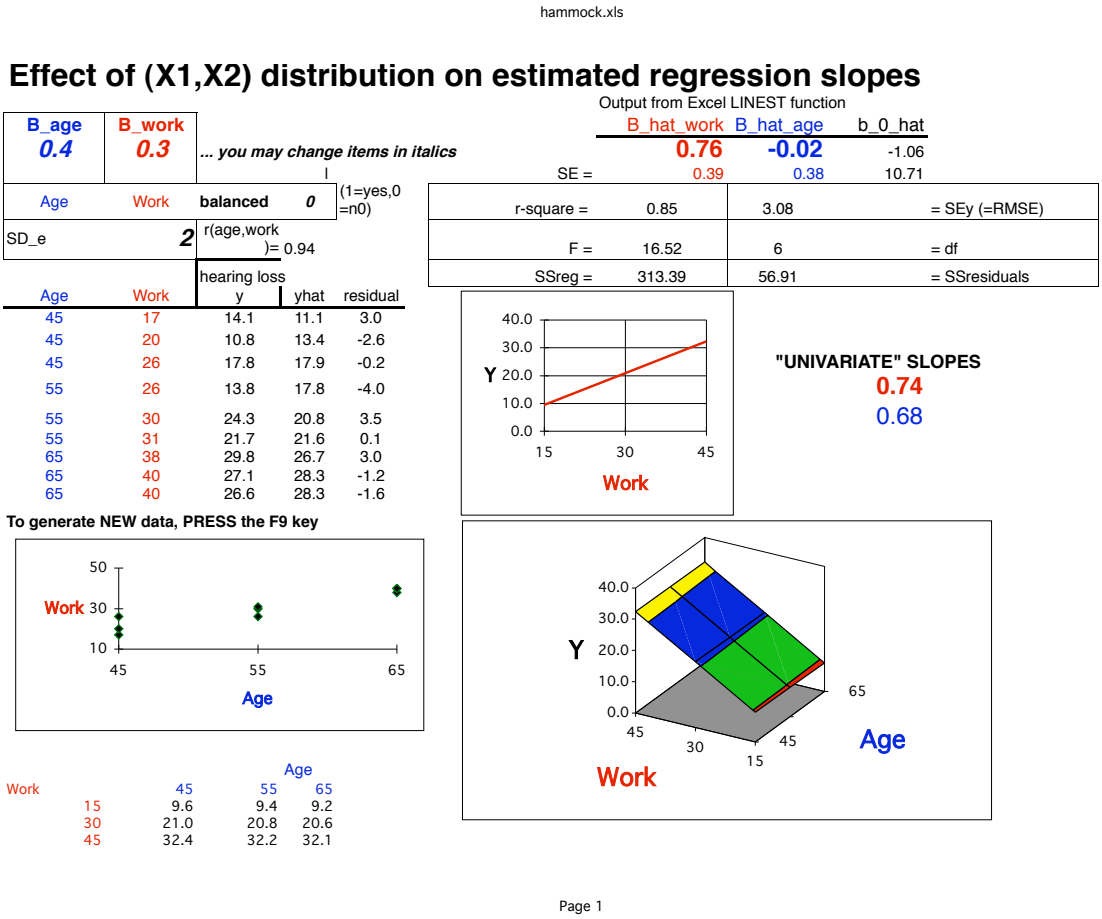



Figure 3: Estimates from a sample: Unbalanced design [`Excel`]

Search

Type here to search | This Folder | Search | Address Book | Options | Help | Log Off

Settings, Address Book, and Help

Toolbar

Reply | Reply to All | Forward | Move | Delete | Junk | Close | Previous Item | Next Item | Close

Content Area

JSE09-061
William Notz [win@stat.osu.edu]
Sent: August 11, 2009 11:47 AM
To: James Hanley, Dr.
Attachments:

Dear Dr. Hanley;

Thank you for submitting your manuscript, JSE09-061, "How collinearity affects fitted regression coefficients: visualization using a statistical `hammock' " to the Journal of Statistics Education (JSE). An Associate Editor and I have read your manuscript. Comments from the Associate Editor are in an attachment to this email.

We like the basic idea of your paper. However, we also have several concerns about the manuscript. These are discussed in detail in the attached review. Among other things, additional detail and clarity are needed to help readers get a more complete sense of the material you present. Also, since you are proposing that your tool be used for instruction and to illustrate a difficult concept, we think it would be helpful to provide more details about how this can be integrated into a classroom discussion or activity.

Because of our concerns, I do not believe the manuscript is ready to go out for additional review and I cannot accept it for publication in JSE. If you are willing to revise the paper in order to respond to the concerns raised by the Associate Editor, we would be happy to review the revision and would likely send it out to reviewers.

I am sorry I cannot be more positive in my decision, but I do hope you will revise and resubmit. If you have any questions, please contact me.

Sincerely,
Bill Notz, Editor
Journal of Statistics Education
jse@stat.ohio-state.edu
http://www.amstat.org/publications/jse

Paging and Bottom Toolbar

Previous Item | Next Item

Navigation

Mail
Calendar
Contacts
Collapse Navigation Pane

Content Area Control

Deleted Items (40)
Drafts [8]
Inbox (2020)
Junk E-mail [23]
Sent Items
Click to view all folders
BiometricsAE
Burchell
Junk (1)
Titanic
biostat
ch_var
mayrand
nserc
osm-bijm
popes

Manage Folders...

Connected to Microsoft Exchange

Review of JSE 09-061

The problem of collinearity affecting the quality of regression estimation and prediction is one that causes conceptual understanding problems for students and practitioners alike. The author has constructed a useful demonstration of this problem with helpful graphical and numerical summaries. I think the basic premise of the paper would be a good one for JSE. However, there are number of substantial problems that would need to be addressed before the paper would be suitable for publication. In particular, here are the key issues:

1. The overview of the data and sampling which produced the summaries of interest needs to be much more clearly described, to give both a context for the problem, as well as to allow the results to be reproducible. Here are some of the questions that should be specifically addressed:
    a. What is the response of interest? What are the two explanatory variables? Spell these out more clearly, and also add a sentence about what variable is of primary interest, and which is a nuisance variable which we wish to mitigate against?
    b. Is this a real data set from which we are drawing 9 values from a finite (but large) population, or is this a simulation from which we are drawing data from an underlying distribution? I suspect it is the second, but this is not clearly stated. This is probably not critical for the students, but for potential instructors this would be helpful.
    c. What are the underlying characteristics of the two explanatory variables for the population from which we are drawing? Ranges and correlations should be clearly identified (this would also help with point 3 below, when we change the correlation to see the effect on estimation).
    d. In the paper, it would be good to highlight the configuration of the data for the "balanced" and "unbalanced" with a plot. The Excel spreadsheet has the key plot for how the data are configured, but this needs to be included in the paper to clarify the layout of the data. The terminology "balanced" and "unbalanced" should also be more precisely defined.
    e. It would be good to discuss what happens if you are not able to select the desired combination of levels from the explanatory variables that give you a balanced design. Certainly in this case you could imagine that there would not be the option to find some of the "older worker" with "small number of years working" or the "younger worker" with "large number of years working". Is all lost in this case? Do we need to restrict our range of years working, or are we better off with a wider range with not perfect balance?
2. Some of the language that the author has chosen to describe the concepts seems confusing to me. In particular, the "hammock" (p.7 - a hammock to me hangs and has a shape that is not possible for the regression equation that is considered here which must be a plane) and the "arrow" (p.3 - I found this part of the paper confusing and missing some key elements to be clear) discussions seem to convey different ideas than I think are desirable for the description of the problem and solution.

3. There is an opportunity with the tool that has been developed (in Excel) to build the exercise to even more fully demonstrate the key concept. I would like to see the author describe a lesson that would allow this concept to be illustrated. One possible organization might be to start with the original problem, and then filling in some additional concepts with the ability to manipulate the level of correlation between the two explanatory variables and how this impacts the two sampling strategies. There are also opportunities to connect the discussion to simple random sampling and stratified sampling, highlighting the opportunity for better estimation in this context as well. The Excel tool has considerable flexibility in how it could be used. Additional discussion needs to be added to the paper about the different elements that are included, what can/should be changed as it is explored, and a legend for the colors of the 3-d plot.
4. I would like to see the discussion of "orthogonality" and "independence of coefficient estimation" included in the presentation of the material. If the students can gain a better understanding of the topic AND gain familiarity with the standard statistical terms used to describe the problems, this is a valuable contribution of the paper.

Overall, the idea of the paper is a helpful one, but additional detail and clarity in the paper will help readers get a more complete sense of the topic presented. Also, since the author seems to be proposing that this be a tool that is used for instruction and to illustrate a difficult concept, I think it would be helpful to provide more details about how this can be integrated into a classroom discussion or activity.