Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

TEACHING REGRESSION WITH SIMULATION

John H. Walker

Statistics Department
California Polytechnic State University
San Luis Obispo, CA 93407, U.S.A.

ABSTRACT

Computer simulations can be used to teach complicated statistical concepts in linear regression more quickly and effectively than traditional lecture alone. In introductory applied statistics classes, the coverage of important statistical topics, such as the nature of the sampling distribution of the simple linear regression slope, the problem of multicollinearity in multiple linear regression, or the danger of extrapolation in predictions from multiple linear regression models, may be shortened or eliminated entirely for lack of time. Simulation can provide a way to introduce these topics in a brief, but memorable, way for introductory students or as the first step in a more thorough treatment for higher level students. This paper describes each simulation, discusses its pedagogical advantages, and gives sample computer output.

1 INTRODUCTION

It is often said that "a picture is worth a thousand words," but this actually understates the power of computer simulations. For many statistical concepts, including those connected with linear regression that are discussed in this paper, there are no words that can equal the pedagogical power of a simulation.

This paper demonstrates three regression concepts that I have taught in both introductory and advanced applied statistics courses using simulations:

- the sampling distribution of the simple linear regression slope,
- the sampling distributions of multiple linear regression coefficients and the effect of multicollinearity on them, and
- the danger of extrapolation in predictions from multiple linear regression models.

The power of simulations is such that they can be used not only to cover these ideas more effectively, but also more quickly. This allows an instructor to cover (even briefly) more advanced topics that otherwise might be skipped due to time constraints.

For each of the topics listed above, the pedagogical goals of the simulation are defined and an example of a simulation is described. The simulations described were programmed in the *Data Desk* 6.1 software program (Data Description, Inc. 1996) and require that program to run. Some of these simulations also exist in the *ActivStats* software program (Velleman 2003). In some cases, many equivalent simulations exist for other software platforms.

2 THE SAMPLING DISTRIBUTION OF THE SIMPLE LINEAR REGRESSION SLOPE

There are many examples of simulations to demonstrate the sampling distribution of the simple regression slope. Figure 1 shows a simulation of 20 sample regression lines (from *ActivStats*) from a population with regression model

$$Y_i = 0 + 10X_i + \varepsilon_i . \qquad (1)$$

The red line in Figure 1 is the population regression line, while the 20 black lines are the sample regression lines. Such a simulation can satisfy several pedagogical goals.

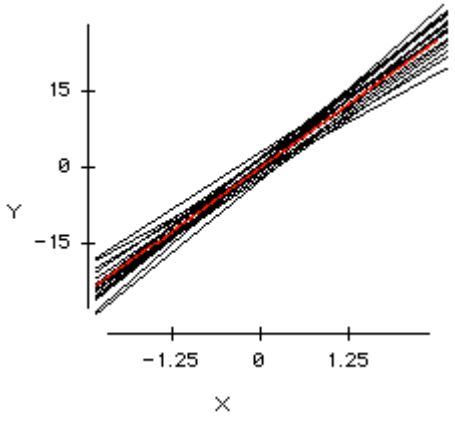

Figure 1: A Simulation of 20 Sample Regression Lines

First, it demonstrates that the regression slope is a sample statistic that varies from sample to sample in the same way as other statistics, such as the sample mean. Introductory statistics students are often so caught up in the complexities of linear regression that they forget that the regression line produced from a sample of data is only an estimate of a theoretical population regression line with its own population slope and intercept. Showing a simulation of the estimated regression lines from many samples emphasizes that the concepts of parameters and statistics still apply for more advanced statistical methods.

A good simulation should start with a single sample for demonstration before repeatedly sampling to create the distribution. It is also helpful if the simulation displays the equations of the simulated sample regression lines. This reinforces that each iteration of the simulation is a linear regression. Table 1 lists the 20 sample regression equations produced by the simulation in Figure 1.

Table 1: Sample Regression Equations for the 20 Samples Simulated in Figure 1

Sample #	Sample Regression Equation
1	Y = -0.37 + 8.31 X
2	Y = 3.02 + 8.9 X
3	Y = -0.33 + 11.01 X
4	Y = -0.19 + 10.86 X
5	Y = 0.56 + 10.37 X
6	Y = 0.71 + 11.47 X
7	Y = -0.52 + 10.56 X
8	Y = -2.54 + 11.34 X
9	Y = -0.29 + 10.33 X
10	Y = -1.66 + 9.67 X
11	Y = 0.22 + 12.22 X
12	Y = 0.37 + 11.22 X
13	Y = 1.93 + 10.74 X
14	Y = -1.02 + 9.75 X
15	Y = 0.21 + 9.43 X
16	Y = -2.18 + 9.42 X
17	Y = -0.71 + 7.63 X
18	Y = 1.44 + 9.55 X
19	Y = 1.41 + 10.72 X
20	Y = 0.26 + 10.57 X

Second, the simulation shows that the method of least squares produces unbiased estimates of the population slope and intercept. In Figure 1, the population regression line (red) can be seen in the middle of the pattern of sample regression lines (black). Figure 2 shows a histogram of these 20 sample slopes with the population slope parameter of 10 near its center. The mean of the sample slopes, 10.2, can also be used as an indication that the least squares method is unbiased. If time permits, multiple simulations for different parameter values and samples sizes could be used to show that the estimates remain unbiased in these situations.

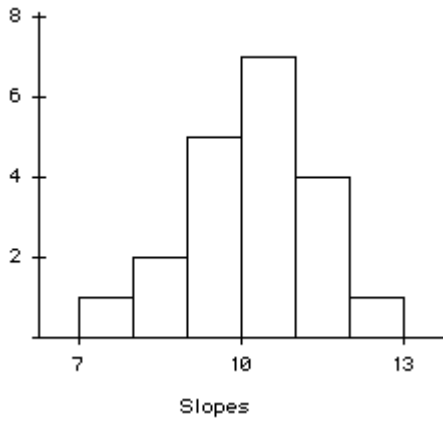

Figure 2: Histogram of the Slopes from the Simulation

Third, Figure 2 also shows that under the assumptions of the simple linear regression model, the sampling distribution of the slope has a Normal distribution. Of course, a greater number of iterations will provide a more Normal shape to the histogram. Although, I have not attempted this, the simulation could be modified to show how violations of the regression error assumptions (Normality, equal variance, and independence) affect the behavior of the sampling distribution.

Finally, the simulation can show how changes to the data affect the standard error of the sampling distribution of the slope. The standard error of the slope is (De Veaux, Velleman, Bock 2004, p. 567)

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\,s_x} \qquad (2)$$

where s_e is the standard deviation of the residuals, and s_x is the standard deviation of the *x*-values in the data.

The equation shows that the standard error of the slope is affected by three inputs—the standard deviation of the residuals, the sample size, and the standard deviation of the *x*-values in the data. An effective simulation will allow the user to alter the values of these inputs and show their effect on the spread of the sampling distribution.

Before showing students formula (2) above, I usually ask for them to guess how changing each input will affect the standard error of the slope. Many students quickly grasp the idea that larger errors or a smaller sample size will produce a more variable slope, but most students incorrectly guess that reducing the spread of the *x*-values will reduce the variability in the slope.

I use a simulation to demonstrate how the standard deviation of the *x*-values (s_x) affects the standard error of the slope. In the previous simulation (Figure 1), s_x = 1. Figure 3 shows the data for one iteration from the set of 20 shown in Figure 1. The red line in Figure 3 is the popula-

tion regression line. The black line is the estimated regression line for the plotted sample observations.

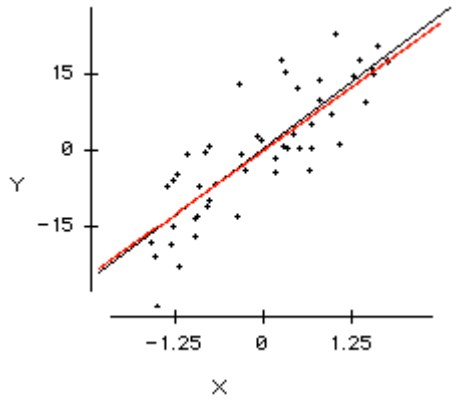

Figure 3: One Sample from the Simulation for $s_x = 1$

Figure 4 shows one sample from a simulation with $s_x = 0.25$. Many programs automatically adjust the scale of the plot, but by keeping the scale fixed, the change in the standard deviation of the *x*-values is more obvious. Again, the red line is the population regression line, and the black line is the estimated regression line of the plotted observations.

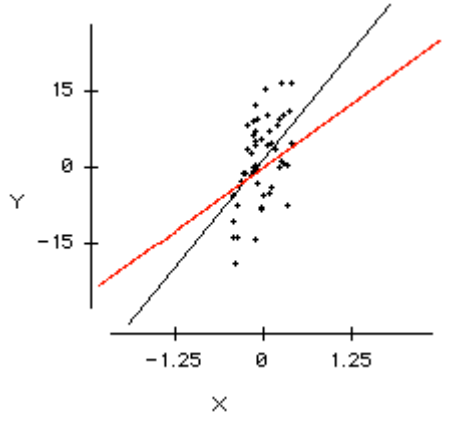

Figure 4: One Sample from the Simulation for $s_x = 0.25$

Figure 5 shows 20 sample regression lines (in black) superimposed over the population regression line (in red). Comparing Figure 5 ($s_x = 0.25$) to Figure 1 ($s_x = 1$), the change in the standard error of the slope is obvious when the *x*-values are narrowly spread. Table 2 shows the mean and standard deviation of the 20 slopes from the two simulations with $s_x = 1$ and $s_x = 0.25$. The means indicates that the least squares slope is unbiased in either case, but the standard deviation of the 20 simulated slopes is approximately 4 times larger when $s_x = 0.25$—as equation (2) would predict. This simulation is an excellent entry into a discussion of experimental design by demonstrating one reason why collecting data at widely spaced design points is superior to collecting data across only a narrow range.

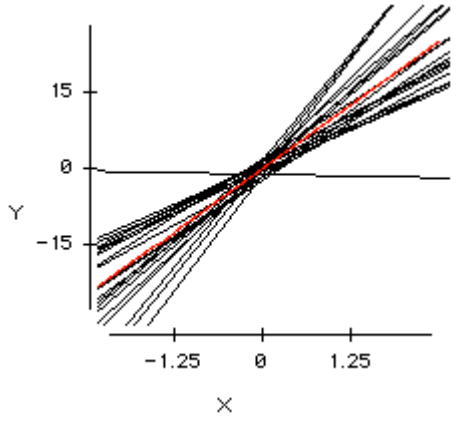

Figure 5: A Simulation of 20 Sample Regression Lines for $s_x = 0.25$

Table 2: Summary Statistics for the Simulated Slopes

s_x	Mean(b_1)	SD(b_1)
1.00	10.20	1.12
0.25	10.05	4.42

3 MULTIPLE LINEAR REGRESSION AND THE PROBLEM OF MULTICOLLINEARITY

Since multiple linear regression is even harder to visualize than simple linear regression, the use of technology to demonstrate important concepts is even more valuable. Not only is multiple regression harder to understand, but it can be affected by data problems, such as hidden extrapolations (Kutner *et al.* 2004, p. 231) and multicollinearity (Kutner *et al.* 2004, pp. 278-289) that are difficult to diagnose and can have a catastrophic effect on the regression results.

In demonstrating multiple regression, I typically use two separate simulations. Both simulate multiple regressions with two predictor variables. The first simulation is a three-dimensional simulation of the behavior of the sample regression function, a two-dimensional function or plane. The second simulation uses partial regression plots (also known as added variable plots, Kutner *et. al.* 2004, pp. 384-390) and so cannot be used unless that topic is also covered.

3.1 Simulation Using the Regression Plane

The goals of this simulation are: first, to demonstrate the variability of the sample regression function from sample to sample; second, to show that this variability increases as the correlation between the predictor variables increases; and third, to show how this correlation between the predic-

tor variables makes prediction using the regression equation extremely dangerous.

The first simulation is done for data with no correlation between X_1 and X_2. Figure 6 shows a simulated sample of 100 observations from a population with a regression relationship described by the model equation

$$Y_i = 0 + 1X_{i1} + 1X_{i2} + \varepsilon_i, \quad (3)$$

which is plotted as the red plane in the graph. The sample regression function estimated from the data is

$$\hat{y} = -0.038 + 1.063x_1 + 1.125x_2, \quad (4)$$

which is plotted as the green plane in the graph. Table 3 summarizes the estimated coefficients, standard errors, and p-values for this regression. The simulation should be repeated many times to demonstrate the variability of the sample regression plane.

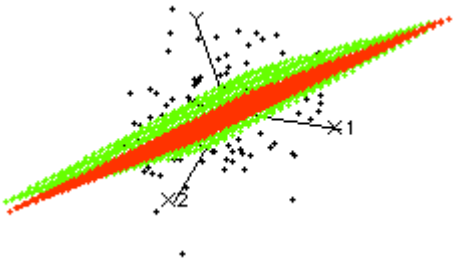

Figure 6: Plot of Y vs. X_1 and X_2 Showing the Population (Red) and Sample (Green) Regression Planes with No Correlation between X_1 and X_2

Table 3: Output for the Regression in Figure 6

Statistic	Value	Std. Err.	P-Value
b_0	-0.038	0.129	0.769
b_1	1.063	0.162	< 0.001
b_2	1.125	0.156	< 0.001

For the next simulation X_1 and X_2 have a 0.995 correlation. Figure 7 shows the same population regression plane as in (3) plotted in red. The equation for the estimated regression equation, which is plotted in green, is

$$\hat{y} = 0.106 + 1.671x_1 + 0.250x_2. \quad (5)$$

The results for this regression are summarized in Table 4. This example allows students to see how much the standard errors of b_1 and b_2 are inflated by the correlation between predictors. In this case, the effect of the multicollinearity is so large that the p-values for the individual predictors are both insignificant at a 5% significance level while the F-statistic of the overall regression (not shown) is a highly significant 83.5 with a p-value < 0.001.

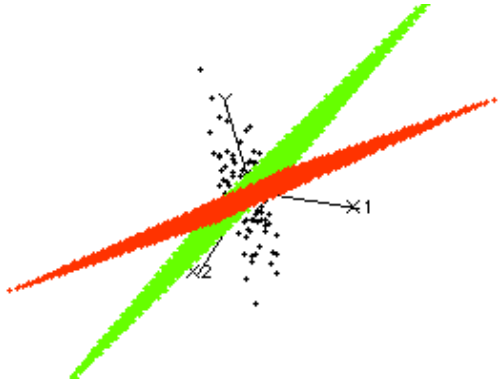

Figure 7: Plot of Y vs. X_1 and X_2 Showing the Population (Red) and Sample (Green) Regression Planes with Correlation 0.995 between X_1 and X_2

Table 4: Output for the Regression in Figure 7

Statistic	Value	Std. Err.	P-Value
b_0	0.106	0.118	0.372
b_1	1.671	1.557	0.286
b_2	0.250	1.551	0.873

This simulation can also be used to highlight the dangers of extrapolation when using the regression equation (5) for prediction. Figure 7 shows a large gap between the position of the population regression plane (red) and the sample regression plane (green). The two planes are widely separated for any combinations of X_1 and X_2 that lie far away from the sample data. Figure 8 is a plot of X_1 vs. X_2. Predictions for values (X_1, X_2) that lie within the sample data are likely to be reliable, despite the multicollinearity, because this is where the planes cross in Figure 7. However, predictions for other values (X_1, X_2) are very unreliable because the prediction $\hat{y}$ on the green plane is far from the mean response $E(Y)$ on the red plane.

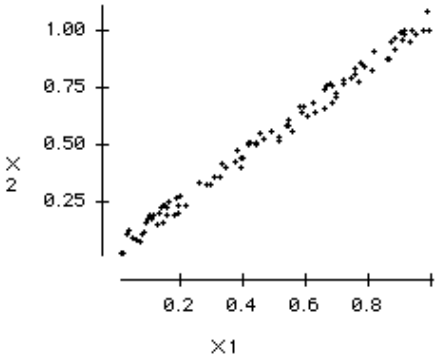

Figure 8: Plot of X_1 vs. X_2 Showing Correlation 0.995

3.2 Simulation Using Partial Regression Plots

A partial regression plot (Kutner *et al.* 2004, pp. 384-390) displays the multiple regression coefficient of a single predictor variable in a two-dimensional graph. (Contrast this with a scatterplot, which shows the simple regression coefficient.) The plot shows the relationship between the response and a predictor variable after removing any linear relationship with the other predictor variables. In the model for a response variable Y and two predictors, X_1 and X_2, there will be two partial regression plots, one for the coefficient of X_1 and another for the coefficient of X_2. The slope of the least squares regression line applied to the data on a partial regression plot of X_1 is the multiple regression coefficient of X_1.

The partial regression plot for X_1 displays "the part of Y not linearly related to X_2" on its vertical axis and "the part of X_1 not linearly related to X_2" on its horizontal axis. These adjusted variables are denoted $Y \bullet X_2$ and $X_1 \bullet X_2$ respectively to represent that Y and X_1 have been adjusted with respect to X_2. This adjustment is made by fitting a linear regression for Y vs. X_2 and X_1 vs. X_2 (in which X_1 is the response) and keeping only the residuals, which are not linearly related to X_2.

The next simulation uses animated partial regression plots to demonstrate the behavior of the multiple regression coefficients from the population regression model

$$Y_i = 0 + 1X_{i1} - 0.5X_{i2} + \varepsilon_i \qquad (6)$$

under different degrees of multicollinearity.

Let r denote the correlation between X_1 and X_2. Figures 9 and 10 show regression lines for 20 samples of n = 50 from the partial regression plots of X_1 and X_2 respectively when r = 0. For each graph, the slope of the red line is the parameter value while the slopes of the 20 black lines are the estimated multiple regression coefficients.

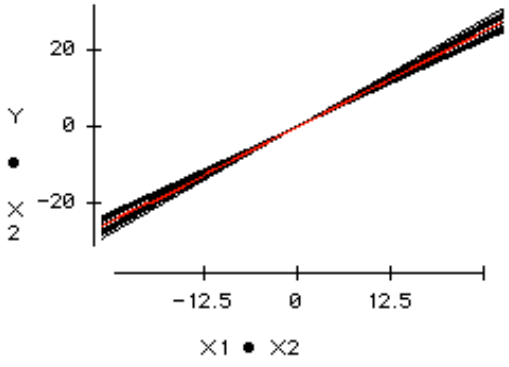

Figure 9: A Summary of Regression Lines from the 20 Partial Regression Plots of X_1 for r = 0

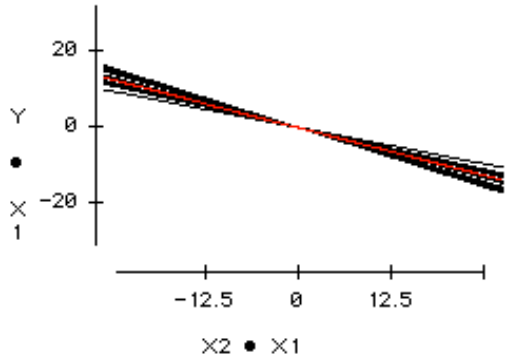

Figure 10: A Summary of Regression Lines from the 20 Partial Regression Plots of X_2 for r = 0

Table 5 shows the mean and standard deviation of the estimated coefficients from the 20 samples. The means can be used to discuss the unbiasedness of the least squares estimators while the standard deviations provide a baseline for later comparisons when X_1 and X_2 are highly correlated. As in the simple regression simulation, a table showing the 20 estimated regression functions is provided but is not shown here.

Table 5: Summaries of the 20 Estimated Coefficients for r = 0

Statistic	Mean	SD
b_1	1.000	0.068
b_2	-0.488	0.082

The simulation in *ActivStats* allows the user to adjust the correlation between the two predictors. To introduce the idea of multicollinearity, I set r = 0.995 and repeat the simulation. When X_1 and X_2 have a very strong linear relationship, the part of each that is not linearly related to the other is very small. Therefore, $X_1 \bullet X_2$ and $X_2 \bullet X_1$ will have a much smaller spread than X_1 and X_2.

Figure 11 shows the partial regression plot of X_1 for one sample when r = 0.995. Note that the spread of $X_1 \bullet X_2$, which is on the horizontal axis of the plot, is very small. Figure 11 is very similar to Figure 4, the simple regression simulation for a narrow spread of x-values.

As a result of the earlier simple regression simulation, students should already be familiar with the concept that a narrow spread of x-values will lead to more variability in the estimated slope. Using this approach, students can correctly guess that multicollinearity will increase the variability of the estimated coefficients after seeing only Figure 11—before they have even seen the full simulation shown in Figures 12 and 13.

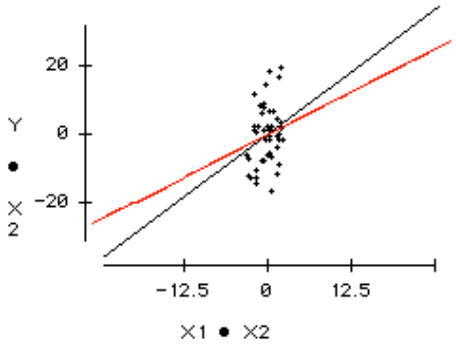

Figure 11: Partial Regression Plot of X_1 for One Sample with $r = 0.995$

Figures 12 and 13 show 20 sample regression lines from the partial regression plots of X_1 and X_2 respectively when $r = 0.995$. Each plot displays the parameter value as the slope of the red line and the estimated coefficients as the slopes of the 20 black lines. These displays demonstrate that although the estimates are still unbiased, their variability has dramatically increased. This can also be seen in Table 6, which summarizes the mean and standard deviation of the estimated coefficients of the 20 samples when $r = 0.995$.

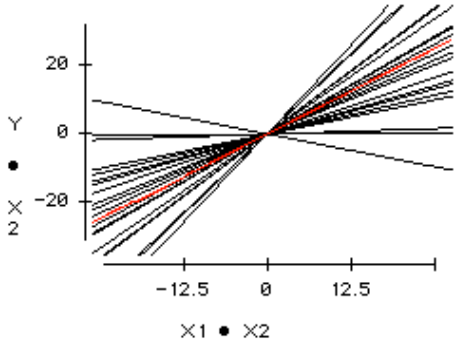

Figure 12: A Summary of Regression Lines from the 20 Partial Regression Plots of X_1 for $r = 0.995$

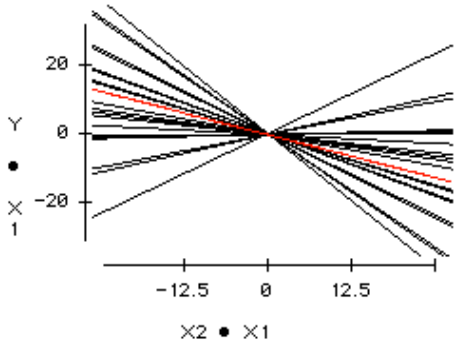

Figure 13: A Summary of Regression Lines from the 20 Partial Regression Plots of X_2 for $r = 0.995$

Table 6: Summaries of the 20 Estimated Coefficients for $r = 0.995$

Statistic	Mean	SD
b_1	0.911	0.651
b_2	-0.412	0.645

This simulation can also be used to show the relationship between the correlation r and the amount of variance inflation. Table 7 displays how the standard deviation of the 20 simulated b_1's changes when $r = 0$, 0.5, 0.75, 0.9, and 0.995. The standard deviation of the simulated b_1's is more than twice as large when $r = 0.9$ than when $r = 0$. However, when $r = 0.995$, the standard deviation is almost 10 times higher than when $r = 0$. The last line of Table 7 shows the ratio of the standard deviations when compared to the standard deviation for $r = 0$.

Table 7: Standard Deviation of Simulated b_1's for Different Correlations

r	0	0.5	0.75	0.9	0.995
$SD(b_1)$	0.068	0.084	0.111	0.143	0.651
Ratio	1	1.235	1.632	2.103	9.574

Although this simulation is a simplification of the problem of multicollinearity in real data, it can give students some insight into how large the correlation between predictors can be before the variance inflation begins to really affect the results of the analysis. If time permits, this can be linked with coverage of the variance inflation factor (Kutner *et al.* 2004, pp. 408-410), a statistic used to detect multicollinearity.

4 CONCLUSION

The examples described above are only a few of the ways that simulation can be used to enhance the teaching of both simple and multiple regression concepts. In addition to these, I have also used simulations to teach the concepts of leverage and influence and to demonstrate the effect of autocorrelated errors on a regression. In more advanced classes simulations could also be used to explore how violations of the assumptions for the regression errors (Normality, equal variance, and independence) affect the sampling distributions of the regression estimates.

Based on my own experience, these simulations increase students' understanding of the material, stimulate class discussion, and in some cases reduce the time required to cover important topics that might otherwise be deemphasized or skipped entirely due to time constraints. They can be incorporated into a traditional lecture or used in a lab setting where students do the simulations themselves. In either environment, I have found these simulations to be extremely valuable in teaching regression.

REFERENCES

Data Description, Inc. 1996. *Data Desk* 6.1. Information online at <www.datadesk.com>.

De Veaux, R. D., P. F. Velleman, and D. E. Bock. 2004. *Stats: Data and Models*. Boston, MA: Pearson Addison-Wesley.

Kutner, M. H., C. J. Nachtsheim, and J. Neter. 2004. *Applied Linear Regression Models, 4th ed.* New York, NY: McGraw-Hill Irwin.

Velleman, P. F. 2003. *ActivStats*. Data Description, Inc. Information online at <www.datadesk.com>.

AUTHOR BIOGRAPHY

JOHN H. WALKER, Ph.D. is Assistant Professor of Statistics at the California Polytechnic State University in San Luis Obispo, CA. He received his Ph.D. in Statistics from Cornell University. His research interests include the use of technology in statistics education, linear models, cluster analysis, and bioinformatics. In addition to his academic work, he is a consultant for Data Description, Inc., which develops the *Data Desk* and *ActivStats* software programs. He serves as an Associate Editor for *The American Statistician* and is a member of the Executive Committee of the Statistical Education Section of the American Statistical Association. His e-mail address is <jwalker@calpoly.edu> and his web address is <statweb.calpoly.edu/jwalker>.