Using Monte Carlo Techniques to Demonstrate the Meaning and Implications of Multicollinearity

Timothy S. Vaughan
University of Wisconsin - Eau Claire

Kelly E. Berry
University of Wisconsin - Eau Claire

Journal of Statistics Education Volume 13, Number 1 (2005), ww2.amstat.org/publications/jse/v13n1/vaughan.html

Copyright © 2005 by Timothy S. Vaughan and Kelly E. Berry, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Multiple regression; Simulation

Abstract

This article presents an in-class Monte Carlo demonstration, designed to demonstrate to students the implications of multicollinearity in a multiple regression study. In the demonstration, students already familiar with multiple regression concepts are presented with a scenario in which the “true” relationship between the response and predictor variables is known. Two alternative data generation mechanisms are applied to this scenario, one in which the predictor variables are mutually independent, and another where two predictor variables are correlated. A number of independent realizations of data samples are generated under each scenario, and the regression coefficients for an appropriately specified model are estimated with respect to each sample. Scatter-plots of the estimated regression coefficients under the two scenarios provide a clear visual demonstration of the effects of multicollinearity. The two scenarios are also used to examine the effects of model specification error.

1. Introduction

The issue of multicollinearity occupies a perplexing position in a comprehensive treatment of multiple regression modeling. The topic itself is a complex one, and as such is typically relegated to latter sections or chapters in textbooks covering multiple regression. While this “back end” treatment is entirely appropriate, the unintended implication communicated to students is that multicollinearity is an “advanced topic”, addressing an infrequent scenario that won’t affect “casual” users of multiple regression analysis.

In contrast, our position is that some degree of multicollinearity is closer to being the rule rather than the exception, when data are not collected under a purposely-structured experimental design. This is particularly true when researchers have a poor understanding of multicollinearity, and hence fail to address the issue in the variable selection and model specification stage of a research study. Indeed, when presented with such an exercise, many students generate laundry lists of highly correlated or even entirely redundant predictor variables. Thus, a lack of understanding with regard to the implications of multicollinearity actually aggravates the problem itself.

The concept itself, moreover, is difficult to teach at an intuitive level. Allison (1999) suggests that typical users’ understanding of the concept is limited to “It’s bad” and “It has something to do with high correlation among the variables.” We find that well motivated students can learn to describe the phenomenon, and apply basic diagnostic tools directed at identifying its presence. Beyond that, a solid conceptual understanding of the practical implications of multicollinearity seems to be difficult to instill.

One reason this concept is difficult for students to comprehend is that the problems arising from multicollinearity arise with respect to the sampling distributions of regression coefficients, whereas the concept is typically addressed in class with respect to a particular dataset. The typical manner of covering multicollinearity in statistics education is to demonstrate the correlation among two or more predictor variables in a regression study, and then to cast some aspersions about the validity of the regression coefficients in response to this. This “single dataset” approach leads students to attempt to answer the question, “How has this multicollinearity affected the regression coefficients on my computer output?” as opposed to “How has this multicollinearity affected the sampling distribution of regression coefficients from which my results were drawn?” While identification of multicollinearity can certainly be taught in this manner, truly understanding the implications of multicollinearity requires attention to the effects on the sampling distribution of regression coefficients.

It is of course possible to simply “tell” the students these effects, e.g. that multicollinearity increases the variance of the regression
coefficients, leads to correlated errors in the regression coefficients themselves, etc. Positively demonstrating these effects, and thus instilling an intuitive level of understanding is more challenging. Kennedy (2002) presents a method of using Ballentine Venn diagrams, which provides a conceptual framework for understanding the implications of multicollinearity.

This paper describes a Monte Carlo exercise that has been used in a second-semester statistics class, to provide a graphic demonstration of the effect multicollinearity has on the sampling distribution of regression coefficients. The basic approach is to generate a number of random datasets, for a scenario where the underlying relationships between the predictor variables and the response variable are known. For each sample, ordinary least squares regression is used to estimate the coefficients of the underlying relationship, using an appropriately specified model. Replication of this experiment generates a collection of regression coefficients which together demonstrate the nature of the underlying sampling distribution. A graphical display of the resulting coefficients under independent vs. correlated predictor variables provides a compelling depiction of the effects of multicollinearity. The basic exercise described here takes about 30 minutes. Additional explanation, side discussions, etc., will naturally impact the total class time expended.

Mills (2002) presents a thorough review and discussion of using computer simulation methods to teach statistical concepts, including regression. Although multicollinearity itself is not explicitly mentioned, there appears to be particular value in applying the technique to the teaching of this topic. The approach gives students the opportunity to observe the behavior of the regression coefficients, relative to the “true” relationship underlying the data.

The demonstration as described here is implemented using a Microsoft Excel spreadsheet. Of course, the technique could be implemented using any statistical package capable of generating independent and correlated samples, computing appropriate response variable values, and running the necessary regressions. Indeed, the entire process could be automated with an appropriately written macro for the software package of choice. While this approach would speed up the demonstration, we feel there is potential for reduced learning impact when the activities described here fall into a “black box”.

2. The Demonstration

The demonstration is given after the students have had significant exposure to the basics of multiple regression. In particular, the students are familiar with the underlying multiple regression model:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \]  

(1)

where \( \varepsilon \) is normally distributed with mean \( E[\varepsilon] = 0 \) and variance \( \text{VAR}[\varepsilon] = \sigma^2 \) As such, it is not difficult to ask them to envision a scenario in which:

\[ y = 300 + 1x_1 + 1x_2 - 1x_3 + \varepsilon \]  

(2)

where \( \varepsilon \) is normally distributed with mean \( E[\varepsilon] = 0 \) and variance \( \text{VAR}[\varepsilon] = 1600 \). We thus have a situation in which the “true” relationship coefficients are known, e.g. \( \beta_1 = 1, \beta_2 = 1, \beta_3 = -1 \). As with any Monte Carlo demonstration of an estimation technique, it is imperative to remind students that in practice we would not know these values, but present them only to compare our estimates with the “right” answers.

We display an Excel spreadsheet (a portion of which is shown in Figure 1) which generates a sample of \( n = 100 \) random observations of \((x_1, x_2, x_3, \varepsilon)\), and computes the corresponding value of \( y \) under the relationship in equation (2). (In our demonstration, the \( x_i \) are generated as independent observations from the uniform \([0,100)\) distribution, then rounded to the nearest integer.) The random generation process is demonstrated by hitting the F9 key a few times, and we verify that the \( y \)’s being computed correctly reflect the underlying relationship.
After a short side discussion reviewing the iterative process of model specification, parameter estimation, diagnostic checking, etc., we proceed to fit the "right" model for the current dataset. The resulting predictor variable coefficients $\beta_1$, $\beta_2$, and $\beta_3$ are then copied into a separate worksheet. This process is repeated 10 or more times, resulting in 10 or more different $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$. (It is important that the students actually witness at least a few replications of the data generation/parameter estimation procedure. Depending on class time available, it is possible to have some of the replications done ahead of time, making it clear that these coefficient sets were produced under the same procedure.)

Before examining the regression coefficients further we point out the correlation matrix displayed in Figure 1, which suggests the $(x_1, x_2, x_3)$ being generated are independent of one another. The independence between $x_1$ and $x_2$, in particular, is graphically demonstrated with a scatter plot as shown in Figure 1. (An instructor might, at this point, note the sampling variability in the correlation coefficients. This provides an opportunity to review the idea that coefficients strictly equal to zero will not be realized, despite the fact that $x_1$ and $x_2$ are independent.)

With independence between $x_1$ and $x_1$ firmly in mind, we turn to a scatter-plot of the various $\hat{x}_1$ and $\hat{x}_2$ as depicted in Figure 2. (The worksheet into which the regression coefficients are copied is already set up to generate the plot.) The grid lines plotted at $\hat{\beta}_1 = 1$ and $\hat{\beta}_2 = 1$ form "crosshairs" on the true underlying parameters $\beta_1 = 1$, $\beta_2 = 1$. As the respective average values of $\hat{\beta}_1$ and $\hat{\beta}_2$ are approximately equal to 1, our procedure is generating unbiased estimates of the "separate effects" of $x_1$ and $x_2$ on $y$. Nonetheless we see a certain degree of estimation error around the true $\beta_1$ and $\beta_2$, with $\hat{\beta}_1$ ranging from 0.63 to 1.27, and $\hat{\beta}_2$ ranging from 0.75 to 1.14. (This is a good time to review the idea that the term "effect of $x_i$" refers to the change in $E[y]$ associated with, but not necessarily

| I. Independent $x_1, x_2, x_3$ | \(\beta_1\) | \(\beta_2\) | \(\beta_3\) | \(\sigma\) | \(\beta_0\) |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | -1 | 40 | 300 |
| $x_2$ | 82 | 90 | 66 | -49.06 | 356.94 |
| $x_3$ | 13 | 78 | 32 | -42.08 | 401.08 |
| \(\varepsilon\) | 53 | 29 | 59 | -8.84 | 314.15 |
| \(\gamma\) | 84 | 19 | 13 | 35.15 | 425.15 |
| $\gamma$ | 61 | 91 | 84 | 35.70 | 403.70 |
| $\gamma$ | 61 | 88 | 77 | -50.60 | 321.40 |
| $\gamma$ | 35 | 66 | 25 | -42.48 | 333.52 |
| $\gamma$ | 3 | 38 | 38 | 1.33 | 304.33 |
| $\gamma$ | 8 | 55 | 54 | 5.30 | 314.30 |
| $\gamma$ | 26 | 28 | 65 | -59.06 | 225.94 |
| $\gamma$ | 66 | 32 | 42 | 53.82 | 409.82 |
| $\gamma$ | 26 | 74 | 84 | 59.71 | 375.71 |
| $\gamma$ | 95 | 49 | 60 | 78.38 | 462.38 |
| $\gamma$ | 98 | 42 | 22 | -10.22 | 407.78 |
| $\gamma$ | 80 | 85 | 16 | 29.64 | 469.64 |
| $\gamma$ | 70 | 86 | 55 | -4.53 | 396.47 |
| $\gamma$ | 62 | 12 | 62 | -12.99 | 299.01 |
| $\gamma$ | 82 | 53 | 84 | 7.73 | 358.73 |
| $\gamma$ | 98 | 55 | 22 | 32.53 | 463.53 |
| $\gamma$ | 68 | 76 | 54 | -18.04 | 371.96 |
| $\gamma$ | 63 | 73 | 49 | -40.42 | 346.58 |
| $\gamma$ | 83 | 92 | 19 | 2.31 | 458.31 |

**Figure 1.** Portion of spreadsheet used to generate data for the independent variables case.
caused by a 1-unit increase in $x_i$ with all other variables held constant.)

\[ \text{Parameter Estimates (}\beta_1 = 1, \beta_2 = 1\text{)} \]

\[ \text{Uncorrelated } x_1, x_2 \]

![Diagram showing regression coefficients](image)

**Figure 2.** Regression coefficients produced by independent data.

Once this point is made, we turn to a second worksheet (shown in Figure 3) which computes y values using equation (2) in a manner identical to that used in the first sampling effort. We then point out the only difference between this sampling mechanism and the previous, i.e. the correlation between the $x_1$ and $x_2$ values reflected by the correlation matrix and scatter plot. In our demonstration, each $x_2$ value is generated from a uniform distribution with mean equal to the corresponding $x_i$ and range governed by a parameter $w$. The data in Figure 3 is generated with $w = 20$, thus each $x_2$ is generated uniformly over the range $x_i \pm 10$. The process of generating 10 or more different observations of ($\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$) is performed with data generated by this worksheet, producing a display such as shown in Figure 4.
II. Correlated $x_1, x_2$

| $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma$ | $\beta_0$ |
|-----------|-----------|-----------|----------|----------|
| 1         | 1         | -1        | 40       | 300      |

| $x_1$ | $x_2$ | $x_3$ | $\varepsilon$ | $\gamma$ |
|-------|-------|-------|---------------|----------|
| 32    | 22    | 80    | -34.71        | 239.29   |
| 3     | 11    | 38    | -93.78        | 182.22   |
| 36    | 38    | 76    | 5.23          | 303.23   |
| 76    | 71    | 8     | -9.37         | 429.63   |
| 64    | 58    | 85    | 17.25         | 354.25   |
| 49    | 43    | 16    | 35.19         | 412.19   |
| 2     | 0     | 67    | -55.42        | 179.58   |
| 73    | 67    | 74    | 63.74         | 429.74   |
| 71    | 11    | 93    | 67.55         | 306.55   |
| 68    | 68    | 4     | 23.39         | 453.39   |
| 38    | 35    | 97    | 32.99         | 308.99   |
| 35    | 43    | 3     | -16.19        | 358.81   |
| 11    | 1     | 33    | 10.02         | 289.02   |
| 54    | 60    | 32    | -28.07        | 353.93   |
| 99    | 53    | 12    | 29.07         | 429.07   |
| 68    | 74    | 8     | -23.38        | 410.62   |
| 98    | 100   | 20    | 0.48          | 478.48   |
| 31    | 36    | 70    | 16.73         | 313.73   |
| 9     | 12    | 88    | 0.76          | 233.76   |
| 45    | 52    | 30    | -41.77        | 408.77   |
| 33    | 42    | 79    | -13.68        | 282.32   |
| 35    | 36    | 40    | 19.09         | 350.09   |
| 40    | 46    | 11    | 30.32         | 405.32   |

**Correlation Matrix:**

|       | $x_1$ | $x_2$ | $x_3$ |
|-------|-------|-------|-------|
| $x_1$ | 1.0000|       |       |
| $x_2$ | 0.9796| 1.0000|       |
| $x_3$ | 0.0163| 0.0016| 1.0000|

*Figure 3. Portion of spreadsheet used to generate correlated $x_1, x_2$.***

---

5 of 9
The striking contrast between Figure 2 and Figure 4 form the primary impact of the exercise. We first note that the average $\hat{\beta}_1$ and $\hat{\beta}_2$ are still approximately 1, implying the regression parameters remain unbiased estimates of the true relationship coefficients. In contrast, we then note the inflated variance of the individual coefficients, as $\hat{\beta}_1$ and $\hat{\beta}_2$ now range over ±1.3 from their expectations. Finally, we draw attention to the manner in which $(\hat{\beta}_1, \hat{\beta}_2)$ are no longer clustered around the point (1,1), but rather tend toward the line $\hat{\beta}_1 + \hat{\beta}_2 = 2$, in the vicinity of (1,1). Thus, the sampling errors in the $(\hat{\beta}_1, \hat{\beta}_2)$ coefficients have themselves become correlated. To explain this, we note we have induced a very simple type of correlation such that $E[x_2] = x_1$. As such, across many datasets we realize a collection of $\hat{\beta}_1$ and $\hat{\beta}_2$ for which $\hat{\beta}_1 + \hat{\beta}_2 \approx 2$. Under any such pair, the expression $\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ adequately represents the net effect of $x_1 + x_2$ when $x_1 \approx x_2$. (Some simple algebraic substitution into (2) with $x_1$ is strictly equal to $x_2$ helps to clarify the point.) In short, our regression coefficients are no longer reliably estimating the “separate effects” of $x_1$ and $x_2$, but rather measure the net effect of $x_1$ and $x_2$ varying together in a manner depicted in Figure 3.

In order to make the comparison of Figure 2 and Figure 4 as compelling as possible, an extreme degree of collinearity between $x_1$ and $x_2$ has been introduced. Time permitting, the latter sampling procedure is repeated under less severe collinearity, generating a display similar to that shown in Figure 5. (The results in Figure 5 were generated with the $w$ parameter discussed above equal to 60.) While the sample coefficients still tend to the line $\hat{\beta}_1 + \hat{\beta}_2 = 2$, the variance inflation effect is reduced under weaker collinearity. Here $\hat{\beta}_1$ ranges from 0.63 to 1.73, while $\hat{\beta}_2$ now ranges from 0.31 to 1.34. Having witnessed the creation of Figure 2 and Figure 4, it is not as critical for the students to witness the sampling involved in this latter case.
Figure 5. Regression coefficients under less extreme collinearity between $x_1, x_2$.

The exercise closes with a discussion of how multicollinearity may affect the utility of a multiple regression model. Here, we turn to the two uses of a regression model, e.g. explanation vs. prediction.

If the model is to be used for explanation (e.g., we are more interested in the $\hat{\beta}$ parameters than the $\hat{y}$ values produced by the model), multicollinearity casts a shadow of doubt over the classic interpretation of $\hat{\beta}$ as “the change in $E[y]$ associated with a 1-unit increase in $x_i$.” As we have seen above, multicollinearity reduces our ability to isolate the change in $E[y]$ associated with a change in any single $x_i$. This concept is used to introduce and motivate the concept of experimental design, the next topic in the course in which the demonstration is used.

If the model is to be used for prediction (e.g., we are more interested in the $\hat{y}$ values produced by the model, and less concerned with the individual $\hat{\beta}$ parameters), multicollinearity is not quite as damning. We do, however, need to base our predictions on the assumption that the predictor variables continue to reflect any relationships present in the data used to fit the model. The scatter plot of $x_2$ vs. $x_1$ from Figure 3 provides an opportunity to present this idea as a special case of the more basic concept, that a regression model should only be used for prediction within the range of $x$’s actually observed. In this case, the basic idea extends to the region of $(x_1, x_2)$ actually observed, rather than simply the range of $x_1$ taken in union with the range of $x_2$.

One last useful point that can be made with this exercise is to clarify the distinction between multicollinearity and interaction. (Students tend to confuse the concepts, especially when relying on a memorized definition as opposed to conceptual understanding.) Returning to the scatter plot from Figure 3, the multicollinearity in this exercise is the relationship between $x_1$ and $x_2$. Interaction between $x_1$ and $x_2$, on the other hand, would exist if we had used the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1 x_3 + \varepsilon$$ (4)
3. Alternatives and Extensions

As with any teaching material, there are numerous ways in which the basic demonstration outlined above could be modified or extended, depending on the learning goals adopted by the instructor and the amount of time one wishes to spend on the topic. It might be useful to ask students to anticipate the effect on model parameters if \( x_2 \preceq 5 + 5x_1 \) (for example) rather than the simpler \( x_2 \preceq x_1 \) correlation model used here. Rather than simulating, this issue can initially be addressed with some simple algebraic substitution for the case \( x_2 \) strictly equal to \( 3 + 5x_1 \) which helps to clear up any impression that the tendency toward \( \hat{\beta}_1 + \hat{\beta}_2 \neq 2 \) is a general result. (In the more general case where \( x_2 \preceq a + bx_1 \), the regression coefficients \( (\hat{\beta}_1, \hat{\beta}_2) \) will tend toward a line passing through \( (\beta_1 + b, 0) \) and \( (0, b) \), in the vicinity of \( (\beta_1, \beta_2) \).

It is important to note that if the Monte Carlo demonstration is conducted with a more general correlated case of the form \( x_2 \preceq a + bx_1 \), the variances of the predictor variables must be held constant across the two scenarios. Otherwise the scenario with the larger predictor variable variance has an advantage in reducing the variance of the corresponding sample coefficient. Of course, increasing the strength of the predictor variable correlation relative to the variance of the response variable magnifies the contrast between the two scenarios.

Some instructors may wish to adopt a realistic setting for the demonstration, rather than simply referring to the variables generically as \( x_1, x_2, x_3, y \). One anonymous reviewer proposed an example in which the predictor variables are exam scores and class attendance, while the response variable is the student’s score on a final exam. While the course in which this demonstration is used is replete with regression examples grounded in realistic scenarios, for this demonstration we prefer the generic approach. This is partially due to the difficulty in identifying a realistic example in which two predictor variables might be independent, and a few minutes later are assumed to be correlated. Moreover, there is potential for student confusion in this change of assumption with respect to a specific variable. One particular exception would be a scenario in which the correlated data arise due to convenience sampling, while the independent observations are gathered under a controlled experimental design.

Our demonstration has used a sample size \( n = 100 \) observations per replication. Time permitting, the effect of sample size on the results demonstrated above would be an interesting extension or out-of-class assignment. A smaller sample size (say \( n = 20 \)) would of course result in greater coefficient variation under both scenarios, but would generate a similar if not more impressive contrast between the independent vs. correlated predictor variables case. Interestingly, a larger number of replications is not required to see the contrast, despite the smaller sample size.

4. A Model Specification Exercise

An interesting follow-up to this demonstration is an exercise in model specification. Using independent \( x_1, x_2, x_3 \) and \( y \) values computed using equation (2), we fit the model

\[
\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_3
\]

(5)
e.g., we omit the \( x_3 \) variable from the model specification. Replicating the procedure a few times, we come to the conclusion that we are still getting unbiased estimates of the “effect of” \( x_1 \) (e.g. \( \hat{\beta}_1 \equiv 1 \)) while the missing \( x_2 \) term is contributing to the error term of the model, as evidenced by larger standard error and smaller adjusted R\(^2\).

When the same procedure is used with correlated \( x_1 \) and \( x_2 \) of the type depicted in Figure 3, we naturally come to the conclusion that the \( \hat{\beta}_1 \) parameter is reflecting the combined effect of \( x_1 \) and \( x_2 \) (e.g. \( \hat{\beta}_1 \equiv 2 \)) thus the missing \( x_2 \) term is contributing to the \( \hat{\beta}_1 \) parameter rather than the error term. The bottom line to this exercise is to recognize that, to the extent any “other variables” vary independently of the variables included in the model, the effects of these variables contribute to the error term. To the extent that other variables are correlated with the variables in our model, we must realize our regression coefficients measure the “net effect” of these simultaneous changes, for each unit change in the variable we have measured and included in our model. This gives additional credence to the regression instructors’ mantra that regression coefficients measure association not cause and effect.

The model specification exercise could also be extended in a number of ways. For example, a “true” functional relationship in which a correlated predictor variable contributes nothing to the observed \( y \) values (e.g. \( \hat{\beta}_2 = 0 \) in the preceding example) provides additional fodder in this direction. Such an exercise could examine the merits of a model containing both variables, relative to a reduced model possibly having greater predictive power.

5. Conclusion
While simulated data should never entirely replace analysis of real datasets in statistics education, selective use of the technique can greatly enhance students’ understanding of certain complex topics. With regard to multicollinearity, the ability for students to actually witness the effect on the regression coefficients, vis-à-vis the “true” relationship parameters, has an impact that is difficult to replicate through other means.

References

Allison, P. D. (1999), *Multiple Regression – A Primer*, Thousand Oaks, CA: Pine Forge Press.

Kennedy, P. E. (2002), “More on Venn Diagrams for Regression,” *Journal of Statistics Education* [Online], 10(1). (ww2.amstat.org/publications/jse/v10n1/kennedy.html)

Mills, J. D. (2002), “Using Computer Simulation Methods to Teach Statistics: A Review of the Literature,” *Journal of Statistics Education* [Online], 10(1). (ww2.amstat.org/publications/jse/v10n1/mills.html)

Timothy S. Vaughan
Management and Marketing
University of Wisconsin - Eau Claire
Schneider Hall 453
Eau Claire, Wisconsin 54702
U.S.A.
VAUGHATS@uwec.edu

Kelly E. Berry
Management and Marketing
University of Wisconsin - Eau Claire
Schneider Hall 423
Eau Claire, Wisconsin 54702
U.S.A.
BERRYKE@uwec.edu