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

USING SIMULATED DATA IN SUPPORT OF RESEARCH ON REGRESSION ANALYSIS

Christopher M. Hill

Center for Army Analysis
6001 Goethals Road
Fort Belvoir, VA 22060

Linda C. Malone

Industrial Engineering and Management Systems
University of Central Florida
4000 Central Florida Boulevard
Orlando, FL 32816-2450

ABSTRACT

Using simulated data to develop and study diagnostic tools for data analysis is very beneficial. The user can gain insight about what happens when assumptions are violated since the true model is known. However, care must be taken to be sure that the simulated data is a reasonable representation of what one would usually expect in the real world. This paper discusses the construction of simulated data sets and provides specific examples using linear and logistic regression analysis. It also addresses the execution of simulation based data studies following data construction.

1 WHY USE SIMULATED DATA?

The research of analytical techniques through simulation analysis provides benefits that are not possible from research based exclusively on theoretical models. Often assumptions are violated in practice when analyzing real data where the true relationships in the data are unknown. Simulation allows a level of knowledge and control that leads to more robust and defendable solutions. Many of our commonly used analytical techniques have existed for long periods of time, are widely accepted, and are used for a wide range of application types. For example, regression analysis and the method of least squares dates back to the early 20th Century, and is used for analysis on everything from biological processes to national economies (Myers 1990). However, if the basic assumptions do not hold, it is difficult to evaluate results or make comparisons with competing techniques. In an application like regression, the early data sets most often studied were characteristically more simple and smaller than representative modern data sets, which are significantly more complex and large. Hence, obtaining a straightforward assessment of the technique's strengths and weaknesses is not clear-cut. These evolving complex applications can drift away from many of the critical theoretical assumptions, so a clear assessment of performance using only real test data sets may not be possible. Using simulated data sets, where the characteristics of the data are controlled and varied, will lead to better research of the technique's abilities in general, and to assessments of its performance in specific applications.

In many modern applications, the researcher truly does not know what they don't know. It is very common to see research where various techniques' performance is compared against a real data set, to see which model performs better. In many of these studies, alterations or improvements of a given technique will be compared to a group of others to determine a best method. This approach seems very common in emerging fields such as so-called data mining applications. Examples of this approach are Foster and Stine (2001); Morrison, Bose, and O'Leary (2000); Shirata and Terano (2000); Reinartz (1998); and John and Langley (1996). This approach has several weaknesses. One significant concern is the increased likelihood of repeated tuning. Salzberg (1997) defines this issue as researchers continually striving for algorithm improvements that achieve improved performance on a given data set. The problem is then modeling nuances of the given data and relating them to global conclusions about algorithmic improvement.

Another concern is this approach may result in a comparison of one model's poor performance to another's, and this may even lead the modeler to overestimate the model performance. Hill and Malone (2003) showed in linear regression applications in large data sets it is possible to obtain consistently bad performance. Simulation allows the separation of Type I and II regression errors; this measurement is not possible using real data. A Type I regression error is selecting a truly insignificant variable as significant, or rejecting a hypothesis when it is true, while a Type II regression error is considering a variable insignificant when it is actually significant (Shaeffer & McClave 1990). The Type II regression error is not mentioned in the literature, probably because of the inability to detect them unless one is using simulated data sets. In these examples, a series of noise models can all exhibit Type I errors while a realistically simulated data set is required to show both Type I and II errors consistently. Without the knowledge of the true model, it is easy to completely miss these types

of errors and consequently see poor performance. Using simulated data allows for better identification of strengths and weaknesses of each approach.

If one had competing techniques available like linear or logistic regression, recursive partitioning, or neural networks, etc., a common approach would be to take a set of real data, specify and estimate the model, then compare the model's performance on prediction using a holdout portion of the data. This type of research would provide a ranking of the performance of the competing techniques against the prediction application. However, we may be comparing models that are all bad. The other issue is that the holdout data set (especially with large n) might be flawed. Also if the researcher is interested in accurate coefficient estimates there is really no way to judge the best model among competing techniques using real data. On the other hand, the same research using simulated data sets with known and controlled relationships can lead to the identification of the specific issues associated with a given technique. This knowledge could lead to development of new solutions to the specific problems, and improved performance in modeling. This would be difficult if not impossible to achieve using real data.

The use of simulation allows multiple comparisons during research of a given analytical technique. The data can be altered methodically to test violations of assumptions of the technique under consideration. For example, in classical linear regression analysis assumptions are that independent variables are not a perfect linear function of other independent variables, independent variables are not correlated with the error term, the errors are normally distributed with mean zero and constant error variance, and the errors are independent (Studenmund 2001). These assumptions could be methodically altered to determine their impacts in a specified setting. Additional problems like outliers or data related problems might also be induced. The setting could be altered as well, such as using varied sample sizes, or different objectives, like prediction versus estimation

Simulation based research also provides the ability to change other characteristics of the technique under study. In the regression case, it allows separation and measurement of Type I and II errors; this approach is not possible using a real data set. It also allows variation of the numbers of independent variables, their distributions, and the type of variable (categorical versus continuous). It allows for manipulation of the number and proportion of variables that influence the response. Another aspect of using simulation specifically to analyze regression performance deals with the effects of selecting a specific alpha level for variable selection. Skipper, Guenther, and Nass (1967) suggest that selection of a significance level is a matter of convention and is somewhat arbitrary, although the nature of the problem under study should dictate the decision. Simulation based modeling enables a scrutiny of these issues.

In short, the ability to control characteristics of the data and to have knowledge about the model and the data characteristics make simulation based research a powerful tool. In fact, this approach will likely lead to broader and more defendable solutions than other approaches. No known technique always outperforms all others regardless of data characteristics. Simulation allows for discovery of issues and detailed comparisons of strengths and weaknesses between competing approaches. This leads to an opportunity for an analyst to bound problems and discover regions where violation of particular assumptions may have a greater impact than others.

2 BUILDING THE SIMULATION MODEL

The process of building a good simulated data set is the most critical step in conducting research in this manner. Hill and Malone (2003) show that data which are too clean or well behaved will provide misleading results. On the other hand, simulated data that has unrealistically large errors, levels of multicollinearity, numbers and magnitude of outliers, etc. will also provide misleading results. Benchmarking the data set can resolve these issues by ensuring the data is realistic. The benchmarking effort should involve comparing the simulated data set against problems that are widely accepted and understood; this will improve the simulated data's credibility. This process can take the form of considering historic studies in the field, application area, or industry under consideration. It can also include benchmarking characteristics against examples from significant or foundational literature. The characteristics of these benchmark studies can be blended with the characteristics of the real data under eventual consideration to create a realistic simulated data set. This is accomplished by finding examples, or precedent, in benchmark cases that are characteristic of the problem under study. This paper will provide both general comments about and specific examples of the construction of simulated data for linear and logistic regression.

2.1 Data Construction

The first step is the construction of the group of independent variables and the response. In simulating the group of independent variables one must consider factors like the number of variables to be in the model, the individual characteristics of the variables, like distribution, range, and variability, the proportion of independent variables that contribute to the response, degree of association or dependence between the variables, and the level of contribution of the variable to the response. These types of decisions should be made by analyzing historic studies, benchmark cases, and example real data sets. The objective is not to create exact replicas of the example cases, but rather to create data that is feasible in real world applications. Analysis of real data as an example requires great caution. Although the data provides much meaningful in-

formation, caution is warranted to avoid over-modeling or repeated tuning problems of the data. The analyst must also guard against obtaining pre-conceived notions about relationships in the data. It is not important to have the correct specific details in the data set, like exactly which variables will really be significant; however, it is important that the simulated data set exhibits the same overall characteristics as in the application field in general.

Once the set of independent variables is formed, the next step is forming the response variable. In the regression example this means determining how many variables to include in the true model, which variables will influence the response, and assigning a coefficient value. It is important to get an adequate proportion and mix of different types of variables. The next component to add into the equation is the error term. As noted earlier, there are several classic assumptions of regression that must be considered in the construction of the error term, with particular focus on the magnitude of the error term. During this portion of the simulated data construction process, it is important to benchmark the data against foundational data. Examples from linear regression and logistic regression will clarify this step.

2.2 Linear Regression

For the linear regression examples, we will provide cases from the Myers (1990) regression text, the Myers and Montgomery (1995) response surface text, and the Studenmund (2001) econometrics text as benchmark examples. These examples provide cases with designed experiments, observed data, engineering applications, econometric examples, mixed variable types, and varied levels of behavior in the data. Several tables and Appendix A provide different aspects of the examples as a general guide for construction of simulated data. These examples contain characteristics of different responses, characteristics of the model, and values for statistics measuring multicollinearity. The reference for each example is provided in Appendix A. It is important to realize the benchmarking step is as much an art as science. The analyst must consider a range of descriptive statistics before settling in on the specifics of simulated data. For example, analysis of the range of a response is not sufficient. The range must be considered simultaneously with the response average, the maximum value, and the minimum value. Likewise, the Mean Squared Error (MSE) in isolation is not as meaningful a benchmark as it is in combination with ratios of the MSE to descriptors of the range. When comparing examples with widely different ranges of response and MSE, it is also useful to use transformations of these statistics, like the natural log. Finally, benchmarks of the statistics describing relationships in the data also warrant care. Measures of multicollinearity and variability must be considered together. This general benchmarking approach will be illustrated with an example.

2.3 Linear Regression Example

Consider an example where the research objective is understanding the effects of large numbers of observations on linear regression analysis. One application of this research will be the United States Army Recruiting Command (USAREC). The first step in the process of constructing simulated data is consideration of reasonable characteristics for the data. In this case, our analysis will center on personnel who have signed an Army contract and are awaiting shipment to training. The first step is to develop a group of variables similar to those used in typical USAREC Delayed Entry Program (DEP) analysis. After consideration of historic studies and real data sets, 37 independent variables, 11 of which were continuous and 26 of which were categorical, were constructed. For the categorical variables, the numbers of categories range from 12 to 2. Some of these are generated via Empirical distributions and some via Bernoulli. The continuous variables were generated from a wide range of distributions including Normal, Beta, Weibull, Empirical, and others. Decisions concerning the specific distributions came from input modeling the distribution of the independent variables in the real data set. The real data was the basis for choices of distributions and parameters. Consideration of a real variable's mean, variance, and distribution led to choices for characteristics of the simulated data. It is important to note that in this example the research was not focused on other data problems, like multicollinearity or outliers, so we could proceed to the next step. However, if these factors were an issue of concern, they should be built into the data set at this point. The data should be created with the objective of reasonableness in a real world problem, not of having the exact and specific characteristics of a data set under consideration.

Once the set of variables was generated, the next step is generation of the response. Eleven arbitrary variables were randomly selected to generate the true model; in addition, each of these variables was given an arbitrary coefficient value. These 11 variables did include a mix of both categorical and continuous variables. The true model was: $G(x) = 5C_{12} - 2A_1 + 3A_2 + 1.5J + 1.9I + .98L - 1.9D_2 + 1.1D_3 + 1.3M - 2.01N + 2.9P$.

The next step in forming the response is the modeling of the error term. In this case, the base of the response was formed with the equation above, and it resulted in a range of the response of approximately 32,000 and a mean response of 1141. The characteristics of the base response were compared to the benchmark cases of Table 1 below.

Although there is not an example that provides a range of response as large as the base for this particular simulated data set, the Mean Y was between the Hospital Data and the SAT Test Model. Our arbitrary selection of true significant independent variables and coefficients resulted in a wide response range. The analyst can reduce the range of response by transformation of the response or changing some of the coefficients used to build the models, or the analyst can proceed to find a reasonable error for the wide range.

Table 1: Linear Response Comparison

Example	Min Y	Max Y	Range	Mean Y
Transistor Gain	852	1636	784	1250
Plating Process	50	790	740	376
Hospital Data	567	18854	18287	4979
Teacher Effectiveness	235	584	349	445
Cleansing Experiment	167	410	243	302
SAT Test Model	590	1430	840	1076

The next step will be to compare the error terms of similar examples to find a level appropriate for the simulated data. Table 2 provides the benchmark metrics for selection of an appropriate error level. In this case, we are looking for a level of error to support the wide range of the response described above. An arbitrary selection of a normally distributed error term with mean zero and standard deviation of 177 resulted in an MSE of 31,000. Upon inspection, it seems a target MSE level of 31,000 is reasonable. This MSE is slightly more than average of the benchmark cases shown in Table 2, but the larger range of response makes this seem logical. The MSE value of 31,000 also falls between the Hospital Data and the SAT Test Model, and this also seems reasonable.

Table 2: Comparison of Linear Metric Errors

Example	MSE	ADJ R^2	$\frac{\sqrt{MSE}}{Range}$	$\frac{MSE}{\ln(Rnge)}$
Transistor Gain	1220	0.973	0.045	183
Plating Process	971	0.983	0.042	147
Hospital Data	412274	.0987	0.035	42009
Teacher Effectiveness	3023	0.591	0.157	516
Cleansing Experiment	363	0.928	0.078	66
SAT Test Model	25472	0.809	0.144	2185

The decision on a level of error fuels the generation of the response and remainder of the data set. In this case, we added a normally distributed error with mean zero and standard deviation of 177 to the base equation shown above. If an appropriate level of error is not clear by using the benchmark approach, decide on a range of possible errors. Simulated data can then be generated and tested against a specific analytical application to determine the effects of error on the analysis. This additional analysis of error levels leads to a more accurate decision concerning the error term. An example of this approach is detailed in Hill and Malone (2003).

After specifying the independent variables, the response, and the errors one can consider altering the data so that it violates assumptions about the error, for example non-normally distributed errors. Another data adjustment might be systematic inclusion of some level of multicollinearity, which would also need to be benchmarked as shown in Appendix A. At this point, the data is ready for analysis. The appropriate numbers of observations must be generated, the error must be generated, and the response calculated by adding the error term to the base regression equation. Replication of the simulated data that capitalizes on different random number schemes is straightforward once the metrics have been established.

2.4 Logistic Regression

Logistic regression analysis is much like linear regression in that we are interested in the relationship of a group of independent variables with a response or dependent variable. Much like in linear regression, the ultimate objective for the study may be either estimation of the coefficient values, or prediction of the response value. One significant difference between the logistic and linear models is that the linear model has a continuous response variable and the logistic model uses a binary or dichotomous response. As a result, the method of estimation uses maximum likelihood as opposed to least squares (Hosmer and Lemeshow 2000). This difference is substantial and brings a different set of issues and statistics to bear on the simulation of the logistic regression problem. The example problems used as benchmark cases for the logistic model come from the Hosmer and Lemeshow (2000) logistic regression text; Myers (1990) regression text; Neter, Kutner, Nachtsheim, and Wasserman (1996) regression text; and the Studenmund (2001) econometrics text. These examples provide a variety of cases including medical, environmental, task success, consumer purchase, and financial applications. As in the linear example, appendices provide the characteristics of problems in the benchmark cases as a guide for construction of the simulated data. Appendix B provides characteristics of variables from the benchmark cases and references. Appendix C gives characteristics of the models. The characteristics and statistics provided in the appendices are meant to be a guide for creating data sets that are realistic in the sense that they are representative of a real problem, just as in the linear regression case.

2.5 Logistic Regression Example

The same USAREC application can be used for creation of a benchmark for the logistic model. The initial portion of creating the independent variables is no different from the linear case—ensure the numbers, types, and distributions

of the independent variables are feasible. The use of 37 variables, 10 of which are continuous and 26 of which are categorical, were created the same way. The distributions used to generate the X's are the same as well. The point of departure for the logistic regression example is generation of the response.

Although the same procedure is used for determination of the "true" significant independent variables and their coefficient values, i.e., random selection, the rest of the process is significantly different. The remaining process for creating the logistic regression model is accomplished by calculating the logit, estimating the conditional mean of Y given X, estimating the base response value, comparing the base value to a random error, and then calculating the simulated response value. The values of the coefficients used in a simulation of a logistic regression do not have the same meaning as in the linear problem and as a result will be smaller than those typically encountered in a linear regression problem. Continuing with this example, the logit equation is determined by: $G(x) = -0.07A + 0.11B_1 - 0.09B_2 - 0.15D + 0.08G - 0.03K + 0.19L_1 - 0.06L_3 + 0.25M - 0.06N - 1.1S$. The equation is different from our linear regression example in that we have different independent variables selected as truly significant, and they have different coefficient values. The next step is to use the G(x) value to estimate the conditional mean value of Y given X. This is given by :

$$\pi(x) = \frac{e^{G(x)}}{1 + e^{G(x)}} \qquad (1)$$

which will range in value between 0 and 1 and represents the probability of observing a 1 response given the group of X's. Although this value could be used to estimate a response value, it should consider the realistic amount of error that will be present in real applications.

The first step in generating the error is to randomly generate an error term, distributed between 0 and 1. In our application, we generated this term with a normal distribution. The parameters of the distribution will affect the amount of error present in the final simulated data. Once the error term is generated, the simulated estimated conditional mean value is compared to the error term value. If it is greater than the error term, the response is coded as a 1; otherwise, the response is coded as zero.

However, consideration of the level of error is not straightforward. As mentioned earlier, the distribution and parameters used for generating the error term directly affect the amount of error that is present in the simulated logistic regression model. A prediction based error measurement will give a feel for the impact of the amount of error present in a simulation model. To do this, compare the calculated response value, which was used to fit the logistic regression model, to the predicted response values from the model, and determine the difference as a prediction-based error assessment. If all other variable values and parameters, except for the error variable, are kept the same, this measurement will provide a sense of the impact of the error. This level of simulation model error is compared to a similar prediction error level in the benchmark cases. In the case of the benchmark examples, a logistic regression model was fitted to the real data. The estimated values were used to predict a response value by using a cut-off score method (Hosmer & Lemeshow 2000). The benchmark errors were obtained by building a model that included all terms under consideration without any subsequent model improvement. Although not exact, this provided a sense of impact of the error that is realistic. The predicted response was then compared to the real data in the benchmark cases to obtain an error rate. The simulation model error rates were then compared to the benchmark error rates, and the level of error in the simulation model was adjusted until the error level fell within the range of the benchmark data. Table 3 below shows varying levels of error in simulation models against the benchmark cases shown in Appendices B and C. In these appendices for the cases of simulated logistic regression models all factors are held constant except for the level of error.

Table 3: Comparison of Error in Simulated and Benchmark Logistic Regression Models

Simulation Case	Error Term	Sim. Error Rate
1	U(0, 1)	0.048
2	N(0, 1.03)	0.439
3	N(0, 1.5)	0.449
4	N(0.5, 0.6)	0.278
Benchmark Case		Bench Error Rate
Infant Birth Weight		0.317
Prostate Cancer		0.102
Average Benchmark		0.218

2.6 Experimentation

Once the process of constructing the simulated data is completed the research can begin. Simulation allows for a unique approach for isolating effects of the random numbers alone before analysis of the other factors begins. This step is accomplished through a noise model and allows quantification of the random effects in a given model that are due to spurious patterns in the random numbers alone. This approach applied to regression models is shown in Freedman (1983), Raftery (1995), and Hill and Malone (2003). As the experimentation continues to determine effects of other factors, it is important to remain aware that creating a simulation model that is too clean in general or choosing an error rate that is too conservative can provide misleading results. An example of this issue is also found in Hill and Malone (2003). In these cases, the models will perform better than the "real" conditions and may lead to overestimation of the performance of the technique under study.

As in other simulation approaches, analysis of competing alternatives, or different factors, can be accomplished by using common random numbers and the resulting group of independent variables, error, and response. This approach allows for variance reduction in general, and isolation of effects that are due to differences in alternatives. Once a hypothesis is formed, it is also possible in simulation experimentation to use varied streams of random numbers, and resulting variables and error, to determine the effect of random numbers on the hypothesis. Both of these abilities provide simulation-based analysis with information that is not available under a standard real data approach. This fosters more accurate and robust results.

3 CONCLUSIONS

Using simulation to create data that serves as a foundation for research of diagnostic tools for regression analysis is powerful. It provides opportunities for analyses that do not exist when using real data alone. The simulated data gives the analyst greater control over the experimental environment and allows measurement and comparisons that are not achievable with real data. In general, simulated data provides more robust and defendable solutions. However, generation of the simulated data has tremendous affects on results. Models that are either too clean and well behaved or are unrealistic with respect to error and other troublesome real world characteristics can provide misleading results. The use of benchmark examples from historic studies, real data, and generally accepted foundational literature to construct the simulated data will help avoid these concerns and lead to more successful and meaningful research.

APPENDIX A: LINEAR REGRESSION BENCHMARK CHARACTERISTICS

Example	Source	Eigen-ratio X'X	Eigen-range X'X	Largest VIF
Transistor Gain	Myers & Montgomery (1995)	1.005	0.005	1
Plating Process	Myers & Montgomery (1995)	1.25	0.2143	1
Hospital Data	Myers (1990)	108385	43.35	9595
Teacher Effectiveness	Myers (1990)	5.49	1.46	1.8
Cleansing Experiment	Myers (1990)	3.06	1.01	1.3
SAT Test Model	Studenmund (2001)	30.15	2.95	5.1

APPENDIX B: CHARACTERISTICS OF LOGISTIC REGRESSION VARIABLES

Description	Source	n	Proportion 1's in Y	Total IVs
Infant Birth Weight	Hosmer & Lemeshow (2000)	189	.0312	4
Prostate Cancer	Hosmer & Lemeshow (2000)	376	0.402	9
Air Restriction	Myers (1990)	39	0.513	2
Mosquito Disease	Neter, Kutner, Nachtsheim, & Wasserman (1996)	98	0.316	4
Mortgage Rates	Studenmund (2001)	78	0.410	6

APPENDIX C: CHARACTERISTICS OF LOGISTIC REGRESSION MODELS

Example	Log Likelihood	Deviance	Cat. IVs	Error Rate	Pseudo R^2
Infant Birth Weight	-111.29	12.099	2	0.317	0.052
Prostate Cancer	-187.27	132	5	0.103	0.261
Air Restriction	-14.89	24.27	0	0.103	0.449
Mosquito Disease	-50.53	21.26	3	0.286	0.174
Mortgage Rates	-39.06	27.48	0	0.218	0.260

ACKNOWLEDGMENTS

The authors would like to thank Kevin Lyman, Gene Paulo, Mansooreh Mollaghasemi, Chuck Reilly, and Linda Trocine for their input, review, and assistance.

REFERENCES

Foster, D. P. & Stine, R. A. 2001. Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy (01-05). The Wharton Financial Institutions Center, Department of Statistics, The Wharton School of the University of Pennsylvania.

Freedman, D. A. 1983. A Note on Screening Regression Equations. The American Statistician, 37 (2) : 152-155.

Hill, C. & Malone, L. 2003. Issues With Linear Regression Analysis In Large Data Sets in Large Data Sets. Manuscript submitted for publication.

Hosmer, D. W. & Lemeshow, S. 2000. Applied Logistic Regression (2nd ed.). New York, John Wiley & Sons, Inc.

John, G. H., & Langley, P. 1996. Static Versus Dynamic Sampling For Data Mining. In *Proceedings, Second International Conference on Knowledge Discovery & Data Mining,* ed. Simoudis, E., Han, J., & Fayyad, U., 367-370. Menlo Park: AAAI Press.

Law, A. M. & Kelton, W. D. 2000. *Simulation Modeling and Analysis* (3rd ed.). Boston: McGraw-Hill Companies Inc.

Morrison, A. M., Bose, G. & O'Leary, J. T. 2000. Can Statistical Modeling Help with Data Mining? A Database Marketing Application for U.S. Hotels. *Journal of Hospitality and Leisure Marketing*, 6 (4): 91-110.

Myers, R. H. 1990. *Classic and Modern Regression With Applications* (2nd Ed.). Boston: PWS-KENT Publishing Company.

Myers R. H. & Montgomery, D. C. 1995. *Response Surface Methodology, Process and Product Optimization Using Designed Experiments*. New York: John Wiley & Sons, Inc.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. 1996. *Applied Linear Regression Models* (2nd Ed.). Chicago: Irwin.

Pampel, F. C. 2000. *Logistic Regression, A Primer*. Sage Publications, Thousand Oaks.

Raftery, A. E. 1995. Bayesian Model Selection in Social Research, *Sociological Methodology*, 25: 111-163.

Reinartz, T. 1998. Similarity-Driven Sampling For Data Mining. In *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98*, Nantes, France, September 1998, Proceedings, 423-431. New York: Springer.

Salzberg, S. L. 1997. Methodological Note, On Comparing Classifiers: Pitfalls To Avoid And A Recommended Approach. *Data Mining and Knowledge Discovery*, 1: 317-328.

Schaeffer, R. L., & McClave, J. T. 1990. *Probability and Statistics for Engineers* (3rd Ed.). Belmont: Duxbury Press.

Shirata, C. Y., & Terano, T. 2000. Extracting Predictors Of Corporate Bankruptcy: Empirical Study On Data Mining Methods. In *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asian Conference, PAKDD 2000* Kyoto, Japan, April 18-20, 2000, Proceedings, ed. Terano, T., Liu, H., & Chen, A. L. P., 204-207, Berlin: Springer-Verlag.

Skipper, J. K., Jr., Guenther, A. L. & Nass, G. 1967. The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science. In Morrison, D. E. & Henkel, R. E. (Eds.) *The Significance Test Controversy—A Reader*, 1970, 155-160. Aldine Publishing Company, Chicago.

Studemond, A. H. 2001. *Using Econometrics, A Practical Guide* (4th ed.). Boston: Addison Wesely.

AUTHOR BIOGRAPHIES

CHRISTOPHER M. HILL is an analyst at the Center for Army Analysis at Fort Belvoir, Virginia. He is a Lieutenant Colonel in the United States Army. This work is part of his dissertation research on the impact of large numbers of observations on traditional regression methods. He can be contacted by e-mail at <Christopher.Hill@caa.army.mil>.

LINDA C, MALONE is a Professor in the Industrial Engineering and Management Systems Department at the University of Central Florida. Her research interests include experimental design, large data analysis, and simulation output analysis. You can reach her by email at <lmalone@mail.ucf.edu>.