Data Quality in Linear Regression Models: Effect of Errors in Test Data and Errors in Training Data on Predictive Accuracy

Barbara D. Klein & Donald F. Rossin
University of MIchigan-Dearborn USA

bdklein@som.umd.umich.edu & drossin@som.umd.umich.edu

Abstract

Although databases used in many organizations have been found to contain errors, little is known about the effect of these errors on predictions made by linear regression models. The paper uses a real-world example, the prediction of the net asset values of mutual funds, to investigate the effect of data quality on linear regression models. The results of two experiments are reported. The first experiment shows that the error rate and magnitude of error in data used in model prediction negatively affect the predictive accuracy of linear regression models. The second experiment shows that the error rate and the magnitude of error in data used to build the model positively affect the predictive accuracy of linear regression models. All findings are statistically significant. The findings have managerial implications for users and builders of linear regression models.

Keywords: Data Quality, Errors, Linear Regression

Introduction

There is strong evidence (e.g., Laudon, 1986; Morey, 1982; Redman, 1992, 1995, 1996) that data stored in organizational databases have a significant rate of errors. The effect of data errors on the outputs of computer-based models has been investigated by a number of researchers (e.g., Ballou and Pazer, 1985; Ballou et al., 1987; Bansal et al., 1993). This investigation builds on this prior research by examining the effect of data quality on linear regression models. A financial application of a linear regression model is used to examine this question.

Data errors may affect the predictive accuracy of linear regression models in two ways. First, the training data used to build the model may contain errors. Second, even if training data are free of errors, once a linear regression model is used for forecasting a user may input test data containing errors to the model.

In general, when claims about the predictive accuracy of linear regression models are made, it is assumed that data used to train the models and data input to make predictions are free of errors. In this study we relax this assumption by asking two questions: (1) What is the effect of errors in test data on predictions made using linear regression models? and (2) What is the effect of errors in training data on predictions made using linear regression models? The first question is focused on the effect of data errors when the model is used for forecasting. The second question is focused on the effect of data errors during model construction.

An understanding of the effect of data errors on linear regression models is particularly important because the availability of inexpensive software packages for personal computers makes the development and use of linear regression models by end-users feasible. Researchers have argued that end-user computing has increased the potential for data errors in computer applications (Boockholdt, 1989). As end users develop applications, it is possible that fewer data validation methods such as logic tests and control totals will be in place and it is likely that less rigorous testing will occur before applications are used in production (Corman, 1988; Davis, 1984; Davis et al., 1983; Panko, 1998).

The remaining sections of this paper present (1) a review of relevant prior research on data quality, (2) a brief explanation of linear regression models, (3) a description of the linear regression models constructed in the study, (4) a discussion of the methodol-

ogy of two experiments, (5) the results of two experiments and (6) conclusions.

Background

Data quality is generally recognized as a multidimensional concept (Wand and Wang, 1996; Wang and Strong, 1996). While no single definition of data quality has been accepted by researchers working in this area, there is agreement that data accuracy, currency, completeness, and consistency are important areas of concern (Agmon and Ahituv, 1987; Ballou and Pazer, 1985; Davis and Olson, 1985; Fox et al., 1993; Huh et al., 1990; Madnick and Wang, 1992; Wand and Wang, 1996; Wang and Strong, 1996; Zmud, 1978). This investigation adopts the conceptualization of data quality proposed by Ballou and Pazer (1985) that includes four dimensions: accuracy, timeliness, completeness, and consistency. This study is primarily concerned with data accuracy, defined as conformity between a recorded data value and the corresponding actual data value.

Prior research has found that organizational databases are not in general free of errors (e.g., Laudon, 1986; Morey, 1982; Redman, 1992, 1995). Between one and twenty percent of data items in critical organizational databases are estimated to be inaccurate (Laudon, 1986; Madnick and Wang, 1992; Morey, 1982; Redman, 1992).

Data quality problems have been found to affect the accuracy and timeliness of economic data published by the United States government (Hershey, 1995; Morgenstern, 1963). Both Standard & Poors Compustat® (with its Price Earnings Dividend tape) and the Center for Research in Security Prices (with its monthly stock return CRSP tape) sell a data base containing monthly price information. Two studies (Bennin, 1980; Resenberg and Houglet, 1974) found large errors possible in each database. Inaccurate data have also been reported in a student loan database maintained by the U.S. Department of Education (Knight, 1992), in records maintained by the U.S. Department of Agriculture (Dead farmer, 1992), and in records maintained by credit reporting bureaus (Consumer enemy, 1991).

Errors in data are acknowledged as a significant problem by at least some information system managers. In a survey of fifty Chief Information Officers of large organizations, half were found to believe that the usefulness of their organization's data is limited because of data accuracy problems (Nayar, 1993). Knight (1992) reports the findings of a study in which two-thirds of surveyed organizations acknowledged problems stemming from inaccurate or incomplete data.

Several studies have investigated the effect of data errors on the outputs of computer-based models. Bansal et al. (1993) compared the effects of errors in test data on a linear regression and a neural network model and found the neural network model to be more robust than the linear regression model as data quality decreases. Ballou and Pazer (1985) present a model for analyzing the effect of errors in data on the outputs of information systems. Ballou et al. (1987) applied the model to a forecasting task and found data errors to have a strong effect on the selection of a forecasting model. In other studies, Ballou and his colleagues have examined the allocation of resources to data quality improvement projects (Ballou and Tayi, 1989), developed a framework for analyzing tradeoffs between the accuracy and timeliness dimensions of data quality (Ballou and Pazer, 1995), and developed a framework applying total quality management concepts to the measurement of data quality (Ballou et al., 1998). O'Leary (1993) analyzed the impact of data accuracy on the performance of an artificial intelligence system designed to generate rules from data stored in a database and found that inappropriate rules may be retained while useful rules are discarded if data accuracy is ignored.

Linear Regression Models

Linear regression is a statistical tool for modeling the relationship between a dependent variable and one or more independent variables. In linear regression models, the dependent variable is a linear function of one or more independent variables as shown in the equation below.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

The parameters of the linear regression model are typically estimated using the least-squares method which results in a line that minimizes the sum of squared vertical distances from the observed data points to the line (Lewis-Beck, 1980; Neter et al., 1990).

Practitioners have been found to be very familiar with linear regression and to employ it as a forecasting tool in tasks such as sales prediction (Mentzer and Cox, 1984; Sanders, 1994). Linear regression is also a recognized forecasting tool for financial applications (Bansal et al., 1993; Chiang et al., 1996; Cole, 1994; Jabbour, 1994; Jankus, 1997; Mark, 1995; Refenes et al., 1994).

Most applications of linear regression models assume that all data used to construct the model and all data input to the model in production are accurate. The remaining sections of this paper present the design and results of an investigation into the performance of linear regression models when this assumption is relaxed.

Model Construction

The application for study in this paper is the prediction of prices or net asset values (NAV) of mutual funds. Mutual funds consist of diversified portfolios of stocks that are managed by professionally trained individuals. They have become the major investment vehicle of choice.

The prediction of NAV of mutual funds was selected as the application domain for examining the research questions for two reasons. First, prior research shows that NAV can be predicted with a reasonable level of error (Chiang et al., 1996). Since the objective in this study is to compare NAV predictions made with data containing no errors to NAV predictions made when data are perturbed, the most important criterion for selecting an example application domain is that predictions made with input data that are free of errors are reasonably good. The prediction of NAV meets this test.

Second, prior research provides insight into a set of relevant input variables that predict the NAV of mutual funds (Chiang et al., 1996). Recent studies (Balvers et al., 1990; Breen et al., 1990; Campbell, 1987; Cochrane, 1991; Fama and French, 1989; Ferson and Harvey, 1993; French et al., 1987; Glosten et al., 1993; Pesaran and Timmermann, 1994, 1995) show that economic variables can be used to predict stock returns. As mutual funds are simply groupings of stocks, prices or net asset values (NAV) of mutual funds should reflect known economic information. Economic variables such as gross national product and the consumer price index have been used as exogenous variables in prior research on the prediction of the NAV of mutual funds (Chiang et al., 1996).

To start the construction of a linear regression model for predicting the net asset value for a mutual fund, 14 economic variables were identified as input. They are specified and defined in Figure 1. A 10-year economic data set (1986-1995) was constructed (Statistical Abstract, 1996). In addition, end-of-year net asset values for 213 U.S. mutual funds were obtained (Individual Investor's Guide, 1997). The criteria for inclusion were having historical net asset value figures back to 1987.

As the purpose of this research is to study is the effect of data quality on linear regression forecasting, it was decided to limit the number of input variables to a more manageable amount. Stepwise linear regression was conducted for the 213 mutual funds to limit the number of input variables. A 5 percent significance level (the SPSS default) was used to bring variables into the models. Four input variables were chosen based on the number of times each had been selected in the regression step. An asterisk in Figure 1 identifies these variables. In addition, it was decided to limit the number of mutual funds to 10 per fund type. Fund type definitions are per *The Individual Investor's Guide* (1997). The randomly chosen 40 funds are indicated in Figure 2.

For construction of the linear regression models, the first nine years of data (the *training* set) are used. Data from the tenth year (the *testing* set) are used to develop the NAV forecast for a specific mutual fund.

Name	Description
GDP	Gross Domestic Product (in billions of dollars). Output attributable to all labor and property supplied by United States residents.
CD*	Consumption Demand (in billions of dollars). Personal consumption expenditures.
ID	Investment Demand (in billions of dollars). Investment spending by firms. Excludes residential investments.
GD*	Government Demand (in billions of dollars). U.S. government spending. Includes consumption expenditures and gross investment.
NEX	Net Exports (in billions of dollars). Net exports of goods and services.
CPI*	Consumer Price Index. Measure of the average change is prices over time in a fixed market basket of goods and services. 1982-84 = 100.
M1*	Money, M1 (in billions of dollars). Includes currency in the hands of the nonbank public, travelers checks, demand deposits, and other checkable deposits.
M2	Money, M2 (in billions of dollars). Includes M1 plus money market funds, savings deposits, and small time deposits.
UR	Unemployment Rate. Percent of the labor force unemployed.
TBR	Treasury Bill Rate. Interest rate for 3-month Treasury bill.
FFR	Federal Funds Rate.
CILEAD	Composite Index - Leading Indicators. 1987 = 100.
CICOIN	Composite Index - Coincident Indicators. 1987 = 100.
CILAG	Composite Index - Lagging Indicators. 1987 = 100.

Note: Asterisk indicates selection for model development.

Figure 1: Potential Independent Variables

Aggressive Growth (out of 64 possible)	Balanced (out of 24 possible)	Growth (out of 80 possible)	Growth & Income (out of 45 possible)
Fairmont	Dodge & Cox Balanced	Fidelity Capital Appreciation	AARP Growth & Income
Fidelity Sel Air Transportation	Fidelity Puritan	Fiduciary Capital Growth	Berger Growth & Income
Fidelity Sel Automotive	Founders Balanced	Founders Growth	Dreyfus Third Century
Fidelity Sel Brokerage & Investment	Greenspring	Janus Fund	Fidelity Sel Utilities Growth
Fidelity Sel Computers	INVESCO Industrial Income	Mathers	IAI Growth & Income
Fidelity Sel Leisure	Northeast Investors Trust	Meridian	INVESCO Value: Value Equity
Fidelity Sel Software & Computer	SAFECO Income	Schwartz Value	SAFECO Equity
INVESCO Dynamics	Strong Asset Allocation	Scudder Equity Trust: Capital Growth	Stratton Monthly Dividend Shares
Kaufmann	USAA Income	Sound Shore	Strong Total Return
USAA Aggressive Growth	Value Line Income	Vanguard/Morgan Growth	T. Rowe Price Growth & Income

Figure2: Randomly Chosen Mutual Funds

Again, a separate linear regression model was constructed for each of the 40 mutual funds using the 9 oldest years of the data for training. The 1995 testing data was then input to the appropriate linear regression model to predict a NAV value for each of the 40 mutual funds for the end-of-year 1996. Actual end-of-year 1996 NAV values and predicted end-of-year NAV values were compared using mean absolute percent error (MAPE) as a measure of accuracy. This comparison formed the base case.

Experimental Methodology

Two experiments were conducted to examine the research questions. Both experiments used the same task (the prediction of NAV of mutual funds), the same dataset, and the same dependent variable. The experimental factors are the same in both experiments, although the levels of the factors are different.

Experiment 1 examines the first research question: What is the effect of errors in test data on predictions made using linear regression models? Experiment 2 examines the second research question: What is the effect of errors in training data on predictions made using linear regression models?

Experimental Dataset

An example dataset for one of the mutual funds used in both experiments is depicted in Figure 3. The training data contains 36

	Year for Economic Variables	Economic Variables				NAV for Fairmont	Year for NAV Variable
		CD	GD	CPI	M1		
Training Data	1986	2892.7	938.5	109.6	724	14.96	1987
	1987	3094.5	992.8	113.6	750	15.19	1988
	1988	3349.7	1032.0	118.3	787	16.02	1989
	1989	3594.8	1095.1	124.0	794	12.17	1990
	1990	3839.3	1176.1	130.7	826	17.02	1991
	1991	3975.1	1225.9	136.2	897	19.41	1992
	1992	4219.8	1263.8	140.3	1024	22.43	1993
	1993	4454.1	1289.9	144.5	1129	24.06	1994
	1994	4698.7	1314.7	148.2	1149	27.02	1995
Test Data	1995	4924.3	1358.5	152.4	1125	26.45	1996

Figure 3: Example Base Dataset for Fairmont Mutual Fund

data items (four economic variables in the columns by nine years in the rows). The test data contains four data items (four economic variables by one year).

Experimental Factors

There are two factors in each experiment: (1) fraction-error and (2) amount-error. Fraction-error is the percent of the data items in the appropriate part of the dataset (the test data in experiment 1 and the training data in experiment 2) that are perturbed. Amount-error is the percent by which the data items identified in the fraction-error factor are perturbed.

1. Fraction-error. Since fraction-error is defined as a percent of the data items in a dataset, the number of data items that are changed for a given level of fraction-error is determined by multiplying the fraction-error by the total number of data items in the dataset.

Experiment 1. The test data used in experiment 1 contain four data items (one value for each of the four economic variables for 1995). This experiment examines all of the possible number of data items that could be perturbed. These four levels for the fraction-error factor are: 25 percent (1 data item perturbed), 50 percent (2 data items perturbed), 75 percent (3 data items perturbed), and 100 percent (4 data items perturbed).

Experiment 2. The training data used in experiment 2 contains 36 data items (one value for each of the four economic variables for nine years). Four levels of the fraction-error factor are tested: 5 percent (2 data items perturbed), 10 percent (4 data items perturbed), 15 percent (5 data items are perturbed), and 20 percent (7 data items are perturbed).

2. Amount-error. For both experiments, the amount-error factor has two levels: (1) plus or minus 5 percent and (2) plus or minus 10 percent.

Experimental Design

The experimental design is shown in Figure 4. Both experiments have four levels for the fraction-error factor and two levels for the amount-error factor. For each combination of fraction-error and amount-error, four runs with random combinations of economic variables were performed for each of the 40 randomly chosen mutual funds. This gives a total of 1,280 runs for each experiment.

Although the levels of the fraction-error factor are different in the two experiments, the sampling procedure is the same. For each fraction-error level, economic variables were randomly selected to be perturbed. This was repeated a total of four times per level. Figure 5 shows the results for experiment 1.

Next, for each level of the amount-error factor, each economic variable was randomly assigned either a positive or negative sign to indicate the appropriate amount-error to be applied. Figure 6 shows the results for experiment 1. The procedure for experiment 2 differs only in the number of economic variables that were randomly selected to be perturbed for the four tested levels of the fraction-error factor.

Experiment 1 (Errors in Test Data):	
Experimental Factors	
Fraction-error levels (25%, 50%, 75%, 100%):	4
Amount-error levels (5%, and 10%):	x 2
Sampling Procedure	
Number of random combinations of economic variables considered within each fraction-error level:	x 4
Number of mutual funds:	x 40
Total number of problems considered:	= 1,280
Experiment 2 (Errors in Training Data):	
Experimental Factors	
Fraction-error levels (5%, 10%, 15%, 20%):	4
Amount-error levels (5%, and 10%):	x 2
Sampling Procedure	
Number of random combinations of economic variables considered within each fraction-error level:	x 4
Number of mutual funds:	x 40
Total number of problems considered:	= 1,280

Figure 4: Experimental Design

Fraction -Error Level	Economic Variable Combination			
	1	**2**	**3**	**4**
25%	(CD)	(CPI)	(GD)	(M1)
50%	(CD, GD)	(CD, M1)	(GD, CPI)	(GD, M1)
75%	(CD, CPI, GD)	(CD, GD, M1)	(CD, GD, M1)	(CPI, GD, M1)
100%	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)

Figure 5: Four Combinations of Economic Variables for Each Fraction-Error Level in Experiment 1

Dependent Variable

In both experiments, actual end-of-year 1996 NAV values and predicted end-of-year 1996 NAV values were compared using mean absolute percent error (MAPE) as a measure of accuracy.

Experimental Results

For both experiments, MAPE results for each combination of fraction-error and amount-error are presented. The results of ANOVA tests and independent samples t-tests conducted to test for the effect of fraction-error and amount-error on MAPE are then discussed. Finally, the findings of tests performed to determine which combinations of fraction-error and amount-error are significantly different than the base case scenario with no data errors are reported.

Experiment 1 Results: Errors in Test Data

Predictive accuracy results, using the simulated inaccuracies for amount-error and fraction-error for the NAV forecasts for 1996 are given in Table 1. Each cell reflects the average MAPE value for 160 estimations (four runs for 40 mutual funds).

Table 1 shows that in general as fraction-error increases from 25 percent to 100 percent, MAPE increases indicating a decrease in predictive accuracy. When amount-error is equal to five percent, MAPE decreases as fraction-error increases from 75 percent to 100 percent. As amount-error increases from 5 percent to 10 percent, MAPE increases also indicating a decrease in predictive accuracy.

A two-factor analysis of variance (ANOVA) test was conducted to test for the effect of the independent variables on MAPE. The independent variables are fraction-error (25 percent, 50 percent, 75 percent, and 100 percent) and amount-error (plus or minus 5 percent, and plus or minus 10 percent).

Table 2 presents the results of the ANOVA test. Significant main effects for fraction-error and amount-error were found ($p<.05$). These results indicate that both fraction-error and amount-error have an effect on predictive accuracy.

When there are more than two levels of a factor, ANOVA results do not indicate where the significant differences occur. For example, while fraction-error is a significant factor, this difference may come as fraction-error changed from 25 percent to 50 percent, 50 percent to 75 percent, or 75 percent to 100 percent. It could also have come from a larger jump, such as 25 percent to 75 percent or 25 percent to 100 percent. Independent samples t-tests were performed in order to determine exactly where significant differences occurred. For the 5 percent amount-error, significant differences

Fraction-Error Level	Economic Variable Combination			
	1	2	3	4
25%	(CD)	(CPI)	(GD)	(M1)
	+	-	+	+
50%	(CD, GD)	(CD, M1)	(GD, CPI)	(GD, M1)
	+, +	-, +	-, +	+, +
75%	(CD, CPI, GD)	(CD, GD, M1)	(CD, GD, M1)	(CPI, GD, M1)
	+, -, -	-, -, -	+, +, -	+, -, +
100%	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)	(CD, CPI, GD, M1)
	-, -, -, +	+, +, -, -	+, -, +, +	-, -, -, -

Figure 6: Randomly Assigned Percentage Increase (+) Over Base Value or Decrease (-) for a Given Amount-Error Level in Experiment 1

($p < .05$) were found between fraction-errors of 25 percent and 75 percent, 25 percent and 100 percent, and 50 percent and 75 percent. For the 10 percent amount-error, significant differences ($p < .05$) were found between fraction-errors of 25 percent and 50 percent, 25 percent and 75 percent, 25 percent and 100 percent, and 50 percent and 100 percent.

The ANOVA results indicate that there are differences in predictive accuracy at different levels of fraction-error and amount-error. However, they do not show which combinations of fraction-error and amount-error have MAPE significantly different than the base case scenario with no data errors. We constructed confidence intervals around the means shown in Table 1 for the experimental conditions to determine which values are significantly different than the base case scenario with MAPE of 16.8 percent. Combinations of fraction-error and amount-error with MAPE different than the base case scenario at a level of significance of .05 are identified with an asterisk in Table 1. When amount error is equal to 5 percent, the scenarios with fraction-error equal to 50 percent, 75 percent, and 100 percent have MAPE significantly higher ($p < .05$) than the base case. When amount-error is equal to 10 percent, the scenarios with fraction-error equal to 25 percent, 50 percent, 75 percent, and 100 percent have MAPE significantly higher ($p<.05$) than the base case scenario.

	Fraction Error				
Amount Error	**0% (0 errors)**	**25% (1 error)**	**50% (2 errors)**	**75% (3 errors)**	**100% (4 errors)**
0%	16.8				
5%		21.4	26.8 *	43.3 *	34.8 *
10%		26.7 *	43.7 *	50.6 *	58.4 *

Notes:

(1) Data used to obtain these results were the test data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified linear regression model. All other cells reflect average accuracy results for 4 simulated estimations involving appropriately simulated data inaccuracies for 40 funds.

(2) Entries marked with an asterisk are values different than the base case MAPE at a significance level of .05

Table 1: Experimental Results: MAPE Values as Accuracy of Test Data Varies

Experiment 2 Results: Errors in Training Data

Predictive accuracy results, using the simulated inaccuracies for amount-error and fraction-error for the NAV forecasts for 1996 are given in Table 3. Each cell reflects the average MAPE value for 160 estimations (four runs for 40 mutual funds).

Table 3 shows that when amount-error is equal to 5 percent, MAPE decreases indicating an increase in predictive accuracy as fraction-error increases. When amount-error is equal to 10 percent, (1) MAPE decreases indicating an increase in predictive accuracy as fraction-error shifts from 5 percent to 10 percent and (2) MAPE increases indicating a decrease in predictive accuracy as fraction-error shifts from 10 percent to 20 percent.

When fraction-error is equal to 5 percent and 10 percent, MAPE decreases as amount-error increases from 5 percent to 10 percent, indicating an increase in predictive accuracy. When fraction-error is equal to 15 percent, MAPE is nearly identical for the scenario with amount-error equal to 5 percent and the scenario with amount-error equal to 10 percent. When fraction-error is equal to

Factor/significance criterion	Predictive Accuracy
Fraction error $F(0.05;3;1272) = 2.60$	12.786 *
Amount error $F(0.05;1;1272) = 3.84$	19.008 *
Fraction error-amount error interaction $F(0.05;3;1272) = 2.60$	1.962

Significant results ($p < .05$) are marked with an asterisk (*).

Table 2: Significance of Varying Amount-Error and Fraction-Error on Predictive Performance – ANOVA Results for Varying Test Data

20 percent, MAPE increases as amount-error increases from 5 percent to 10 percent, indicating a decrease in predictive accuracy.

A two-factor analysis of variance (ANOVA) test was conducted to test for the effect of the independent variables on MAPE. The independent variables are fraction-error (5 percent, 10 percent, 15 percent, and 20 percent) and amount-error (plus or minus 5 percent, and plus or minus 10 percent).

Table 4 presents the results of the ANOVA test. An interaction effect was found between fraction-error and amount-error, and a main effect was found for fraction-error ($p<.05$). The interaction between fraction-error and amount-error is viewed as an important interaction, and an analysis of the dependent variable suggests that a transformation of the variable is not appropriate (Neter et al., 1990). At lower levels of fraction-error (5 percent and 10 percent), predictive accuracy is best at the higher level of amount-error (10 percent). At the highest level of fraction-error (20 percent), predictive accuracy is best at the lower level of amount-error (5 percent).

When there are more than two levels of a factor, ANOVA results do not indicate where the significant differences occur. For example, while fraction-error is a significant factor, this difference may come as fraction-error changed from 25 percent to 50 percent, 50 percent to 75 percent, or 75 percent to 100 percent. It could also have come from a larger jump, such as 25 percent to 75 percent or 25 percent to 100 percent. Independent samples t-tests were performed in order to determine exactly where significant differences occurred. For the 5 percent amount-error, significant differences ($p < .05$) were found between fraction-errors of 5 percent and 10 percent, 5 percent and 15 percent, and 5 percent and 20 percent. For the 10 percent amount-error, no significant differences were found ($p < .05$).

The ANOVA results indicate that there are differences in predictive accuracy at different levels of fraction-error and amount-error. However, they do not show which combinations of fraction-error and amount-error have MAPE significantly different than the base case scenario with no data errors. We constructed confidence intervals around the means shown in Table 3 for the experimental conditions to determine which values are significantly different than the base case scenario with MAPE of 16.8 percent. Combinations of fraction-error and amount-error with MAPE different than the base case scenario at a level of significance of .05 are identified with an asterisk in Table 3. When amount-error is equal to 5 percent, the scenarios with fraction-error equal to 5 percent, 10 percent, 15 percent, and 20 percent have MAPE significantly lower than the base case scenario ($p < .05$). When amount-error is equal to 10 percent, the scenarios with fraction-error equal to 5 percent, 10 percent, 15 percent, and 20 percent have MAPE significantly lower than the base case scenario ($p < .05$).

	Fraction Error				
Amount Error	**0%** **(0 errors)**	**5%** **(2 errors)**	**10%** **(4 errors)**	**15%** **(5 errors)**	**20%** **(7 errors)**
0%	16.8				
5%		13.2 *	10.7 *	10.1 *	9.5 *
10%		11.0 *	9.9 *	10.2 *	12.0 *

Notes:

(1) Data used to obtain these results were the training data. The 0% fraction error and 0% amount error cell reflects the accuracy of the unmodified test data used in conjunction with the unmodified linear regression model. All other cells reflect average accuracy results for 4 simulated estimations involving appropriately simulated data inaccuracies for 40 funds.

(2) Entries marked with an asterisk are values different than the base case MAPE at a significance level of .05.

Table 3: Experimental Results: MAPE Values as Accuracy of Training Data Varies

Factor/significance criterion	Predictive Accuracy
Fraction error F(0.05;3;1272) = 2.60	3.042 *
Amount error F(0.05;1;1272) = 3.84	0.046
Fraction error-amount error interaction F(0.05;3;1272) = 2.60	3.812 *
Significant results ($p < .05$) are marked with an asterisk (*).	

Table 4: Significance of Varying Amount-Error and Fraction-Error on Predictive Performance – ANOVA Results for Varying Training Data

Conclusion

Two conclusions about the linear regression models built to predict the NAV of mutual funds can be drawn from this research. The first conclusion addresses the effect of errors in test data, and the second addresses the effect of errors in training data.

1. Errors in test data. For errors in test data, it is demonstrated that predictive accuracy decreases as the magnitude of errors (amount-error) increases and as the error rate (fraction-error) increases. All scenarios with data errors except the case of 25 percent amount-error and 5 percent fraction-error have predictive accuracy significantly worse than the base case scenario without data errors.

This finding is as expected. Decreases in data quality lead to poorer accuracy. This finding is also consistent with the work of Bansal et al. (1993) who found the predictive accuracy of a linear regression model built to predict the prepayment rate of mortgage-backed security portfolios to be affected by the fraction of the test data set containing errors and the size of errors in test data.

2. Errors in training data. For errors in training data, it is demonstrated that the predictive accuracy of a linear regression model built to forecast the NAV of mutual funds is better when errors exist in training data than when training data are free of errors. All of the scenarios with errors have predictive accuracy significantly better than the base case scenario without data errors. This is a significant contribution to the literature on data quality because this is the first research finding on the effect of errors in training data on linear regression models.

This finding demonstrates that perfectly accurate data may not always provide the best forecast. If the data point to be predicted does not fall along the general trend line of earlier data points, imperfect data may tilt the regression line such that a better prediction results. It is important to note that each economic variable selected to be perturbed in this experiment was randomly changed in either a positive or negative direction. Thus the improvement in predictive accuracy when the training data were perturbed does not stem from a systematic tendency to tilt the regression line in a particular fashion.

The findings of this study have implications for practitioners working in a variety of settings characterized by imperfect data. They suggest that an understanding of the error rate and the magnitude of errors in a dataset should be important considerations for users of linear regression models and that devoting resources to lowering the error rate in test data is likely to be beneficial.

The findings also suggest that the error rate of a dataset used to build a linear regression model should be an important consideration. The finding that lowering the error rate of training data can decrease predictive accuracy is of particular practical importance given the potential cost required to lower the error rate. Under some conditions, devoting resources to lowering the error rate of training data may be harmful.

Much of the literature on data quality assumes that improvements in data quality are always beneficial (e.g., Redman, 1992, 1995). The results of this study show that there is at least one case in which this assumption does not hold. Given the expenditures of time and money required to improve data quality in many organizations, this result merits further study. Research findings providing guidance to practitioners about the conditions under which it is not worthwhile to expend the resources required to improve data quality could be quite valuable.

Although it would be rash to rely on the findings of a single study as the basis for conclusions about the effect of errors on linear regression models in general, such conclusions may be drawn on the basis of a body of evidence collected through additional research. This study demonstrates that the outputs of one linear regression model are sensitive to data errors. The results suggest that additional studies examining the effect of data errors on the outputs of linear regression models in other application domains are worthwhile.

Until a body of evidence addressing the research questions across application domains has been built, we suggest that designers and users of linear regression models who are interested in understanding the relationship between data errors and predictive accuracy for a specific linear regression model follow the methodology outlined in this paper. We also suggest that a module for analyzing the effect of data errors be added to statistical analysis software packages so that users working in other domains can more easily understand the effect of data errors on their work.

Acknowledgements

The first author wishes to acknowledge support from a UMD School of Management summer research stipend. Both authors wish to thank the editor and anonymous reviewers for their constructive comments and suggestions that made this a stronger paper.

References

Agmon, N., & Ahituv, N. (1987). Assessing data reliability in an information system. *Journal of Management Information Systems*, **4**, 34-44.

Ballou, D., & Pazer, H. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, **31**, 150-162.

Ballou, D., & Pazer, H. (1995). Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research*, **6**, 51-72.

Ballou, D., Pazer, H., Belardo, S., & Klein, B. (1987). Implications of data quality for spreadsheet analysis. *Data Base*, **18**, 13-19.

Ballou, D., & Tayi, G. (1989). Methodology for allocating resources for data quality enhancement. *Communications of the ACM*, **32**, 320-329.

Ballou, D., Wang, R., Pazer, H., & Tayi, G. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, **44**, 462-484.

Balvers, R., Cosimano, T., & McDonald, B. (1990). Predicting stock returns in an efficient market. *Journal of Finance*, **45**, 1109-1128.

Bansal, A., Kauffman, R., & Weitz, R. (1993). Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. *Journal of Management Information Systems*, **10**, 11-32.

Bennin, R. (1980). Error rates in CRSP and COMPUSTAT: A second look. *The Journal of Finance*, **35**, 1267-1271.

Boockholdt, J. (1989). Implementing security and integrity in micro-mainframe networks. *MIS Quarterly*, **13**, 135-144.

Breen, W., Glosten, L., & Jagannathan, R. (1990). Predictable variations in stock index returns. *Journal of Finance*, **44**, 1177-1189.

Campbell, J. (1987). Stock returns and the term structure. *Journal of Financial Economics*, **18**, 373-399.

Chiang, W., Urban, T., & Baldridge, G. (1996). A neural network approach to mutual fund net asset value forecasting. *Omega*, **24**, 205-215.

Cochrane, J. (1991). Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance*, **46**, 209-238.

Cole, C. S. (1994). Forecasting interest rates with eurodollar futures rates. *Journal of Futures Markets*, **14**, 37-50.

Consumer enemy no. 1. (1991, October 28). *Newsweek*. pp. 42, 47.

Corman, L. (1988). Data integrity and security of the corporate data base: The dilemma of end user computing. *Data Base*, **19**, 1-5.

Davis, G. (1984). Caution: User developed systems can be dangerous to your organization. *MISRC Working Paper 82-04*, MIS Research Center, University of Minnesota.

Davis, G., Adams, D., & Schaller, C. (1983). *Auditing & EDP*. New York: American Institute of Certified Public Accountants, Inc.

Davis, G., & Olson, M. (1985). *Management information systems: Conceptual foundations, structure, and development*. New York: McGraw-Hill Book Company.

Dead farmer syndrome haunts efforts to trim USDA offices. (1993, April 19). *Minneapolis Star Tribune*, p. 5A.

Fama, E., & French, K. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, **25**, 23-49.

Ferson, W., & Harvey, C. (1993). The risk and predictability of international equity returns. *Review of Financial Studies*, **6**, 527-566.

Fox, C., Levitin, A., & Redman, T. (1993). The notion of data and its quality dimensions. *Information Processing & Management*, **30**, 9-19.

French, K., Schwert, G., & Stambaugh, R. (1987). Expected stock returns and volitatility. *Journal of Financial Economics*, **19**, 3-30.

Glosten, C., Jagannathan, R., & Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance*, **48**, 1779-1802.

Hershey, R. D. (1995, January 16). US is considering a large overhaul of economic data. *New York Times*, pp. A1 and D3.

Huh, Y., Keller, F., Redman, T., & Watkins, A. (1990). Data quality. *Information and Software Technology*, **32**, 559-565.

The individual investor's guide to low-load mutual funds. 16th ed. (1997). Chicago, IL: American Association of Individual Investors.

Jabbour, G. M. (1994). Prediction of future currency exchange rates from current currency futures prices: The case of GM and JY. *Journal of Futures Markets*, **14**, 25-36.

Jankus, J. C. (1997). Relating global bond yields to macroeconomic forecasts. *Journal of Portfolio Management*, **23**, 96-101.

Knight, B. (1992). The data pollution problem. *Computerworld*, **26**, 81-83.

Laudon, K. (1986). Data quality and due process in large interorganizational record systems. *Communications of the ACM*, **29**, 4-11.

Lewis-Beck, M. S. (1980). *Applied regression: An introduction*. Newbury Park, CA: Sage Publications, Inc.

Madnick, S., & Wang, R. (1992). Introduction to the TDQM research program. *Total Data Quality Management Research Program Working Paper #92-01*.

Mark, N. C. (1995). Exchange rates and fundamentals: Evidence on long-horizon predictability. *The American Economic Review*, **85**, 201-218.

Mentzer, J. T., & Cox, J. E. (1984). Familiarity, application, and performance of sales forecasting techniques. *Journal of Forecasting*, **3**, 27-36.

Morey, R. (1982). Estimating and improving the quality of information in a MIS. *Communications of the ACM*, **25**, 337-342.

Morgenstern, O. (1963). *On the accuracy of economic observations*. Princeton, NJ: Princeton University Press.

Nayar, M. (1993). Achieving information integrity: A strategic imperative. *Information Systems Management*, **10**, 51-61.

Neter, J., Wasserman, W., & Kutner, M. (1990). *Applied linear statistical models*. 3rd ed. Homewood, IL: Irwin.

O'Leary, D. (1993). The impact of data accuracy on system learning. *Journal of Management Information Systems*, **9**, 83-98.

Panko, R. R. (1998). What we know about spreadsheet errors. *Journal of End User Computing*, **10**(2), 15-21.

Pesaran, M., & Timmermann, A. (1994). Forecasting stock returns: An examination of stock market trading in the presence of transaction costs. *Journal of Forecasting*, **13**, 335-367.

Pesaran, M., & Timmerman, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, **50**, 1201-1228.

Redman, T. C. (1992). *Data quality: Management and technology*. New York: Bantam Books.

Redman, T. (1995). Improve data quality for competitive advantage. *Sloan Management Review*, **36**, 99-107.

Redman, T. C. (1996). *Data quality for the information age*. Norwood, MA: Artech House, Inc.

Refenes, A. N., Zapranis, A., & Francis, G. (1994). Stock performance modeling using neural networks: A comparative study with regression models. *Neural Networks*, **7**, 375-388.

Rosenberg, B., & Houglet, M. (1974). Error rates in CRSP and COMPUSTAT data bases and their implications. *The Journal of Finance*, **29**, 1303-1310.

Sanders, N. R. (1994). Forecasting practices in US corporations: Survey results. *Interfaces*, **24**, 92-100.

Statistical abstract of the United States. (1996). Washington, D.C.: U.S. Bureau of the Census, Government Printing Office.

Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, **39**, 86-95.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, **12**, 5-34.

Zmud, R. (1978). An empirical investigation of the dimensionality of the concept of information. *Decision Sciences*, **9**, 187-195.