Business Valuation Review
Volume 28 • Number 2

BVR

A Second Course in Regression Analysis as Applied to Valuation and Lost Profits

Mark G. Filler, CPA/ABV, AM

I first address the initial part of an earlier Hawkins article where he demonstrates the use of regression analysis to determine a subject company's value. Herein, it is shown that even a simple, basic application of regression analysis requires more fundamental analysis than Hawkins put forth. My analysis presents both flaws in the data sets and a conceptual flaw in the use of revenue alone to predict value. I propose and demonstrate corrective methodologies for both of these errors. I then deal with the use of regression analysis to determine the amount of active versus passive appreciation in the growth in value of a company during a marriage. I explain why this is a misuse of regression analysis and demonstrate better regression techniques for forecasting revenues using time trend and housing starts as independent variables, and, at the same time, I transform the model to correct for the serial correlation problem raised in Trout's letter to the editor.

Introduction

In a well-presented article in a previous issue of the *Business Valuation Review*,[1] author George B. Hawkins lays out nicely the basic precepts of regression analysis and their application in a straightforward setting and in an easy to understand format. As this current article will refine, extend, and elaborate on Hawkins' article, the reader is requested to read or refer to it, as well as the letter to the editor in a subsequent issue of this journal,[2] before continuing on.

While regression analysis (RA) is the ideal tool to explore relationships between and among variables, it has two other uses and a major benefit. Those uses are:

- It uses one or more input variables to predict the value of an output variable.
- It determines which input variables are the best predictors for a particular output variable.

The major benefit of choosing RA over an average or median valuation multiple is the reduction in dispersion obtained by RA. Dispersion, or variation about the mean or trend line, needs to be minimized if the accuracy of our prediction is to be contained within a more narrow rather than a broader range. This is an important consideration since valuation is always considered to be a range,[3] and, consequently, we would like that range to be as small as possible. Of the three possible ways of developing a valuation multiple (mean, median, and RA), RA will consistently deliver more accurate predictions than its rivals.

To compare the degree of variation that exists among competing procedures, we can use a common-size metric known as the coefficient of variation (CoV), which is a ratio that is computed by dividing the standard deviation by the mean. If we use the data sets in Table 1 of the Hawkins' article, we can select the average price/Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) ratio (9.66) and the average price/revenue ratio (.86). By multiplying each of these ratios against the appropriate EBITDA and revenue amounts, we will derive 11 predicted prices for each category. By averaging the 11 prices so attained and then calculating their standard deviation, we will find the inputs necessary to compute the CoV. For the two regression models presented in the Hawkins article, the CoV is calculated by dividing each standard error[4] of the estimate[5] (SEE)

[1]Hawkins, George B. "Regression Analysis in Valuation Engagements", *Business Valuation Review* 27(1) (Spring 2008).
[2]Robert R. Trout, "Letter to the Editor", *Business Valuation Review* 27(3) (Fall 2008).

Mark G. Filler, CPA/ABV, CBA, AM, CVA leads Filler & Associates' Litigation and Claims Support practice in Portland, Maine.

[3]This is true even if the client or the trier of fact wants or needs a discrete point estimate of value.
[4]The summary output in Hawkins' Figure 3 is incorrect because it reflects the use of only nine of the eleven available observations. The SEE for the eleven observations is 13,563.439.
[5]This is a measure of the accuracy of the prediction obtained from a regression model. It measures the amount of scatter, or variation, in the actual data around the fitted regression line. Think of it as the standard deviation of the trend line.

Table 1
Comparative Dispersion Metrics

	Price/revenue	Price/EBITDA	RA–revenue	RA–EBITDA
Average price	28,728	39,522	31,479	31,479
Standard dev.	21,842	38,668	13,563	4,714
CoV	76.0%	97.8%	43.1%	15.0%

by the average of price paid. The results of the four computations are shown in Table 1 here.

The conclusion to be drawn from this table is obvious—RA reduces variation. Would you rather say that you are 68% confident that your point estimate of value is ABC, plus or minus 76%, or plus or minus 15%? Later in this article, we will explore procedures such as outlier removal and transformations that will reduce the CoV even further.

The determination of value as a function of revenue is a generally accepted valuation method, whether one uses average or median multiples or RA. However, a noted author and teacher[6] has written that the use of revenue alone without controlling for profit margin can lead to misleading valuations. For example, should two companies, each with $1,000,000 in revenue but with operating profits of $150,000 and $75,000, respectively, sell for the same multiple of revenue? Since one wouldn't think so, we need to employ a model that will control, or account for, the differences in profitability while still recognizing the level of sales that has been obtained.[7]

The RA model that will accomplish this uses the ratio of EBITDA/revenue as its input, or independent, or x variable, and price/revenue as its output, or dependent, or y variable. By converting our variables from real numbers to rational numbers (ratios), we have in effect transformed the data, which in turn requires us to compute our goodness-of-fit metrics in natural language or space by ultimately back-transforming our results. Let's walk through this process. The setup data for the ratio model is shown on Table 2. The summary output of the regression is found on Table 3. To predict price paid, we add the intercept to the product of the EBITDA/revenue coefficient and the EBITDA/revenue ratio. This results in a revised revenue multiple, which, when multiplied by revenue, results in the predicted price. For example,

Table 2
Set-up Data for Ratio Model

Transaction no.	x variable EBITDA/revenue	y variable Price/revenue
1	0.220	1.546
2	0.119	0.833
3	0.096	0.787
4	0.097	0.782
5	0.068	0.790
6	0.144	1.054
7	0.107	0.791
8	0.125	0.862
9	0.022	0.627
10	0.210	0.811
11	0.053	0.573

the formula for transaction number five would be as follows: .4859 + 3.2612 * .068 = .7064 * 42,705 = 30,167. Figure 1 is a scattergraph showing actual price paid on the y-axis and predicted price paid on the x-axis. While the results we have achieved so far are an improvement over the use of revenue alone, a glance at Figure 1 indicates a data point in the upper right corner that is obviously an outlier. Rather than removing it from the data set, let's try another procedure to pull it and the other data points in closer to the trend line.

The data sets of sales transactions that are available to us by way of Standard Industrial Classifications (SIC) Code Numbers and North American Industrial Classification System (NAICS) Code Numbers are rarely distributed in such a manner that the application of simple linear regression will give us a relevant and reliable answer. This is because the data sets are (a) hardly ever linear, (b) infrequently homogeneous as to variance (the larger the x variable, the greater, or smaller, the dispersion about the regression line), and (c) not often normal, or even symmetrical. If the data are linear, then we can proceed to use simple linear regression without having to resort to more complex models, i.e., we can stick with the basic tools that Excel provides us. The reason that homogeneity and normality, or at least symmetry, are good things is beyond the scope of this article, but suffice it to say that without these qualities, standard statistical tests and confidence intervals will not be reliable, nor will you be

[6] Aswath Damodaran, *Valuation: Tools and Techniques for Determining the Value of Any Asset*. 2nd ed. (New York: John Wiley & Sons, Inc., 2002), chapter 20.

[7] Using cash flow or earnings as the sole value driver will usually undervalue the low cash flow/high sales business in these circumstances. In these cases, the buyer has only to contain and control costs without the added burden of necessarily increasing sales. Therefore, there is compensable value in having a certain level of sales, even if the commensurate profits do not currently exist.

Table 3
Summary Output for Ratio Model

Regression Statistics	Transformed	Back-transformed				
Multiple R	0.760	0.981				
R^2	0.578	0.962				
Standard error	0.177	7,305				
CoV	20.6%	23.2%				
Observations	11	11				
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	0.387	0.387	12.333	0.007	
Residual	9	0.282	0.031			
Total	10	0.669				
	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	0.4859	0.119	4.080	0.003	0.217	0.755
EBITDA/revenue	3.2612	0.929	3.512	0.007	1.160	5.362
Transaction No.	Intercept	EBITDA/revenue Coefficient	EBITDA/revenue Ratio	Revised Revenue Multiple	($000) Revenue	Predicted Price ($000)
1	0.4859	3.2612	0.220	1.2020	61,283	73,661
2	0.4859	3.2612	0.119	0.8748	62,444	54,629
3	0.4859	3.2612	0.096	0.7978	61,508	49,074
4	0.4859	3.2612	0.097	0.8030	60,119	48,273
5	0.4859	3.2612	0.068	0.7064	42,705	30,167
6	0.4859	3.2612	0.144	0.9547	32,000	30,549
7	0.4859	3.2612	0.107	0.8349	26,674	22,270
8	0.4859	3.2612	0.125	0.8935	8,704	7,777
9	0.4859	3.2612	0.022	0.5591	5,901	3,300
10	0.4859	3.2612	0.210	1.1715	3,144	3,683
11	0.4859	3.2612	0.053	0.6580	3,139	2,065

able to explain away the variation in your data as noise, or ordinary and expected random error: simple tests will make it apparent that your model is deficient.

Fortunately, to fix these three problems we only need one procedure, and that is transformation of either or both the x and y variables. This is true because data that are not normally distributed are also often neither linear nor homogeneous. Thus, transformation provides a simple way both to fix statistical problems (nonsymmetrical and heterogeneous distributions) and to fit curves to data (curvilinear regression).

Transformation of variables is not new to business valuation, as both Jay Abrams and Roger Grabowski, in their work with the Ibbotson and Duff & Phelps databases, respectively, have shown. In both instances, the x variable, market size, was transformed logarithmically to straighten out the curved distribution of data points that is generated when discount rates are plotted against market size. However, a transformation by logarithms, which converts changes in data from absolute changes into percentage changes, does not work as well with the transaction data sets as transformation by exponents, since we can select the exponent that works best in the situation, while the logarithm of any number is fixed so that the data set at hand may not conform to a logarithmic fix. Therefore, because of the flexibility afforded our transformation process by exponents, this option will be the transforming process we demonstrate in this article.

This next procedure entails transforming both the x and y variables by raising each of them to its own individual power such that the model's SEE is minimized. We constrain this optimizing model by making the standardized residuals[8] sum to zero, as they should in any circumstance. The Excel tool we use is Solver, which can be added to the Tools drop-down menu by checking the Solver box in the Tools, Add-ins menu.

Table 4 shows the setup data, Solver results, and forecasted price that is calculated using the same procedures demonstrated on Table 3 and discussed already, except

[8]A residual is the difference between actual price paid and predicted price paid. To standardize a residual, we divide it by the SEE, the standard deviation about the trend line.

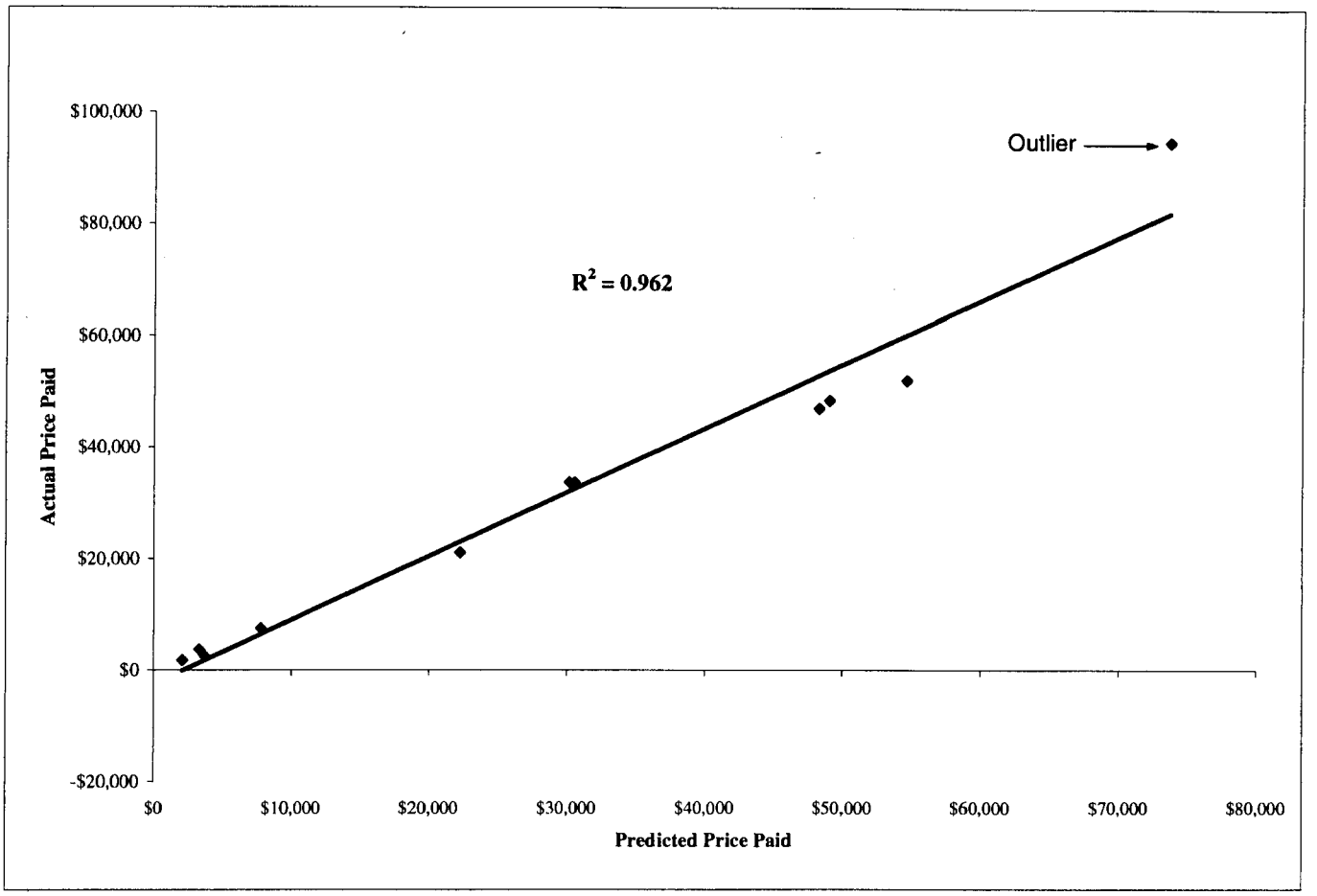

Figure 1
Ratio Model

Table 4
Set-up Data for Transformation Model

	Exponents				
	92.99555831	4.79386269	($000)	($000)	
Transaction No.	EBITDA/revenue	Price/revenue	Revenue	Forecasted Price	Standardized Residual
1	5.9223E–62	8.084	61,283	94,762	(0.00)
2	1.3279E–86	0.416	62,444	51,234	(0.27)
3	1.6394E–95	0.317	61,508	50,466	0.74
4	7.4355E–95	0.307	60,119	49,327	0.83
5	1.6129E–109	0.323	42,075	35,039	0.46
6	4.5843E–79	1.284	32,000	26,255	(2.67)
7	5.5763E–91	0.325	26,674	21,886	0.28
8	1.0291E–84	0.490	8,704	7,142	(0.13)
9	5.0865E–154	0.107	5,901	4,842	0.41
10	1.0372E–63	0.366	3,144	2,745	0.07
11	1.6030E–119	0.070	3,139	2,575	0.28
				Residual sum =	(0.0)
	SEE =	2,795		CoV =	8.9%

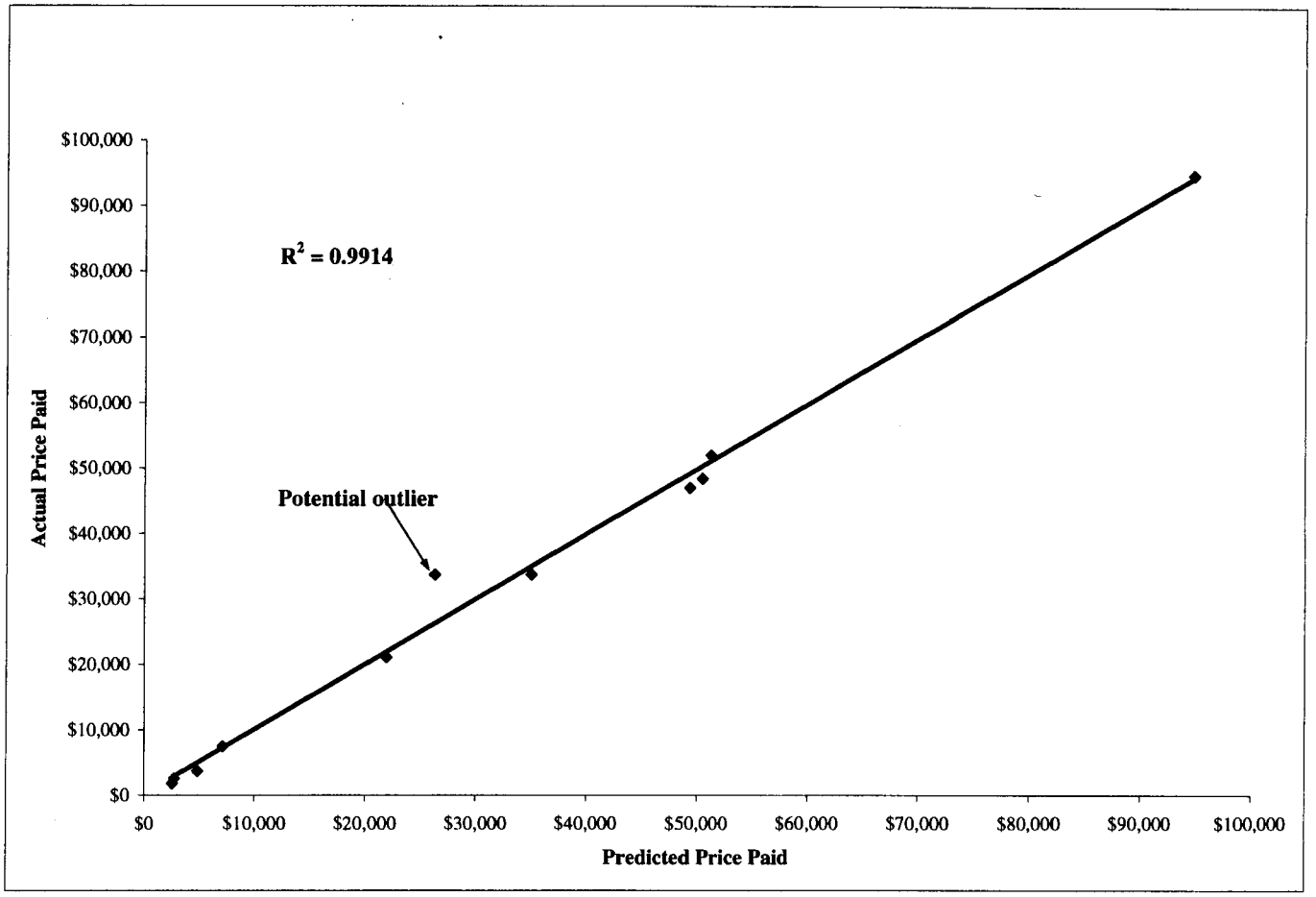

Figure 2
Ratio Model, Transformed

that we must employ a back-transformation process[9] for this model.

Once again, we create a scattergraph with actual price paid on the *y*-axis and predicted price paid on the *x*-axis and show the results in Figure 2. While these results are superior to those shown in Figure 1, we still have one potential outlier. Transaction six, while not more than three standard deviations from the trend line, is still greater than 2.6 standard deviations, an outcome we would expect only 1% of the time.[10]

Therefore, without showing the setup table, we removed transaction six and reran Solver, obtained new optimizing exponents while again constraining the standardized residuals to sum to zero. Forecasted price was calculated the same way as shown on Table 3, with the addition of the back-transforming procedure. This resulted in an SEE of 1,030 and a CoV of 3.3%. The scattergraph of actual versus predicted price paid is shown in Figure 3, with an attendant R^2 of .999. Although the increase in R^2 over Figure 1 is de minimis, the decrease in SEE is on the order of 63%, i.e., from 8.9% to 3.3%, thereby justifying the removal of transaction six. While results this excellent will not often be achieved in practice, this example indicates the type of superior outcomes that the transformed ratio model can accomplish.

Since the application of RA is almost always an iterative process—the first model chosen is rarely the best model—let's apply this process to Hawkins' revenue and price paid data set, for pedagogical purposes only, because in practice, a revenue-only model would be

[9]Since we have transformed the *y* variable, in order to have our predicted price paid result be in natural language, we must back-transform the initial result by raising it to the power of (1/4.79386269). This is the same procedure involved in squaring the integer 2 to get 4, and then raising 4 to the power of (1/2), or .5, to back-transform it to 2.

[10]In a normal (bell-shaped) distribution, the total area under the curve equals 100% probability. One standard deviation from the mean will encompass 68.26% of the area under curve, leaving 31.74% probability in both tails of the curve. Two standard deviations encompass 95.44%, leaving 4.56% probability in both tails. Three standard deviations encompass 99.74%, leaving .26% probability in both tails, and 2.6 standard deviations encompass 99.06%, leaving about 1.0% probability in both tails.

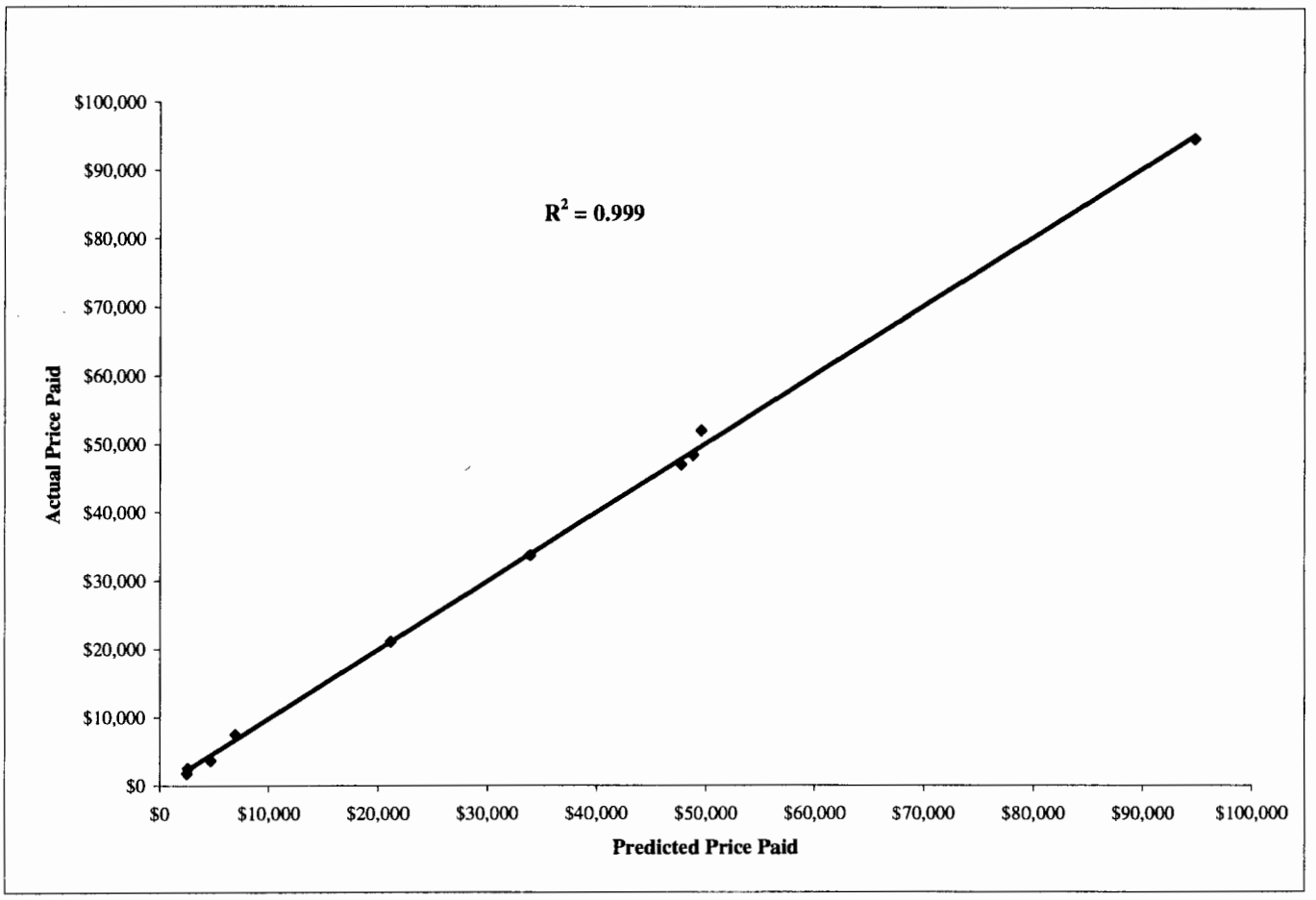

Figure 3
Ratio Model, Transformed and Outlier Removed

ineffective in determining value. We will make attempts at devising a model that meets the following criteria: (a) highest R^2, (b) lowest SEE, (c) residuals that add to zero, and (d) no standardized residual greater than 2.6. The first attempt is the regression of price paid against revenue. The second attempt is a similar regression but with transaction number one removed because it is an outlier—it's 2.75 standard deviations from the trend line. However, this regression, while an improvement over the first, produced another outlier—original transaction number six has a residual that is 2.69 standard deviations from the trend line. Rather than remove it, we have transformed both the x and y variables in the manner previously discussed and demonstrated and run a third regression. This transformation process allows us to keep original transaction number six (it is now only 2.17 standard deviations from the trend line) and have a higher R^2 and a lower SEE than the second model. The results of the three models are shown in Table 5.

Table 5
Goodness of Fit Model

	First Attempt	Second Attempt	Third Attempt
SEE	13,563	2,946	2,588
CoV	43.1%	11.7%	10.3%
R^2	.797	.981	.986

Next, let's turn our attention to Hawkins' EBITDA and price paid data set and explore the issues of outliers and influential and leverage points in RA. First, let us define our terms. An outlier is any really unusual observation, while only influential and leverage points have large residuals. Typically, not all outliers are both influential and exert leverage. An observation is influential if removing it would markedly change the trend line. Leverage points are those observations that have values far from the mean of the x variable. Observations that are far from the mean of the y variable are just plain outliers, since they are usually not influential in relation to the regression results. Only observations that have high leverage and are outliers tend to be influential. See Figure 4 for a demonstration of these concepts.

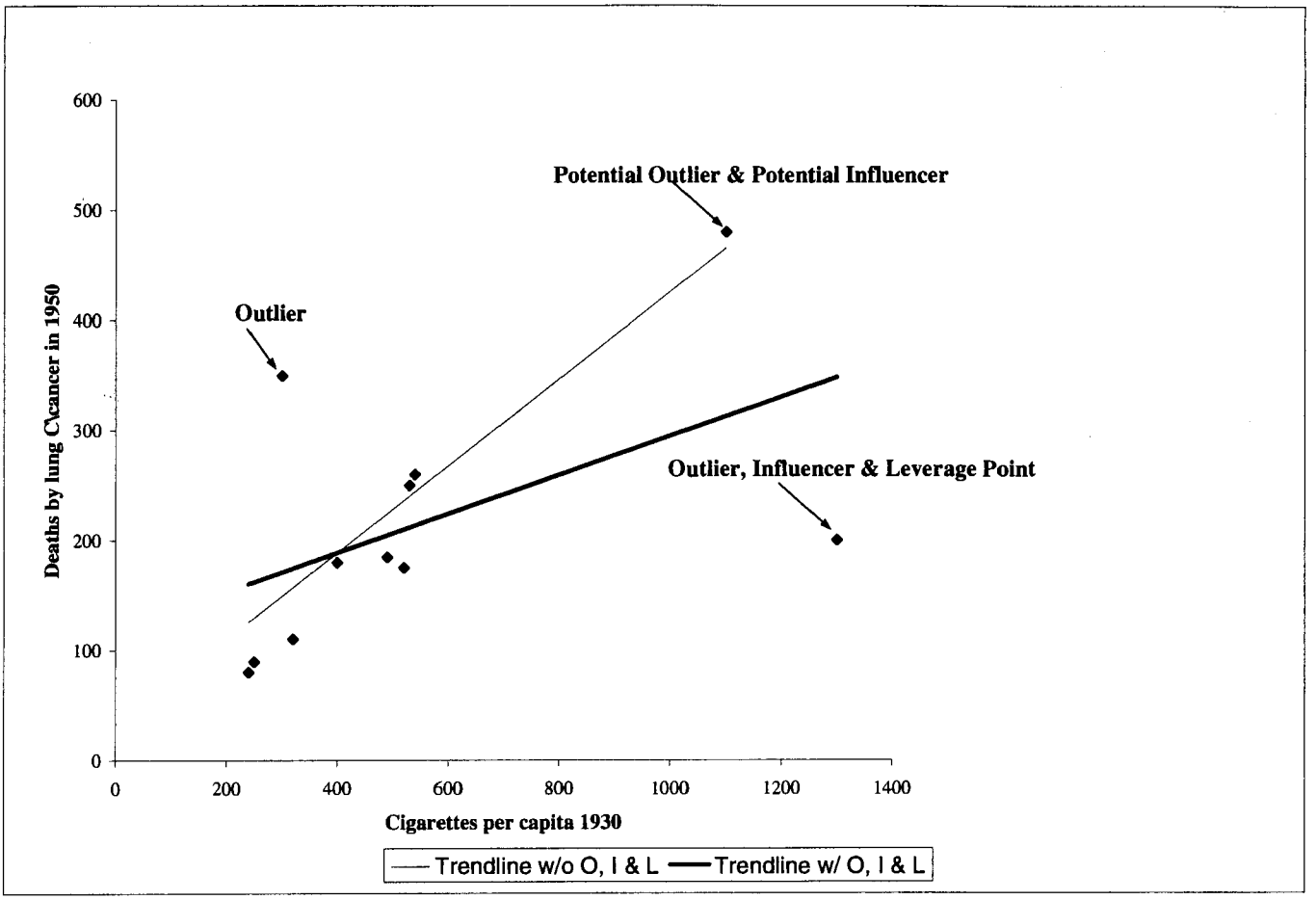

Figure 4
Outliers, Influencers, and Leverage Points

As can be seen on Figure 4, the data point in the southeast corner of the graph has a tremendous influence on both the slope and the intercept of the trend line, and it should either be removed and/or investigated further. The data point in the northwest corner, while an outlier, would have little, if any, effect on the trend line if it were removed. The data point in the northeast corner is an observation of interest because it has the potential to influence the regression outcome. This same situation exists in Hawkins' Figure 2, where the distance from the average of x of the data point in the northeast corner of the chart, which represents transaction number one, causes us to question its effect on the trend line. Therefore, we run a second regression without that observation in the data set, and then test whether the results are statistically and practically significant from the original regression. Table 6 compares the two regression outputs for those items of importance regarding influence and leverage.

While this comparative table seems to indicate no substantial difference between the two models, let us exam the results of two statistical tests to be absolutely sure.

Table 6
Regression Output

	Full Model	Model Without Trans. No. 1
Intercept	2,686.27	1,632.17
EBITDA coefficient	7.03	7.45
SEE	4,714.48	4,765.39
CoV	15.0%	18.4%
R^2	.975	.951
Largest standardized residual	2.4	2.4

The first significance test was a regression model that included a dummy variable[11] of 1 for transaction one and a 0 for the other ten transactions. If transaction number one is influential, then the t-statistic for the dummy variable coefficient will be more than 2.262, and its p-value will be less than 5%. We can see on Table 7 that it

[11]Dummy, or binary, variables are defined as either a 1 or a 0 and are used to measure variables that cannot be stated in quantitative terms of either how many or how much, but instead measure qualitative attributes such either/or.

Table 7
To Include or Not Include Transaction No. 1. Tests of Significance

SUMMARY INPUT			
Transaction No.	EBITDA ($000)	Dummy	($000) Price
1	13,457.00	1	94,769
2	7,448.00	0	52,000
3	5,884.20	0	48,400
4	5,845.60	0	47,000
5	2,888.10	0	33,740
6	4,600.00	0	33,715
7	2,854.80	0	21,100
8	1,087.90	0	7,500
9	132.60	0	3,700
10	661.00	0	2,550
11	165.70	0	1,800

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.989
R^2	0.978
Adjusted R^2	0.972
Standard error	4,765
Observations	11

ANOVA

	df	SS	MS	F	Significance F
Regression	2	7,954,516,286	3,977,258,143	175.14	2.4858E–07
Residual	8	181,671,756	22,708,970		
Total	10	8,136,188,043			

	Coefficients	Standard Error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	1,632.17	2,410.54	0.677	0.517	(3,926.54)	7,190.89
EBITDA	7.45	0.60	12.500	0.000	6.08	8.82
Dummy	(7,118.85)	7,916.17	**(0.899)**	**0.395**	(25,373.60)	11,135.89

t-Test: Paired Two Sample for Means

	Forecast w/ 1	Forecast w/o 1
Mean	24,892.53	25,150.50
Variance	351,501,282.95	394,265,896.24
Observations	10	10
Pearson correlation	1	
Hypothesized mean difference	0	
df	9	
t-stat	**(0.736)**	
$P(T \leq t)$ two-tail	**0.480**	
t critical two-tail	2.262	

fails those tests, meaning that the dummy variable, and by inference, transaction number one, is statistically insignificant to the regression model.

The second test of significance is to match up the regression predictions for the other ten transactions as predicted by the two models and test whether their combined differences are statistically significant. We use Excel's t-test for paired samples,[12] the output of which is shown on the bottom half of Table 7. If there were a statistically significant difference between the two different sets of predicted selling prices, then the t-statistic would have been greater than 2.262, and the p-value would have been less than 5%. The null hypothesis of no difference between the predicted values cannot be rejected, meaning we accept it and conclude that transaction number one can stay in the regression model—we gain nothing by removing it. Of course, there was also no practical significance to the change in both the slope and the intercept of the trend line brought about by dropping transaction

[12]The data are paired because each of the two predicted prices for each transaction is based on the same amount of EBITDA.

number one from the regression model. Therefore, if an outlier lies on the plane of the trend line, no matter its distance from the average of x, it should stay in the model.

While there is more we can say about these regression models, especially about residual diagnostics and remediation techniques, let us turn our attention to the second half of Hawkins' article, the section dealing with accounting for active versus passive appreciation of a nonmarital, or "separate" asset. However, before we deal specifically with that topic and the application of RA to it, let's use the data in Hawkins' Table 6 as a basis for forecasting revenues using RA for a valuation or lost-profits assignment.[13] Assuming that we have access to data from the National Association of Home Builders and various state and federal government agencies such that we have accurate forecasts for national housing starts for the years after 2006, and if we have an accurate RA model, then we can forecast post-2006 revenues for our subject company. So, let's build that accurate model, recognizing that it will be a combined model, i.e., it will contain elements of both time-series and explanatory forecasting models.[14]

The first thing one should do when given a data set of observations is to graph the data, the y variable against each x variable. See Figure 5 for those graphs based on Hawkins' Table 6. Figure 5A indicates that the relationship between revenue and time, in this case years, is not linear, but curvilinear, and while Figure 5B shows a linear trend, there is substantial noise, or random error, in the data. While annual housing starts for the first six years range from 894,9020 to 1,146,300, Woodco's revenues are stuck in a rut—they range only from $12,100 to $16,900. This pattern of nonlinearity, or noise, is repeated for subsequent years through 2003, but to a much lesser degree. One minus R^2 value for each model indicates the unexplained variance, or unaccounted noise, in each model, corroborating what our eyes tell us—our RA forecasting model will have to account for curvilinearity and noisy data.

For expediency's sake, let us concede that the models listed in Table 8, while presenting reasonably good metrics, are deficient for one or more reasons, including serial correlation[15] and nonlinearity in the residuals.[16]

That leaves just two models that we need to compare and contrast in detail—a transformation of model three and a second-degree polynomial, or quadratic model, using only time and time-squared as the x variables. The setup data and summary output information for the quadratic model are shown on Table 9.

While the output metrics on Table 9 all look quite good—high R^2, low CoV, and statistically significant p-values for all the coefficients—there is a fly in the ointment. Examining the two charts in Figure 6, we note that there is no serial correlation per Figure 6B, but there is a problem showing in Figure 6A. One of the residuals does not fit inside the imaginary box—we do not have random dispersion without any outliers or patterns. This tells us that our model is missing something, and we need try again, this time with a transformation process utilizing both years and national housing starts in a combined time-series and explanatory model. The setup data and summary output information for this model are shown on Table 10.

Not only are the output metrics better than those of the quadratic model in all respects, but also the residuals in Figure 7A illustrate almost perfect random dispersion, and the residuals in Figure 7B demonstrate a lack of serial correlation. Table 11 tabulates the results of the

[13]Sample size is always an issue in RA, and the way it is dealt with depends on whether one is (a) testing a hypothesis (e.g., eating two pounds of rutabaga a day for six months will increase one's score on a standardized test) or (b) predicting or forecasting (e.g., what might a business' value be based on a given amount of cash flow).

For hypothesis testing, the researcher wants a sample size large enough so that a type II error is not committed (failing to reject the null hypothesis when logic indicates it should be) but not so large that a type I error is committed (rejecting the null hypothesis when there is no reason to do so). In terms of the courtroom, the researcher neither wants to convict the innocent nor free the guilty. Since the number of observations, or cases, in the researcher's experiment, needed to avoid these errors is unknown, there are as many rules of thumb that suggest sample size as there are statistical textbook authors.

For the valuation analyst, choice of sample size is rarely an option. The data available are a given (e.g., ten years of sales history, 25 observations in a transaction database, etc.), and more data are rarely available in order to get the RA model to perform better. However, we do have a simple test that enables us to determine if we have a sufficiently large sample. The p-value of the x-coefficient's t-statistic is a function of, among other things, sample size. So, if the p-value is .05 or less, the x-coefficient will be statistically significant, i.e., different from zero, whether the sample size is 11, 18, or 100. Therefore, if the sample size is too small, the analyst will know it immediately—the p-value will be greater than .05. If it is less than or equal to .05, the model is good to go, regardless of how few observations there are in the sample.

In either case, a model's ability to account for the data should ultimately be evaluated by examining the deviations between predicted and observed values. Sample size should not be the factor determining the success or failure of the model. Sample size reflects the resources available to the researcher and the valuation analyst; it does not reflect the verisimilitude of the model. William Gossett, the inventor of the t-distribution, worked at Guinness Brewery as head of quality control and used sample sizes of three to five observations.

[14]The type of data used in the valuation models discussed above was cross-sectional, which refers to data collected by observing many subjects, such as company data, at the same point of time, or without regard to differences in time. For the succeeding section, we will focus on time series, which is a sequence of observations that are ordered in time. If observations are made on some phenomenon throughout time, it is most sensible to display the data in the order in which they arose, particularly since successive observations will probably be dependent.

[15]Serial, or auto, correlation exists when the residuals are correlated over time. It is often found in time-series data when one period's value is related to a preceding period's value. It is usually not a problem with cross-sectional data, unless those data are ranked by size.

[16]When the residuals are plotted against the fitted values, they should form a square or rectangle around their mean value of zero. The distribution of the residuals in this square or rectangle should be random, without any patterns or outliers presenting.

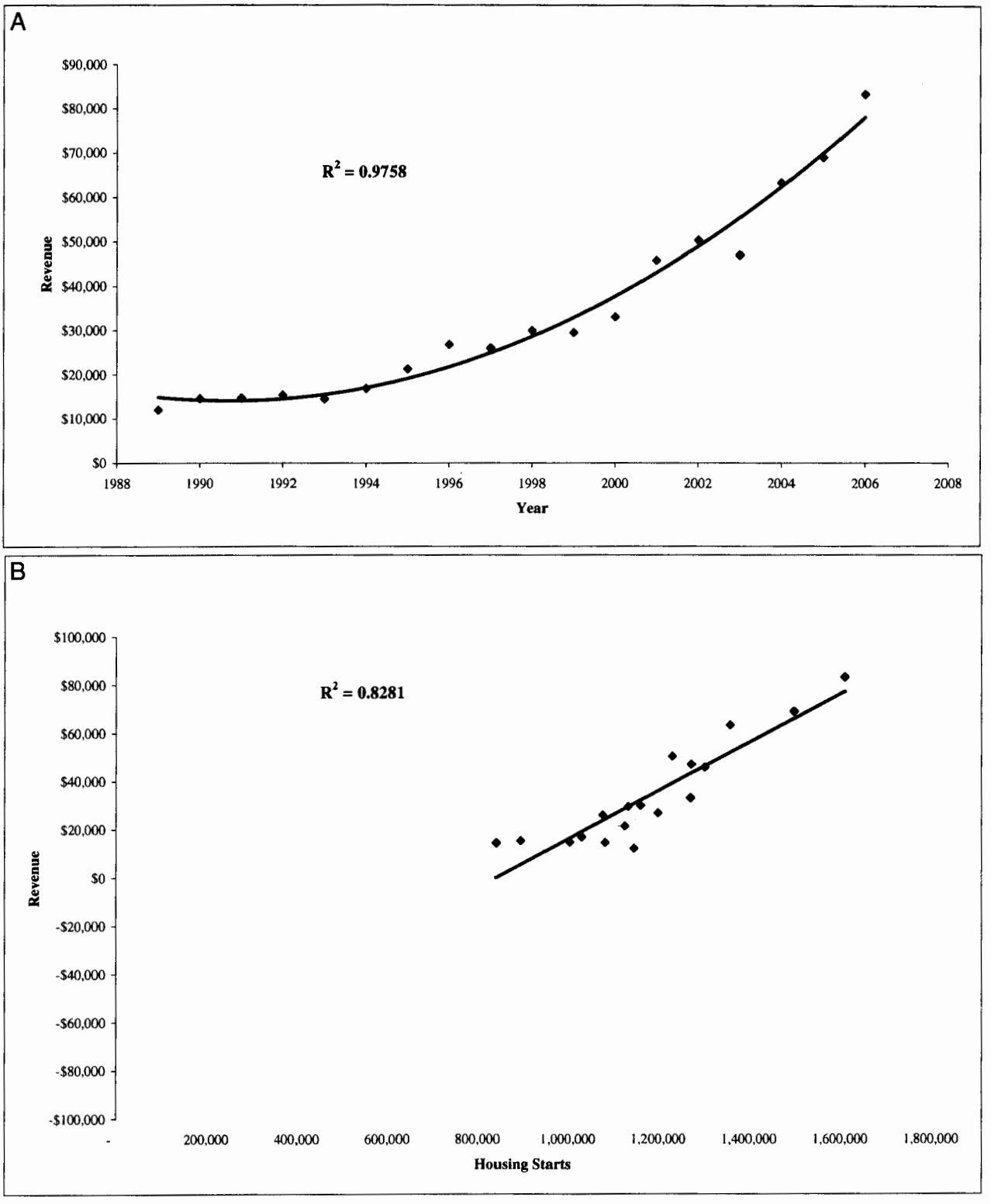

Figure 5
(A) Scattergraph – Revenue vs. Time
(B) Scattergraph – Revenue vs. Housing Starts

Table 8
Comparative Data and Statistics

	Model 1	Model 2	Model 3
X_1 variable	Housing starts	Time	Housing starts
X_2 variable	–	–	Time
SEE	9,083	7,748	5,850
CoV	26.6%	22.7%	17.1%
R^2	.8281	.8749	.9332
$1 - R^2$	.1719	.1251	.0668

contest between the two models. The three line-fit plots in Figure 8 graphically show how well the combined model fits the data, as well as giving a preliminary insight into which of the two x variables is more important than the other.

For a multiple regression model like the combined model we have just demonstrated, the analyst often wants to know which of the variables is the most important, both relatively and absolutely, i.e., he or she wants to know its effect size. A beta coefficient, or weight, of an x variable measures the relative importance of that variable against the others in a multiple regression model. It is the average amount the y variable increases when the x variable increases one standard deviation and the other x variables are held constant. The beta weight formula for any particular x variable is: x-coefficient * standard deviation of the x variable/standard deviation of the y variable.

Table 9
Woodco Predicted Revenues
Time-Series Analysis—Quadratic Model

SUMMARY INPUT

Year	Year^2	Revenues ($000)
1989	3,956,121	12,100
1990	3,960,100	14,600
1991	3,964,081	14,700
1992	3,968,064	15,400
1993	3,972,049	14,500
1994	3,976,036	16,900
1995	3,980,025	21,300
1996	3,984,016	26,800
1997	3,988,009	26,000
1998	3,992,004	30,000
1999	3,996,001	29,500
2000	4,000,000	33,100
2001	4,004,001	45,900
2002	4,008,004	50,400
2003	4,012,009	47,100
2004	4,016,016	63,400
2005	4,020,025	69,100
2006	4,024,036	83,400

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.9878
R^2	0.9758
Adjusted R^2	0.9726
Standard error	3,517
CoV	10.3%
Observations	18

ANOVA

	df	SS	MS	F	Significance F
Regression	2	7,494,332,549	3,747,166,275	303	7.47803E–13
Residual	15	185,578,562	12,371,904		
Total	17	7,679,911,111			

	Coefficients	Standard Error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	1,085,180,317.29	138,042,739.49	7.861	0.000	790,949,002	1,379,411,633
Year	(1,090,235.75)	138,216.16	(7.888)	0.000	(1,384,837)	(795,635)
Year^2	273.83	34.60	7.915	0.000	200	348

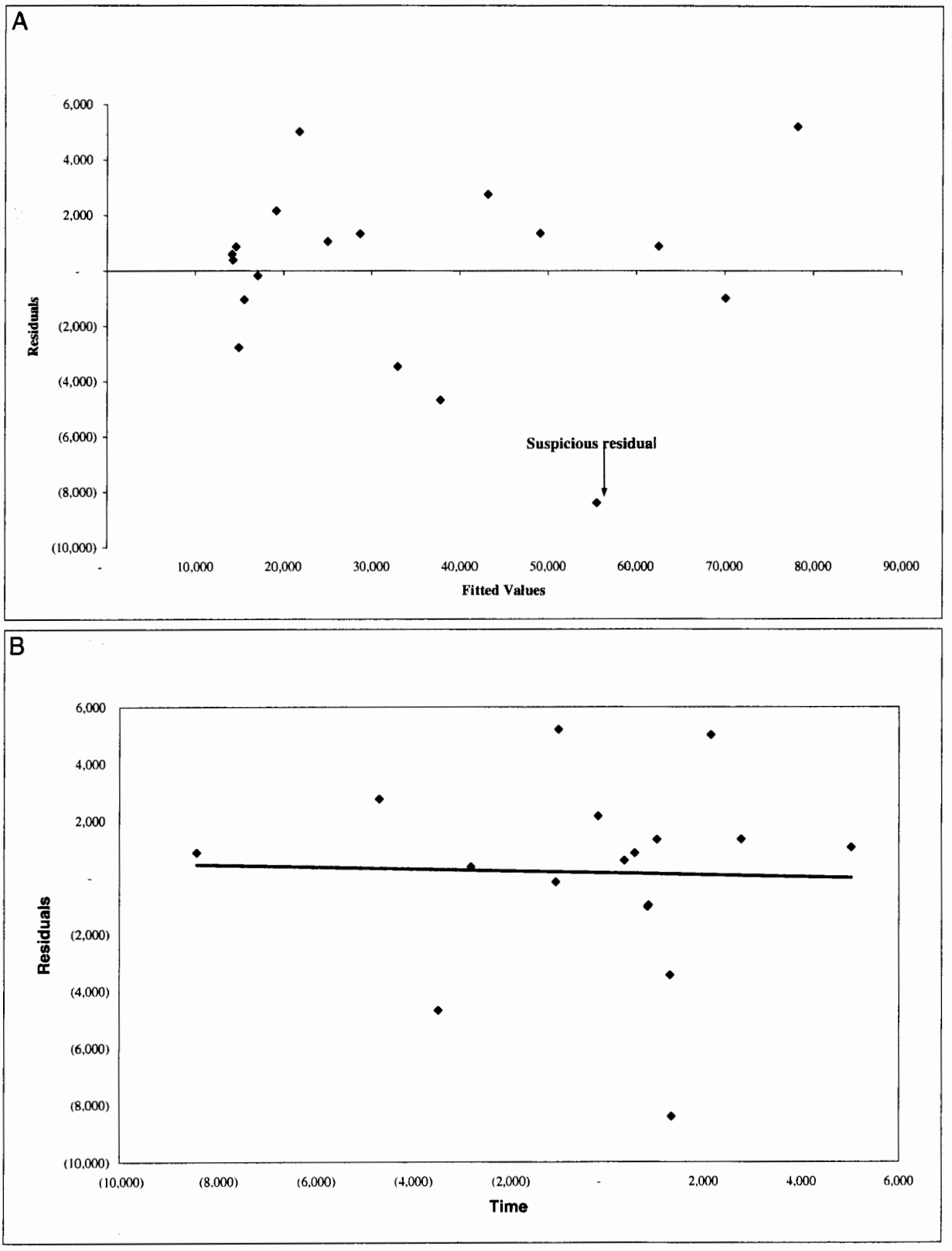

Figure 6
(A) Residuals versus Fitted Values
(B) Residuals versus Time

Table 10
Woodco Predicted Revenues
Combined Time-Series Analysis & Explanatory Model

SUMMARY INPUT

				Exponents		
				1.03249	–1.52424	–0.20912
	Natural			Transformed		
Year	*x* Time	*x* Starts	*y* Revenue ($000)	*x* Time	*x* Starts	*y* Revenue
1989	1	1,146,300	12,100	1.000	5.810E–10	0.140
1990	2	1,081,400	14,600	2.046	6.350E–10	0.135
1991	3	1,003,400	14,700	3.109	7.117E–10	0.134
1992	4	894,900	15,400	4.184	8.474E–10	0.133
1993	5	840,400	14,500	5.268	9.325E–10	0.135
1994	6	1,030,100	16,900	6.360	6.838E–10	0.131
1995	7	1,125,600	21,300	7.457	5.974E–10	0.124
1996	8	1,198,400	26,800	8.559	5.429E–10	0.119
1997	9	1,076,300	26,000	9.666	6.396E–10	0.119
1998	10	1,161,000	30,000	10.777	5.698E–10	0.116
1999	11	1,133,600	29,500	11.891	5.909E–10	0.116
2000	12	1,271,400	33,100	13.009	4.961E–10	0.113
2001	13	1,302,500	45,900	14.130	4.782E–10	0.106
2002	14	1,230,900	50,400	15.253	5.212E–10	0.104
2003	15	1,273,200	47,100	16.380	4.951E–10	0.105
2004	16	1,358,500	63,400	17.508	4.485E–10	0.099
2005	17	1,499,000	69,100	18.639	3.860E–10	0.097
2006	18	1,610,500	83,400	19.772	3.460E–10	0.094

SUMMARY OUTPUT

Regression Statistics

	Transformed	Back-Transformed
Multiple R	0.99276	0.9927
R^2	0.98557	0.9855
Adjusted R^2	0.98364	0.9835
Standard error	0.00188	2,731
CoV	1.59%	8.00%
Observations	18	18

ANOVA

	df	SS	MS	*F*	Significance *F*
Regression	2	0.003612588	0.001806294	512.1541988	1.56712E–14
Residual	15	5.29028E–05	3.52686E–06		
Total	17	0.003665491			

	Coefficients	Standard Error	*t*-stat	*P*-value	Lower 95%	Upper 95%
Intercept	0.131	0.004	33.638	0.000	0.123	0.140
Time	(0.002)	0.000	(17.873)	0.000	(0.002)	(0.002)
Starts	14,882,064.161	4,861,054.118	3.061	0.008	4,520,966.201	25,243,162.120

RESIDUAL OUTPUT

Observation	Transformed Predicted Revenue	Residuals	Back-transformed Predicted Revenue ($000)
1	0.1379	0.0021	13,023
2	0.1364	–0.0018	13,705
3	0.1353	–0.0008	14,278
4	0.1349	–0.0018	14,437
5	0.1339	0.0010	15,005
6	0.1278	0.0028	18,732

Table 10
Continued

RESIDUAL OUTPUT

Observation	Transformed Predicted Revenue	Residuals	Back-transformed Predicted Revenue ($000)
7	0.1241	0.0003	21,529
8	0.1209	–0.0024	24,394
9	0.1200	–0.0006	25,344
10	0.1165	–0.0007	29,135
11	0.1144	0.0018	31,786
12	0.1106	0.0029	37,415
13	0.1079	–0.0019	42,105
14	0.1061	–0.0022	45,625
15	0.1033	0.0021	51,928
16	0.1001	–0.0011	60,195
17	0.0967	0.0005	70,948
18	0.0937	–0.0002	82,706

Another effect size measure is the elasticity of the x variables. Conventionally, the elasticity of an x variable in a model is defined as the percent change in the y variable for a one percent change in the x variable. The elasticity formula for any particular x variable is: x-coefficient * average of the x variable/average of the y variable. A third effect size measure is level importance, which contrasts the "potential" influence of the raw x-coefficient with its "actual" influence. The formula for level importance is: x-coefficient * x variable mean. It is understood that the larger the x-coefficient is, the more the y variable will change for each unit increase in the x-coefficient, but the smaller the x variable average is, the fewer actual unit changes will be expected. By taking both the size of the x-coefficient and the size of the average value into account, level importance is a better indicator of expected actual influence of the x variable on the y variable. A review of Table 12 reveals that time has larger values than housing starts for all three effect size measures, indicating that it is the more important of the two x variables by a factor of at least two to one. However, this does not mean we should remove housing starts from the model. On the contrary, we have seen that time alone, while a necessary ingredient in the model, is not sufficient by itself to pass all the required tests. It is the combination of the two variables, their joint effect, joined with the transformation process, which makes the model so powerful.

Table 11
Comparative Data and Statistics

	Quadratic Model	Time series & Explanatory Model
X_1 variable	Time	Time
X_2 variable	Time^2	Housing starts
SEE	3,517	2,731
CoV	10.3%	8.0%
R^2	.9758	.9855

Having used Hawkins' Woodco data to explore various forecasting models, let us now turn our attention to the purpose for which Hawkins introduced the data set—the segregation of growth in value into passive and active categories. The fundamental inference that Hawkins draws from regression analysis and brings to bear on this issue is that the R^2 statistic "explains a significant percentage of the variation in Woodco's results," or as he puts it another way, "82.8% of variations in Woodco revenues [can be] explained by changes in national homebuilding activity." Let's contrast these two statements with a quote from a statistics 101 text book[17]: "The more closely x and y are linearly related, the more the variability in the y-values can be explained by variability in the x-values . . . ." Thus, it is variability that is explained, or accounted for, not some basic relationship. Perhaps this will be made clearer if we demonstrate just what the textbook quote really means in a statistical, rather than a literal sense. The more that y is a strict function of x, the more their individual variances will match up, or covary, and the more strongly they will be correlated. For example, if we look at Table 13, we see the price and EBITDA data sets that were first introduced in Table 7. Without creating a scattergraph, we can't tell from just looking at the raw data whether or not there is a relationship between the x and y variables. However, if we common-size the data by standardizing[18] it, we make it easier to see if the two

[17]Ott, L. and Mendenhall, W. *Understanding Statistics*. 5th ed. (Boston: PWS-Kent Publishing, 1990), 421.

[18]Standardizing a value, or creating a z score, is accomplished by subtracting the value from the average of the data set and dividing the result by the standard deviation of the data set.

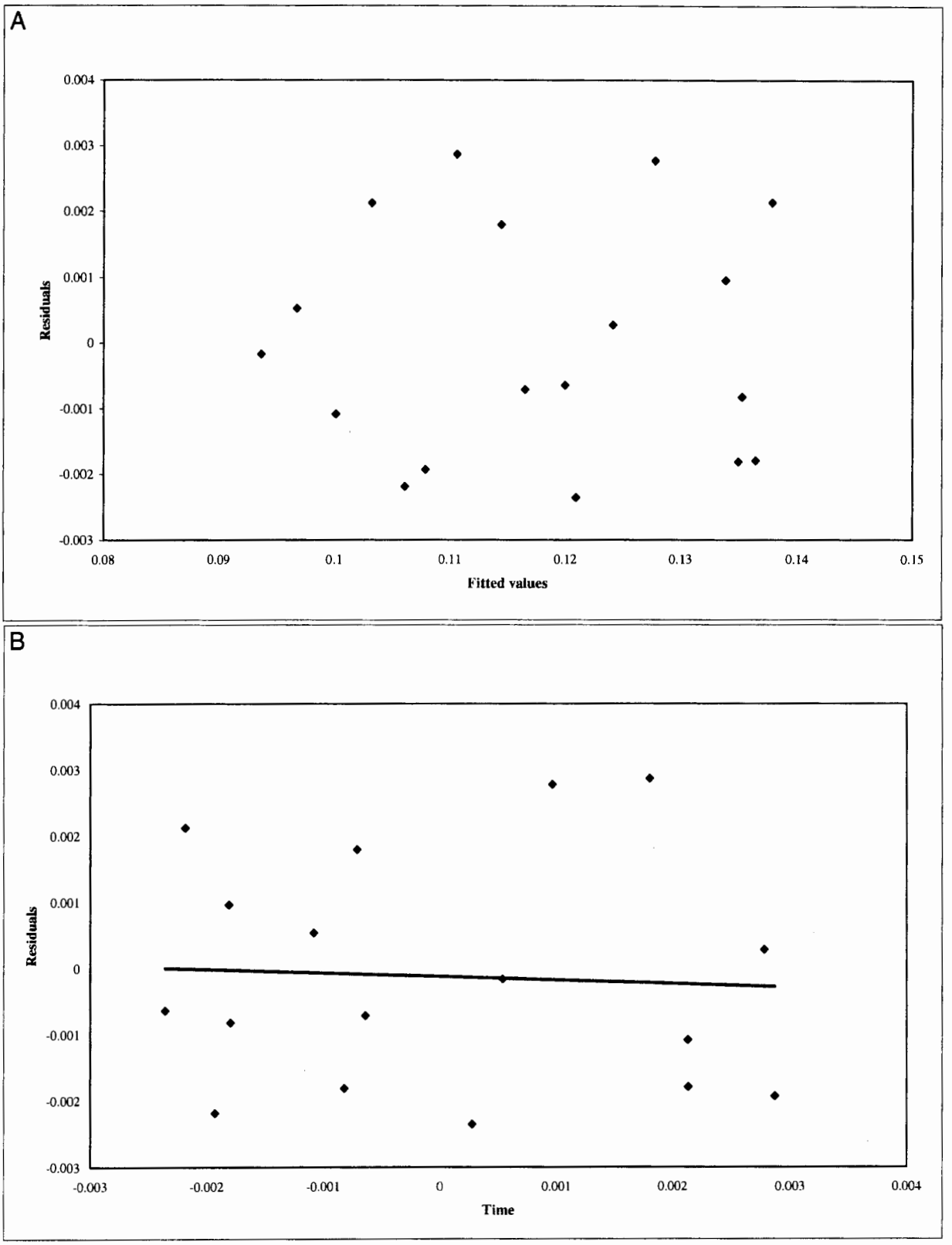

Figure 7
(A) Residuals versus Fitted Values
(B) Residuals versus Time

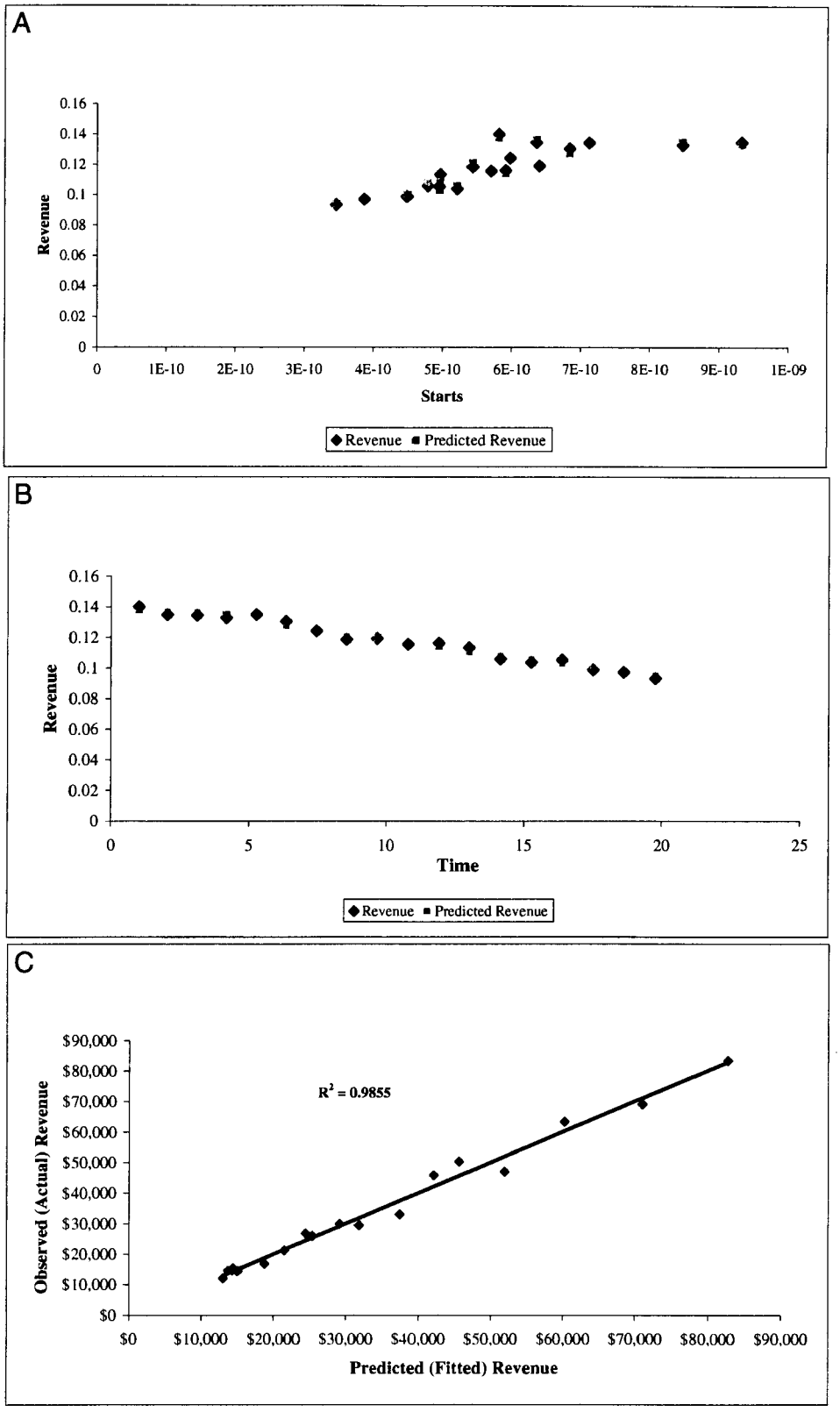

Figure 8
(A) Combined Model, Starts Line Fit Plot
(B) Combined Model, Time Line Fit Plot
(C) Combined Model, Observed vs. Predicted

Table 12
Combined Time-Series Analysis & Explanatory Model
Beta Coefficients, Elasticities, and Level Importance

Year	Revenue (Y)	Starts	Time
1989	0.1400	5.8100E–10	1.0000
1990	0.1346	6.3497E–10	2.0456
1991	0.1344	7.1173E–10	3.1090
1992	0.1331	8.4736E–10	4.1843
1993	0.1348	9.3253E–10	5.2684
1994	0.1306	6.8380E–10	6.3597
1995	0.1244	5.9736E–10	7.4568
1996	0.1186	5.4294E–10	8.5592
1997	0.1193	6.3957E–10	9.6660
1998	0.1158	5.6982E–10	10.7768
1999	0.1162	5.9095E–10	11.8913
2000	0.1134	4.9614E–10	13.0090
2001	0.1060	4.7820E–10	14.1298
2002	0.1039	5.2124E–10	15.2534
2003	0.1054	4.9508E–10	16.3796
2004	0.0990	4.4848E–10	17.5082
2005	0.0973	3.8601E–10	18.6392
2006	0.0935	3.4602E–10	19.7723
Mean	0.118	5.83511E–10	10.278
Standard deviation	0.015	0.000000000	5.909
Beta coefficient		0.149	–0.873*
Elasticities		0.074	–0.189†
Level importance		0.01	(0.02)
	Constant	Starts	Time
Coefficients	0.131408657	14,882,064	–0.0022

* One standard deviation change in time will result in a –.873 standard deviation change in revenue (Y).
† One percent change in time will result in a –.189% change in revenue (Y).

data sets compare, or covary. A cursory examination of the two columns of standardized values indicates that they closely match up, a conclusion verified by the chart in Figure 9. If the two data sets perfectly covaried, each of the two columns in each pair would be of equal length, and R and R^2 would each be 1. However, they are not 1, but .9876 and .9754, respectively, which can be computed from the standardized values. By multiplying each pair of standardized values, we find the amounts in the product column on Table 13. Summing this column and dividing by 10 $(n-1)$ gives us the coefficient of correlation, R, which, when raised to the power of 2, results in an R^2 value of .9754.

A second way of computing R^2 is to first calculate total sum of squares, or total variance, by subtracting each individual value of the y variable from the average of y and squaring the results, and then summing the eleven squared values, which gives 8,136,188,043. The second step is to compute the regression sum of squares by subtracting each individual predicted value from the average of y and squaring the results, and then summing the eleven squared values, which gives 7,936,151,449. If these two numbers matched perfectly, then R^2 would be 1, i.e., the variance, or degree of dispersion about the mean, of the actual values will match the variance of the predicted values. However, they do not, so the ratio of TSS/RSS gives us .9754. What this means is that 97.54% of the total variance is accounted for, or matched, by the variance of the predicted values. Notice in Table 13 how the covariance of price and predicted price are exactly the same as that for price and EBITDA. This is another way of showing that R^2 indicates the fraction of the variation in y (price in this case) that is accounted for by the model.

Some of the confusion surrounding the term "explain" arises from two sources. The first is a literal interpretation of the term. As we have seen, the regression sum of squares RSS is often called the "explained variation" in y, and the residual, or error, sum of squares ESS is called the "unexplained variation." The coefficient R^2 then is interpreted in terms of the proportion of the total variation in y (TSS) that has been "explained" by x. Unfortunately this terminology frequently is taken literally and, hence, misunderstood. Remember that in a regression model,

Table 13
Analysis of Variance and Derivation of R and R^2

	Actual ($000)		Standardized			($000)		Variance		Price ($000)		Standardized	
Trans no.	EBITDA	Price	EBITDA	Price	Product	Average Price	Predicted Price	Total Sum of Squares	Regression Sum of Squares	Predicted	Actual	Predicted	Actual
1	13,457	94,769	2.34	2.22	5.19	31,479	97,349	4,005,566,564	4,338,763,548	97,349	94,769	2.34	2.22
2	7,448	52,000	0.84	0.72	0.60	31,479	55,079	421,092,786	556,928,605	55,079	52,000	0.84	0.72
3	5,884	48,400	0.45	0.59	0.27	31,479	44,078	286,304,858	158,731,673	44,078	48,400	0.45	0.59
4	5,846	47,000	0.44	0.54	0.24	31,479	43,807	240,887,331	151,963,470	43,807	47,000	0.44	0.54
5	2,888	33,740	(0.30)	0.08	(0.02)	31,479	23,002	5,110,066	71,859,751	23,002	33,740	(0.30)	0.08
6	4,600	33,715	0.13	0.08	0.01	31,479	35,045	4,997,663	12,710,983	35,045	33,715	0.13	0.08
7	2,855	21,100	(0.31)	(0.36)	0.11	31,479	22,768	107,733,077	75,886,051	22,768	21,100	(0.31)	(0.36)
8	1,088	7,500	(0.75)	(0.84)	0.63	31,479	10,339	575,014,240	446,917,103	10,339	7,500	(0.75)	(0.84)
9	133	3,700	(0.99)	(0.97)	0.96	31,479	3,619	771,698,095	776,202,755	3,619	3,700	(0.99)	(0.97)
10	661	2,550	(0.86)	(1.01)	0.87	31,479	7,336	836,913,340	582,904,579	7,336	2,550	(0.86)	(1.01)
11	166	1,800	(0.98)	(1.04)	1.02	31,479	3,852	880,870,022	763,282,931	3,852	1,800	(0.98)	(1.04)
Sum					9.88			8,136,188,043	7,936,151,449	346,274	346,274		
Average	4,093	31,479								31,479	31,479		
Standard deviation	4,005	28,524								28,171	28,524		
Average $(n-1) = R$					0.9876			Regression SS/total SS = 0.9754					
Average $(n-1)$ squared $= R^2$					0.9754								

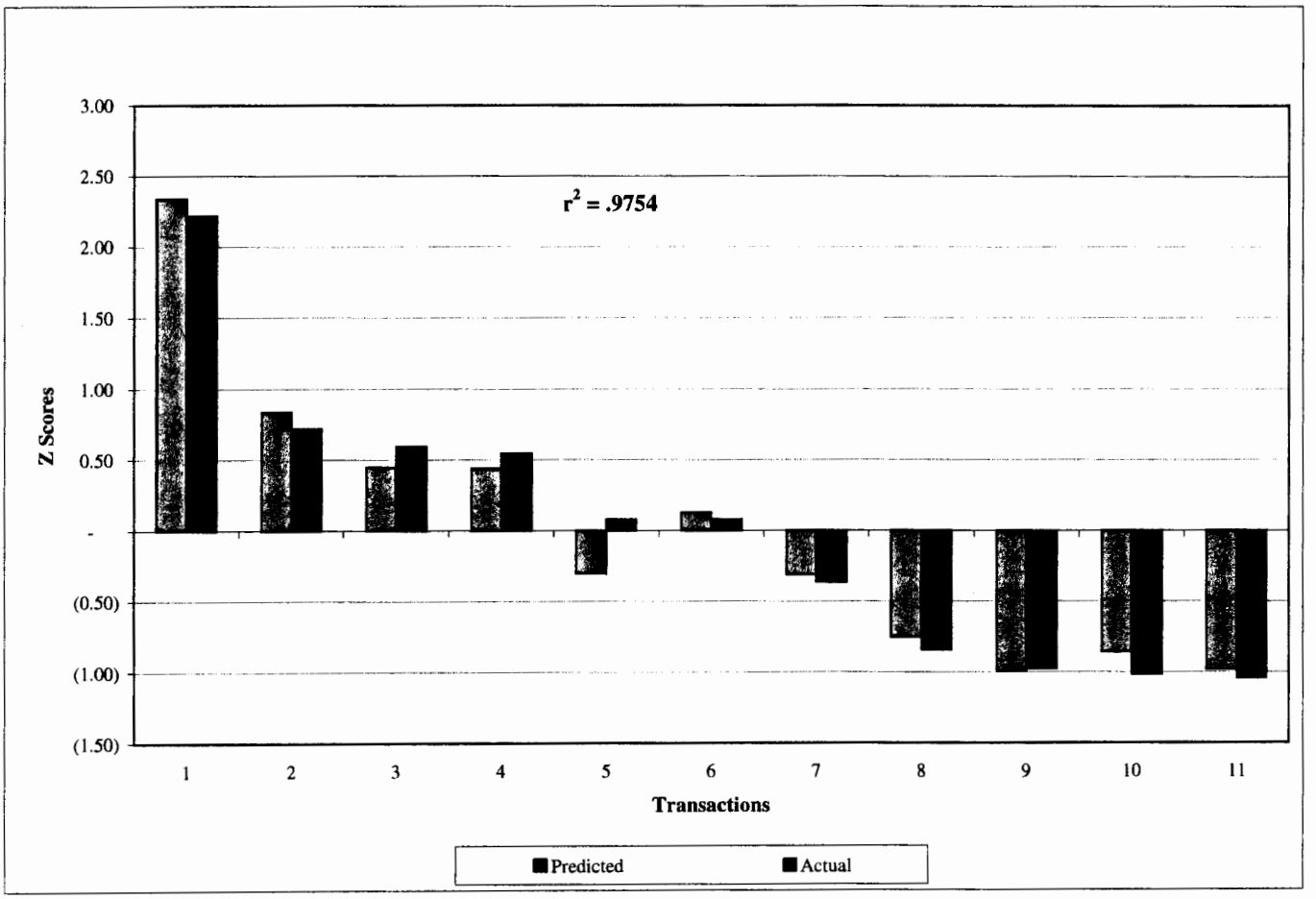

Figure 9
Comparative Variances

there is no implication that y necessarily depends on x in a causal or explanatory sense. This translates into the well-known admonition that correlation does not equal causation.

The second source of confusion is the conflation of the terms causation and prediction. Put another way, all causation models are prediction models, but prediction models are not necessarily causation models. For example, while we can predict revenue with time as the independent variable, and we can predict value using EBITDA as the independent variable, we would never assert that time drives or "causes" revenue, but we are very comfortable saying that a company's EBITDA drives, or "causes" its value. Causation, or explanatory, models attempt to match a theory with data and explain how x causes y. Prediction models have no such agenda— they merely wish to use independent variables that are associated with the dependent variable in such a manner as to make the former valid predictors of the latter. They do not "explain" anything more than the degree of covariance between and among the variables, i.e., the degree to which y and x move in tandem.

There are additional problems associated with the attempt to use the strength of the relationship between national housing starts and company revenues as a proxy for passive appreciation in value. The first is that revenue alone should never be a value driver. The better value driver is cash flow to equity or invested capital, because much that happens between the top line and the bottom line reflects management's ability to cope with exogenous forces. A reliance on revenue alone diminishes management's role in value creation. A second problem is that the effect size analysis of the combined, transformed model output indicates that time is the substantially more important predictor variable. This makes housing starts a necessary, but not sufficient variable for "explaining" revenue. Hawkins himself has definitively addressed the third, and last problem.[19] In that article, he argues cogently and convincingly that the multiplicity of factors that create

[19]George B. Hawkins, "Active versus Passive Appreciation—The Same Old Inflation Argument—But is It Valid?" *CCH Business Valuation Alert* 5(3) (April 2004).

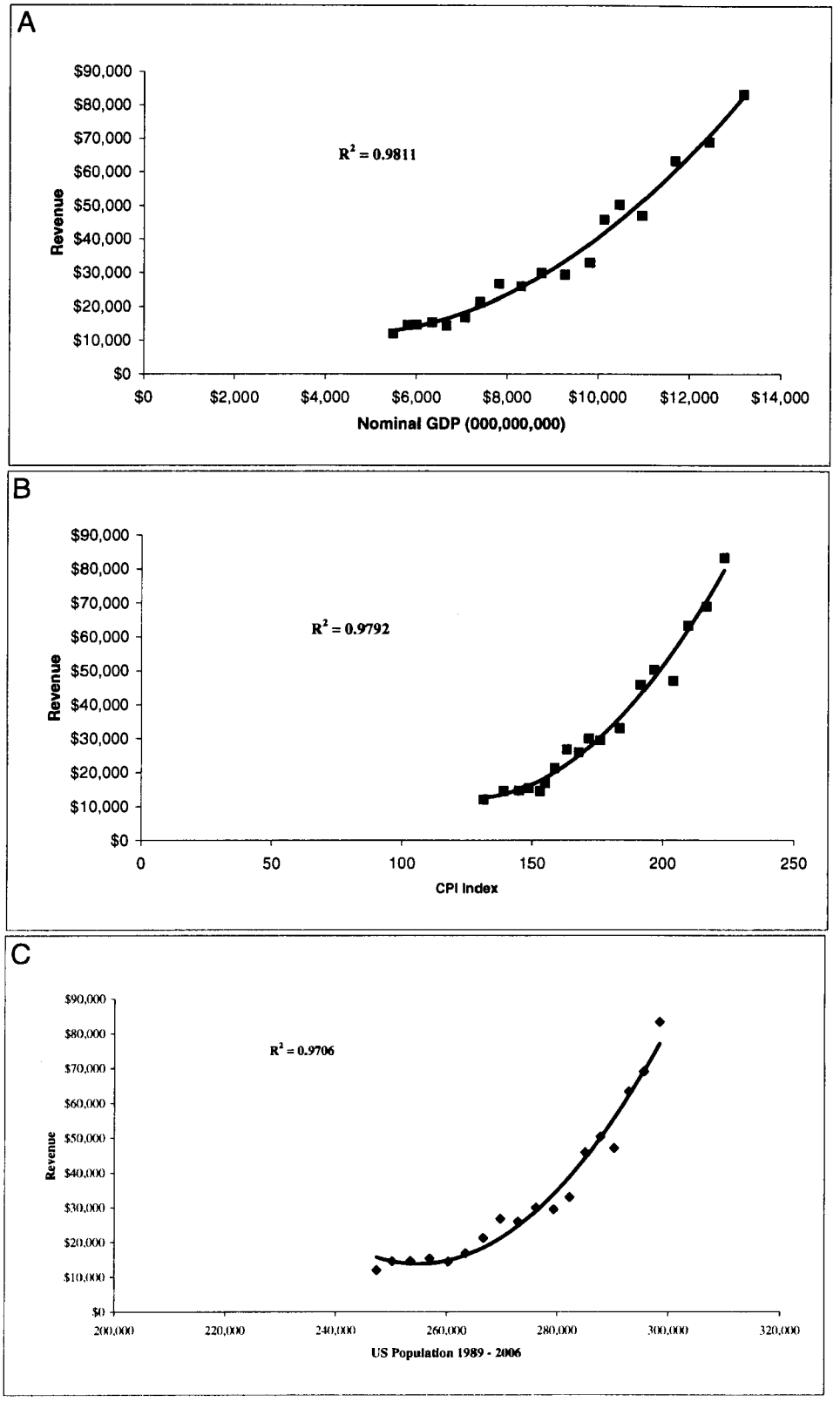

Figure 10
(A) Revenue versus Nominal GDP
(B) Revenue versus CPI Index
(C) Revenue versus Population

value cannot be overridden by "simplistically ascrib[ing] emphasis to one factor such as inflation." If he had substituted "national housing starts" for inflation, his conclusions would not have changed.

Figure 10A–C presents three scatterplots that "explain" the relationship between revenue and nominal Gross Domestic Product (GDP), the Consumer Price Index (CPI) and inflation, respectively. The fact that inflation, or any other exogenous factor that can be associated with revenue growth is now dressed up in the guise of RA cannot serve to negate Hawkins' previous arguments. None of these factors drives revenue, and revenue does not drive value. RA is a powerful tool for exploring relationships—it cannot prove causation. At best, in some circumstances, it can demonstrate it. The attempt to exploit R^2 as a measure of passive forces is a misuse of RA.

The concluding section of this article deals with Trout's letter to the editor regarding the Hawkins' article and its methodology and tools. As to methodology, I have no quarrel with the procedures broadly outlined in the letter that Trout used to deal with curvilinearity and serial correlation. I also agree with his finding that housing starts is a less important independent variable than time. What I disagree with is his assertion that Excel is not up to the task of performing the necessary statistical analysis that a "real" statistical program can carry out. It is true that the tools that come with Excel's Analysis ToolPak and the statistical functions that are accessed through the Paste Function button are inadequate. However, there is available a host of freeware add-ins,[20] very inexpensive add-ins that can be purchased,[21] and CD-ROMs that come with textbooks[22] that contain all sorts of analytical tools included with the price of the textbook. Also, the standalone programs do not come with an optimizer like Solver. Unlike the standalone programs, which can cost up $2,500, Excel is a sunk cost—the marginal cost, if any, of the add-ins is de minimis. In fact, all of the analysis for this article was done with Excel and its add-ins.

Regression analysis is a tool that should find its way into every valuation analyst's toolkit. It can be used to determine value under the market approach, and it can be used to forecast sales, cost of sales, and overhead costs under the income approach. However, its power is directly related to its proper use. An approach that always defaults to simple linear regression of y against x will only coincidentally give a proper response. A rigorous, time-consuming, analytical process that tests various models is the only way to get a result that will stand up in court.

[20]For example, including but not limited to: Essential Regression, Gerry's Stats Tools.
[21]For example, including but not limited to: Analyze-it, SigmaXL, and WinSTAT.
[22]For example, including but not limited to: Managerial Statistics, Data Analysis with Microsoft Excel.