EXHIBIT 2

LITIGATION SERVICES HANDBOOK

The Role of the Financial Expert

Third Edition

Edited by

ROMAN L. WEIL

MICHAEL J. WAGNER

PETER B. FRANK

JOHN WILEY & SONS, INC.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

effective way for the opposing party to probe the damages analysis prior to trial. Using a *Daubert* challenge to disable a damages analysis is relatively new, and it remains to be seen if this tactic is a successful way to disqualify an expert whose analysis, although novel in some way, nonetheless uses standard economic principles. Chapter 2 also discusses *Daubert* and laws governing expert witness testimony.

5.3 ISSUES COMMON TO MOST DAMAGES STUDIES. Throughout this discussion, I assume that the plaintiff is entitled to compensation for losses sustained from a harmful act of the defendant. The harmful act may be one whose occurrence itself is wrongful, as in a tort, or it may be a failure to fulfill a promise, as in a breach of contract. In the first instance, damages have traditionally been calculated under the principle that compensation should place the plaintiff in a position economically equivalent to the plaintiff's position had the harmful event never occurred. In applications of this principle, either *restitution damages* or *reliance damages* are calculated. These two terms are essentially synonyms with respect to their economic content. The term *restitution* is used when the harmful act is an injury or theft and the defendant is unjustly enriched at the expense of the plaintiff. The term *reliance* is used when the harmful act is fraud and the intent of damages is to place the plaintiff in as good a position as if no promises had been made. In breach of contract, damages are generally calculated under the expectation principle, where the compensation is intended to replace what the plaintiff would have received if the promise or bargain had been fulfilled. These types of damages are called *expectation damages.*

In this section, I review the elements of the standard loss measurement shown in Exhibit 5-1. For each element, there are several areas of potential dispute. The sequence of questions posed in this section should identify most of the areas of disagreement between the damages analyses of opposing parties.

(a) Characterization of the Harmful Event

(i) How Was the Plaintiff Harmed, and What Legal Principles Govern Compensation for the Harm? The first step in a damages study translates the legal theory of the harmful event into an analysis of the economic impact of that event. In most cases, the analysis considers the difference between the plaintiff's economic position if the harmful event had not occurred and the plaintiff's actual economic position. The damages study restates the plaintiff's position "but for" the harmful event; this step is often called the *but-for analysis.* Damages, then, are the difference between the but-for value and the actual value.

In cases where damages are calculated under the restitution-reliance principle, the but-for analysis[5] posits that the harmful event did not occur. In many situations—such as injuries resulting from accidents—the but-for analysis presumes no contact at all between the parties. Damages are the difference between the value the plaintiff would have received had there been no contact with the defendant and the value actually received.

Expectation damages[6] generally arise from the breach of a contract. The harmful event is the defendant's failure to perform. Damages are the difference between the value the plaintiff would have received had the defendant performed his obli-

(b) Regression Analysis. In simple terms, regression analysis attempts to find a relationship between cost drivers or volumes (called the *independent variables*) and a particular cost of interest (called the *dependent variable*). Regression analysis can handle relationships between a dependent variable and multiple independent variables. A simple regression has only one independent variable; multiple regression has more than one.

Regression usually requires a computer program. Numerous general statistics packages and specific regression packages exist for most computers. For a sample output of a regression program, see Exhibit 7-5 in Section 7.6(g).

In practice, most regression analysis is linear. The term *linear* refers to the fact that the measured relationship can be drawn as a straight line on a graph. (With more than one independent variable, the graph becomes difficult or impossible to draw, but mathematically the result is equivalent to a straight line.) Thus, to apply linear regression, the relationship between the independent and dependent variables in the relevant range should approximate a straight line. This means, for example, that a one-unit change in production should have the same effect on costs with low output and high output.

This restriction is not as onerous as it may seem. First, in practice, many costs approximate linear behavior. If they do not, often they are linear in some limited range of interest or a mathematical transformation can make them linear (see Section 7.6(f)). In addition, a regression analysis provides diagnostic measures that enable the analyst to ascertain whether the assumption of linearity appears reasonable in a particular instance.

Regression analysis refers to the particular hypothesized relationship between dependent and independent variables as a *model*. A simple linear regression model is expressed as:

$$Y = aX + c$$

where Y is the dependent variable (say, costs), X is the independent variable (say, units produced), a is the regression coefficient of X, and c is the constant term of the regression.

Consider a simple example. Suppose we produce widgets. A production run costs \$100 to set up, and each widget costs \$2 to produce after setup. The resulting equation is

$$\text{Cost} = \$2 \times (\text{Number of widgets}) + \$100$$

Exhibit 7-2 shows a graph of this simple linear relation.

In regression, we simply turn the problem around. We would know, say, the number of widgets produced each month and the total monthly cost. We would hypothesize a model of the form

$$\text{Cost} = a \times (\text{Number of widgets}) + c$$

The goal of regression is to estimate values for a and c. The estimation uses a set of mathematical equations embedded in a regression computer program.

Analysts do not limit themselves to a single independent variable when constructing a model. For example, if on the same production line we produced both widgets and gizmos, the model might look like

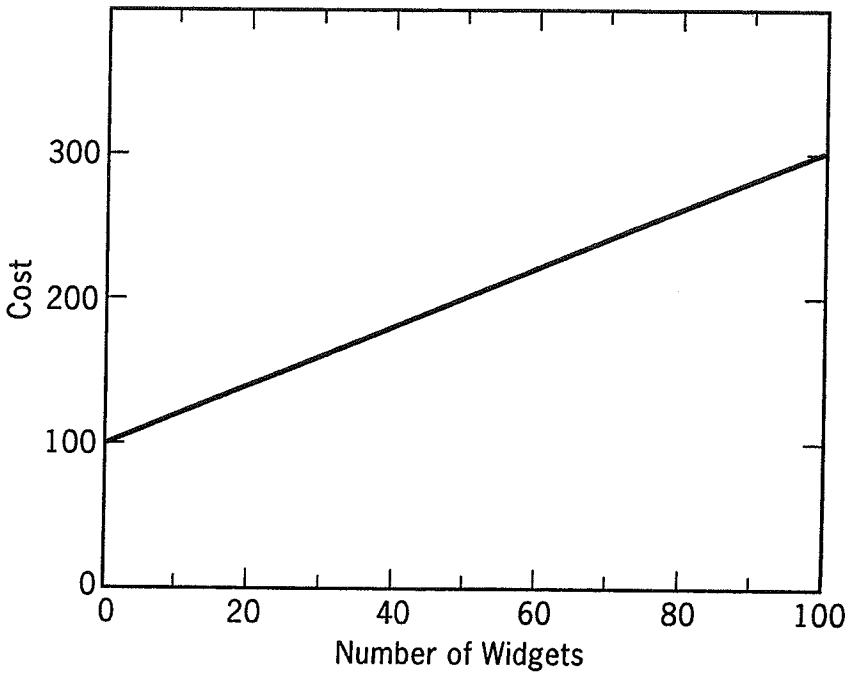

Exhibit 7-2. Example of Linear Cost Relationship

$$\text{Cost} = a_1 \times (\text{Number of widgets}) + a_2 \times (\text{Number of gizmos}) + c$$

In this case, a_1 equals the incremental cost per widget produced, a_2 equals the incremental cost per gizmo produced, and c equals the constant (fixed) cost per period.

We may have any number of independent variables. The equation for the general linear model with n independent variables is

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + c$$

In applying regression analysis to cost estimation problems, the analyst must specify an explicit model for analysis. Many naive users of regression make the mistake of simply plugging data into a program and using the results. This can lead to nonsense results. Acknowledged or not, an underlying model and assumptions always exist. The careful analyst recognizes this and makes sure that the model describes the intended type of cost relation.

In specifying a model, the analyst draws on knowledge of the economic and physical behavior of the variables involved. For example, as production volume increases, one might reasonably expect costs to increase directly, to increase but at a decreasing rate, to increase directly with a percentage change in volume, or to increase with the combined change in two or more different products.

One might also consider volume discounts, overtime costs, cost changes expected during expansion compared with changes expected during contraction, and technological change, for instance. Each of these implies a different model.

Sometimes the an
problem and app
other aspects.

Building an ex
sions of a model
the dependent va
found to have a n
and refine the mo
just to see if they
lead to substantia
lem, the analyst n

(c) Volume Data. perior to sales (in
tion (see Section 7
but not units of
price, on a period
take care to matcl

(d) Inflation and C will undoubtedly
analysis must sep
els. Typically, the

Adjusting for i
dex. The most cor
Price Index (PPI),
base year into wh
affect the analysi
when one uses th
cost data by the
question. (In this
monthly, quarterl

A

To see why thi
lowing productio

Year	Units
1	10
2	15
3	20

Without consider
from $20 in Year

Sometimes the analysis needs to treat some aspects of the situation as a regression problem and apply different techniques (as discussed earlier in the chapter) to other aspects.

Building an explanatory model often becomes an iterative procedure. Early versions of a model should include those variables suspected of having an effect on the dependent variable. In subsequent iterations, the analyst can delete variables found to have a negligible effect or found to duplicate the effects of other variables and refine the model as necessary. A model should not, however, include variables just to see if they are significant, a practice sometimes called *data mining*, which can lead to substantial errors in estimates of statistical significance. To avoid this problem, the analyst needs a theory of how costs behave before running regressions.

(c) Volume Data. The preferred measure of production volume is units. This is superior to sales (in dollars) because it avoids problems of price changes and inflation (see Section 7.6(d) below). When the sales dollar value of production is known but not units of production, one might divide the dollar volume by the selling price, on a period-by-period basis, to estimate units produced. However one must take care to match costs to production volumes, not to sales volumes.

(d) Inflation and Cost Data. Inflation has been part of the economy for decades and will undoubtedly remain so in the future. To obtain meaningful cost estimates, the analysis must separate the effects of inflation from those of changing activity levels. Typically, the analyst does this by restating cost data to *constant* or *real* terms.

Adjusting for inflation can be simple. First, choose an appropriate inflation index. The most common ones are the Consumer Price Index (CPI) and the Producer Price Index (PPI), both issued monthly by the U.S. government. Then choose some base year into which to convert all dollar amounts. The choice of base year will not affect the analysis, but interpreting the magnitude of the results becomes easier when one uses the most recent year as the base year. Next, multiply each period's cost data by the ratio of the base year inflation index to the index for the year in question. (In this discussion, we talk about *years*. The principle applies equally to monthly, quarterly, or other periodic data.) In equation form,

$$\text{Adjusted cost} = \text{Original cost} \times \frac{\text{Base year index}}{\text{Current year index}}$$

To see why this adjustment is necessary, consider an example. Assume the following production volume, costs, and inflation:

Year	Units	Total Cost ($)	Average Cost/Unit ($)	Inflation Index (Yr 1 = 100)	Average Cost/Unit (Deflated $)
1	10	200	20	100	20
2	15	500	33	200	17
3	20	900	45	300	15

Without considering inflation, it appears that the average cost per unit increased from $20 in Year 1 to $33 in Year 2 to $45 in Year 3. A regression with units as the

independent variable and total cost as the dependent variable would generate the following equation:

$$\text{Total cost} = -\$517 + \$70 \times \text{Units}$$

The regression leads us to believe that the incremental cost per unit equals $70. The last column in the tabulation shows this result to be misleading. After adjusting for inflation, we find in fact that the average cost per unit declines over time. The regression results on the adjusted costs yield the following:

$$\text{Total cost} = \$100 + \$10 \times \text{Units}$$

This indicates that in inflation-adjusted (Year 1) dollars, the incremental cost per unit equals $10. If we want to know the estimated incremental cost per unit in Years 2 and 3, we simply reinflate the estimate, getting an incremental cost per unit of $20 in Year 2 and $30 in Year 3. These results differ significantly from the $70 measured without an inflation adjustment.

As a rule of thumb, one should consider an inflation adjustment whenever the analysis uses more than two years of data or whenever inflation exceeds 10 percent during the time span of the data. The inflation adjustment has three steps:

1. Make the inflation adjustment to restate the data in *constant* dollars.
2. Perform the analysis.
3. Restate the results in *current* dollars of the time period(s) of interest.

(e) Indicator (Dummy) Variables. The use of indicator (sometimes called *dummy*) variables provides a valuable regression technique. An indicator variable is simply a variable set equal to zero for some observations and 1 for others. It serves as a flag to indicate that something is different about certain observations.

For example, consider a series of monthly data from accounting records spanning many years. In many accounting records, the entry in December may reflect year-end adjustments and so not reflect a typical month. Rather than ignore the December data, one can create an indicator variable to attempt to capture the year-end effect by setting a new variable with the value 1 in each December and 0 in all other months. The regression can then include this variable (along with the accounting data and other relevant variables). By considering the statistical relevance of the indicator variable's coefficient (see following discussion), the magnitude and importance of the December effect can be measured and extracted from the measurement of the other variables.

Analysts may use other useful indicator variables, such as the following:

- A variable that takes the value 1 for the affected period (e.g., periods of alleged harm) and the value 0 otherwise
- A variable that takes the value 1 during an anomalous period (say, a fire or strike) and 0 otherwise
- A variable that takes the value 1 during a particular season of the year and 0 otherwise
- A variable that takes the value 1 for each period after the firm has built a new factory and 0 for each period before then

Dummy variables can capture seasonal effects. One can use more than one indicator variable at once to measure effects of each season separately. To measure a quarterly effect, establish three dummy variables, each taking the value 1 for a different season of the year. For example, the Spring dummy variable has the value 1 for each period that occurs in April or May or June or the value 0 otherwise. Then put all three dummy variables into the regression equation. The resulting coefficients and statistics measure the effect of each season.

In using dummy variables, we usually define one fewer variable than there are possible situations, such as seasons. If we assign indicator variables to all situations, the model will be unable to estimate a constant term separate from the effects of the indicator variables.

(f) Transformations. A transformation is a mathematical manipulation applied to the data or the model before fitting the regression. Transformations convert the data into a form appropriate for the regression as specified. The transformation can (1) make the model amenable to analysis or (2) make the data more suitable to the model.

For example, consider an exponential model: $Y = aX^b$. This is not linear; it does not fit into the general equation discussed above in Section 7.6(b). By using a transformation, however, we can make it linear. We take the logarithm of both sides, getting

$$(\ln Y) = (\ln a) + b(\ln X)$$

By treating the variable $\ln Y$ as the new dependent variable, $\ln X$ as the new independent variable, and $\ln a$ as the new constant term, we have transformed the equation into one suitable for estimation by linear regression.

We have already seen an example of the second kind of transformation: the adjustments for inflation discussed in Section 7.6(d). Other examples of transformations include:

- Multiplying two variables together, as when we have data on price and quantity sold, with the product representing total revenue
- Dividing one variable by another
- Raising a variable to a power, as when economists estimate production functions involving the interplay of capital and labor
- Taking the logarithm of a variable (see Section 7.6(f) above)
- Creating an indicator variable (see Section 7.6(e))
- Computing the change in a variable from one period to the next

These transformations allow the linear regression procedure to fit what would otherwise be nonlinear situations.

(g) Sample Regression Output. Exhibit 7-3 shows a sample set of data listing monthly production and cost data of widgets from the Widget Manufacturing Company for a three-year period. A graph of the data appears in Exhibit 7-4. The graph shows that the data are well-behaved and exhibit a linear relation and are therefore suitable for regression.

Month	Units	Cost
January '96	42.3	3406.8
February '96	53.6	4367.4
March '96	51.6	3957.5
April '96	60.1	3659.4
May '96	41.5	2535.5
June '96	59.4	3685.7
July '96	59.1	3650.8
August '96	52.1	3863.7
September '96	58.4	4555.4
October '96	53.1	3871.5
November '96	50.4	2995.3
December '96	57.3	4490.8
January '97	70.6	5300.0
February '97	35.6	4287.8
March '97	49.5	3967.2
April '97	52.1	3573.2
May '97	63.2	4484.3
June '97	63.6	3760.2
July '97	63.2	4446.8
August '97	81.8	5600.0
September '97	73.8	4875.7
October '97	72.8	4378.7
November '97	81.3	5274.3
December '97	82.8	5276.1
January '98	74.0	4217.8
February '98	64.6	4938.0
March '98	65.1	5051.7
April '98	72.4	4397.4
May '98	88.8	4914.3
June '98	97.2	5852.3
July '98	95.1	6515.9
August '98	80.5	5250.9
September '98	101.4	5867.1
October '98	85.0	5331.0
November '98	81.0	4796.0
December '98	100.3	5886.3

Exhibit 7-3. Monthly Production and Cost Data

A regression was fit of the form Cost = $a \times$ Units + c. An extract from the computer output appears in Exhibit 7-5. At the top of the output, we see that the dependent variable is Cost. After some identifying information, we come to the regression results. The first column shows that the regression includes two variables: the constant (C) and the real variable, Units. The next column shows the computed

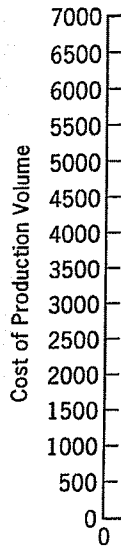

Exhibit 7-4. N

Dependent V
Number of O

Variable	C
Constant	
Units	

R-squared, 0.78
1.78.

Exhibit 7-5. S

coefficients. R
equals

In other w
term equals $
cost. Althougl
pendent varia
Section 7.6(j)(i

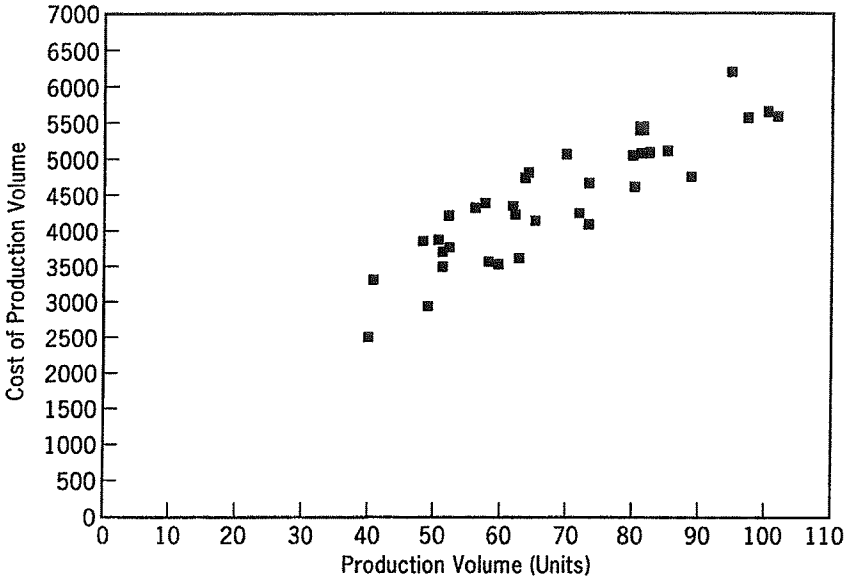

Exhibit 7-4. Monthly Production and Cost Data

Dependent Variable is COST
Number of Observations: 36

Variable	Coefficient	Std. Error	*t*-Statistics	Two-Tailed Significance
Constant	1256.50	308.82	4.07	0.00
Units	47.90	4.39	10.91	0.00

R-squared, 0.78; adjusted *R*-squared, 0.77; S.E. regression, 421.83; Durbin-Watson statistic, 1.78.

Exhibit 7-5. Sample Regression Results—Monthly Production and Cost Data

coefficients. Reading these values, we find that the resulting regression equation equals

$$\text{Cost} = \$47.90 \times \text{Units} + \$1{,}256.50$$

In other words, the incremental cost per unit equals $47.90, and the constant term equals $1,256.50. (Analysts often loosely interpret the constant as the fixed cost. Although this holds true if the regression is relevant for values of the independent variable near zero, the analyst should cautiously make this assertion. (See Section 7.6(j)(*i*), *relevant range*.)

(h) Statistical Measures and Statistical Tests. The remaining output provides various measures of how well the regression equation fits the data. One advantage of regression is that it not only provides estimates of the costs and other parameters but also measures the accuracy of those estimates. These measures fall into two groups:

1. How well the model as a whole fits the data
2. The importance of specific variables

The important measures in the first category are the square of the multiple correlation coefficient (*R*-squared or R^2), adjusted R^2, the standard error of the regression, and the Durbin-Watson statistic. In the second category, we find the standard error of each variable and the related *t*-statistic and measures of significance.[3] Each of these appears in the sample output in Exhibit 7-5.

R-squared or R^2, the square of the multiple correlation coefficient, ranges from 0.0 to 1.0. The number R^2 measures the percentage of total variation of the dependent variable from its mean value that the regression equation explains.

If the regression equation explains the variation in the dependent variable perfectly, then R^2 equals 1. If, on the other hand, the regression equation cannot explain the variation any better than the mean itself, then R^2 equals zero.

A high value of R^2 leads us to say that the regression equation explains a large portion of the variation in the dependent variable. Thus, R^2 invites the interpretation that it measures the *goodness* of fit of a regression. This need not be so. In practice, use caution when drawing conclusions about a regression from the value of R^2. Whether R^2 provides useful information depends on whether the question at issue involves the prediction of the dependent variable or the coefficient of some independent variable. If the issue involves one of the independent variables (for example, variable cost), then the R^2 will have lesser relevance. Interpreting R^2 requires experienced statistical judgment.

In our sample printout (Exhibit 7-5), the R^2 equals .78, or 78 percent. That figure, combined with a look at the graph and the residuals (see later discussion), indicates that the regression equation predicts fairly well.

The *adjusted* R^2 is a simple variant of the R^2. In the basic formulation of R^2, adding any independent variable to the regression equation, even if correlated randomly with the dependent variable, cannot lower the R^2. The adjusted R^2 addresses this problem. In a loose sense, it indicates whether adding another independent variable provides information worth its cost. Adjusted R^2 penalizes the computation for adding an uncorrelated variable. For this reason, analysts usually prefer the adjusted R^2 to simple R^2. The adjusted R^2 makes an important difference only if one has a small sample size. When the number of data points exceeds the number of independent variables by at least 20, the R^2 and adjusted R^2 have nearly equal values. In our example, the adjusted R^2 equals 77 percent, almost exactly equal to the unadjusted R^2.

The *standard error (SE)* of the regression measures the size of the estimation errors made by the regression equation. Loosely speaking, the standard error of the regression measures the accuracy with which the derived equation predicts the dependent variable for a given value of the independent variable. The units are the same as those of the dependent variable. It is analogous to the standard deviation of a variable; in fact, it is the standard deviation of the residuals. Because one can

think of this value
are met) that 95 pe
actual value. In ou
mately $422. Thus
actual value.

In economic da
ple, quantity outpu
cessive months oft
order) *serial correla*
Serial correlation c
sults. For example,
November and De
estimated equatior
also overpredicted
rection as in the pr

The *Durbin-Wat*
in the residuals. Tl
to 4 (perfect negati
Interpreting the D
bles. As a rule of th
a few independent
cause for concern
higher-order seria
ple, the Durbin-W
order serial correl

These statistics
second category c
tioned, the import
lated *t*-statistics ar

The computed
the data do not lie
measures the accu
dard error is analo
can construct conf
dard error. For exa
of the coefficient l
value. In our exam
cient of *units* equa
(i.e., two times the

The *t-statistic* is
coefficient differs
t-statistic for *units*

The analyst con
statistics books an
the *t*-statistic excee
(unknown) value

think of this value as a standard deviation, it is roughly true (if certain assumptions are met) that 95 percent of all predictions will lie within two standard errors of the actual value. In our example, the standard error of the regression equals approximately $422. Thus approximately 95 percent of our estimates lie within $844 of the actual value.

In economic data, successive values of a variable may be correlated. For example, quantity output or the inflation rate tends to change slowly over time, with successive months often close together. In statistical terms, this is called positive (first-order) *serial correlation*. (Negative serial correlation also occurs, but less commonly.) Serial correlation can cause problems in drawing conclusions from regression results. For example, consider a retail store with seasonal sales, larger than average in November and December and smaller than average in January and February. The estimated equation will likely overpredict sales in a period following one where it also overpredicted sales. The tendency of errors in prediction to be in the same direction as in the preceding period is serial correlation (of residuals).

The *Durbin-Watson statistic* measures the extent of (first-order) serial correlation in the residuals. The statistic can range from 0 (perfect positive serial correlation) to 4 (perfect negative serial correlation). A value of 2 indicates no serial correlation. Interpreting the Durbin-Watson statistic requires looking up values in several tables. As a rule of thumb, however, in an analysis with at least 50 observations and a few independent variables, a Durbin-Watson value below 1.5 or above 2.5 signals cause for concern. (Correcting for a serial correlation problem and identifying higher-order serial correlation go beyond the scope of this chapter.) In our example, the Durbin-Watson statistic equals 1.78, indicating no problems with first-order serial correlation.

These statistics apply to the regression equation as a whole. We now turn to the second category of statistics, those that describe individual variables. As mentioned, the important statistics are the standard error of each variable and the related *t*-statistics and measures of significance.

The computed regression coefficient is subject to some degree of error because the data do not lie perfectly on a straight line. The *standard error* of each variable measures the accuracy of the variable's coefficient estimate. In this way, the standard error is analogous to a standard deviation. Under the usual assumptions, we can construct confidence bounds for the coefficients using a multiple of the standard error. For example, we can be roughly 95 percent confident that the true value of the coefficient lies within plus or minus two standard errors of the computed value. In our example, an approximate 95 percent confidence bound for the coefficient of *units* equals $47.90 per unit plus or minus approximately $8.78 per unit (i.e., two times the standard error of 4.39).

The *t-statistic* is a statistical test of the hypothesis that the true value of the coefficient differs from some specified number, typically zero. In our example, the *t*-statistic for *units* equals approximately

$$47.90/4.39 = 10.91$$

The analyst compares the *t*-value to a standard table of *t*-statistics found in most statistics books and included in most statistical computer software. In general, if the *t*-statistic exceeds 2.0, we can conclude with 95 percent confidence that the true (unknown) value of the coefficient does not equal zero.

If the coefficient were truly zero, the corresponding variable would contribute nothing to the equation. Accordingly, a test of the hypothesis that the coefficient differs from zero is a test to ascertain whether the corresponding variable contributes to the predictive power of the regression results.

Loosely speaking, the *level of significance* is the probability that the true value of the coefficient equals zero (or some other specified critical value). This is just the probability value (from the standard table of *t*-statistics) that corresponds to the observed *t*-statistic. Significance level is usually measured as two-sided or two-tailed, meaning that we are equally interested in whether the computed value lies above or below the true coefficient. In our example (Exhibit 7-5), the *units' t*-statistic of 10.91 corresponds to a significance level of 0.00, indicating that the probability is near zero that *units* does not relate to the dependent variable.

(i) Examination of Residuals. The analyst can use the regression equation to predict values for the dependent variable using the known value(s) of the independent variable(s). The differences between the predicted values of the dependent variable and the actual values are called *residuals*. The pattern of the residuals can help diagnose the regression equation's appropriateness.

The analyst usually examines residuals graphically. Plot the residuals against other variables to see if any patterns exist. If the data have a natural sequence (e.g., they are time sequenced), analysts most commonly use a plot against time. Sometimes analysts plot the residuals against the estimated value of the dependent variable. In a properly specified regression equation, the residuals should just be the random errors inherent in the data. Accordingly, the residuals should not exhibit any particular pattern but should demonstrate a random (normal) distribution around zero. If the pattern does not look random, the regression may not be appropriate.

Consider the patterns in Exhibit 7-6. In part (a), the residuals appear normal, indicating a satisfactory regression fit. Analysts hope for this pattern.

In part (b), the residuals begin negative, turn positive, and then become negative again. This pattern suggests that the data are not linear. Fitting a straight line is not appropriate. The analyst should construct a different model with an appropriate transformation or nonlinear form.

In part (c), a positive residual will likely be followed by another positive residual, and similarly for negative residuals. This indicates the presence of first-order serial correlation, meaning that one high value will likely be followed by another high value, as happens with time-series data subject to seasonal patterns. Details of corrective action go beyond the scope of this chapter.

In part (d), the variability of the residuals grows over time. This problem is called *heteroscedasticity*. It may indicate that the regression results are not as significant as the diagnostics indicate. Again, the solutions to this problem go beyond the scope of this chapter.

Finally, in part (e), we see that one residual value appears misbehaved. Analysts refer to this value as an *outlier*. The analyst should investigate this data point. It may be an error. If not an error, it may include significant adjusting entries. (Commonly, the last month of the year is an outlier.) Analysts may decide to rerun the analysis without the offending point (if substantive reasons exist) to spread the year-end adjustment throughout the year, or to analyze the adjustment separately.

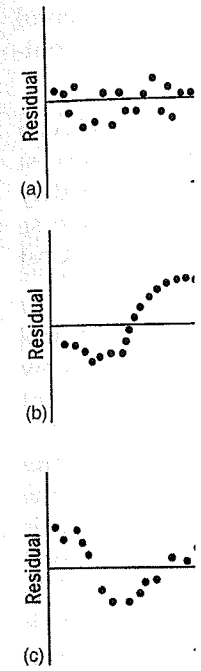

Exhibit 7-6. Ex

If outliers regu
indicator varia

(j) Other Statist

(i) Relevant Rang
For example, o
5,000 to 10,000
precisely, the r
analyst could
In our exam
in the range of
sions significa
reasonably est
accounting an
Often there
quired inferen
recognize that
other forms o
volved in the
As mention
simple cost–v

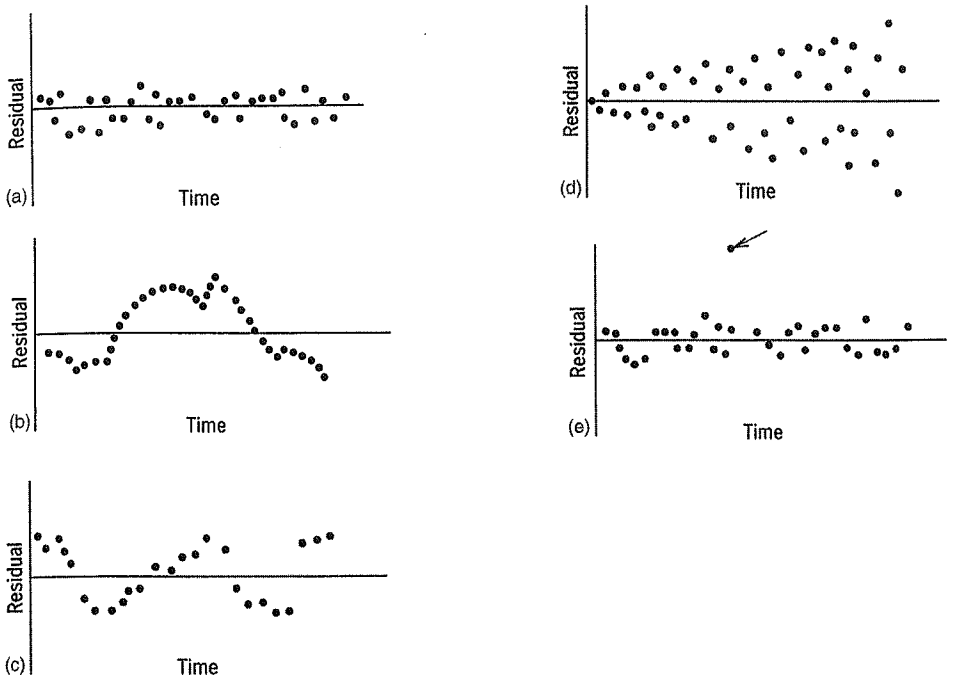

Exhibit 7-6. Examination of Residuals

If outliers regularly recur (every December, for example), one may want to use an indicator variable to measure the year-end effect.

(j) Other Statistical Issues

(i) Relevant Range. Cost data come from a specific range of production volumes. For example, over the last few years a particular factory may have produced from 5,000 to 10,000 units per month. This defines the approximate *relevant range*. More precisely, the relevant range equals the range of production volumes for which the analyst could reliably extrapolate cost.

In our example, we could make conclusions more confidently about production in the range of 5,000 to 10,000 units per month than we can extrapolate the conclusions significantly above or below this range. Although the range of production reasonably estimates the relevant range, ascertaining the relevant range requires accounting and statistical judgment.

Often there is no choice but to extrapolate outside the relevant range; the required inference lies there, and no other reasonable approach presents itself. We recognize that this often occurs and merely counsel caution. The analyst should use other forms of evidence, including substantive judgments from personnel involved in the production process in question to validate the extrapolation.

As mentioned in Section 7.6(g), analysts sometimes refer to the constant in a simple cost–volume regression as the estimate of fixed costs per period. One must

use caution with this interpretation, particularly when, as is usually the case, production volumes significantly exceed zero so that production of zero units lies outside the relevant range.

(ii) Nonlinear Cost Relationships. Not all costs, even within a relevant range or over a specified period, change linearly. In many operations, a diminishing marginal cost of production exists where variable costs per unit may become disproportionately lower with each new level of production. In other operations, costs may increase at certain levels of production within the existing capacity of the facility. Cost analyses should not assume linear costs without reason.

(iii) Causation vs. Correlation. In regression, one of the measures of goodness-of-fit is the square of the multiple correlation coefficient, or R^2. This number measures how closely the dependent variable changes in response to changes in the independent variable(s). This is called *correlation.*

This number does not measure causation. Simply because two variables correlate, or change in the same direction at the same time, does not mean that the changes in one cause the changes in another. Without other information, it is equally likely that the first variable causes changes in the second, that the second causes changes in the first, or that some third, unidentified variable causes both. Accordingly, in reaching a conclusion about correlation, one can say that the variables move together but not that changes in one cause changes in the other. Conclusions about causation generally come from nonstatistical analyses or *a priori* knowledge.

(iv) Number of Data Points Required. No hard-and-fast rule exists about the number of data points required for a regression. Certainly, the more the better, provided that the expected relationship between the variables remains stable over time. Some observers suggest a minimum of 15 or 20 data points. One can, however, run valid regressions on as few as five or six points. If you use fewer than the recommended minimum, consult a professional statistician to make sure your methods and conclusions are valid.

(v) Multicollinearity. When two data series change almost in unison, they are highly correlated. In multiple regression, problems will occur if two (or more) independent variables are highly correlated. Analysts call this *multicollinearity.* In that case, the regression formula cannot assign valid coefficients for both of the independent variables.

An example will clarify this problem. Consider sample data as follows:

	Independent Variables	
Dependent Variable	**#1**	**#2**
10	1	100
20	2	200
30	3	300
40	4	400

In this case, one possible equation would be

$$\text{Dependent variable} = 10 \times \text{Independent \#1}$$

An equally valid choice is

$$\text{Dependent variable} = .1 \times \text{Independent \#2}$$

If the analyst includes both independent variables in the regression, no mathematical way exists to assign coefficients to the two variables. Instead, the regression algorithms usually cannot compute anything at all. (A realistic example of this would occur if one were attempting to derive the separate cost of producing suit jackets and suit pants in a factory producing only men's two-piece suits. The data would support a reasonable estimate of the total cost of a suit, but not separate estimates for coats and jackets.)

Only rarely will an analyst encounter two variables 100 percent correlated, as with the two independent variables in our example. With two highly (but not perfectly) correlated variables, regression usually works and gives apparently correct results. Because the technique could not evaluate how much weight to assign each of the two variables, however, the *t*-statistics of the correlated variables will signal difficulty. Typically, correlations of over .80 between two independent variables will cause difficulty.

Analysis can detect multicollinearity in several ways. Most regression program packages will provide statistics for correlations between variables in a multiple regression equation. If so, look for high correlations between two independent variables. If the variables vary together, it is unlikely that the regression can separate their effects. Second, look at the standard error of the variables. If multicollinearity exists, the regression diagnostics will indicate large standard errors, indicating little confidence in the estimates.

When multicollinearity exists between several variables, combine the correlated variables or eliminate all but one of them.

(vi) Changes in Cost Behavior over Time. An underlying assumption in regression is that the model describes the cost relationship. Consider a regression of the form: Cost = $a \times$ Units + c. This equation asserts that the coefficients are constant over time. If a cost relation shifts or evolves over time, the specified model becomes inappropriate.

Cost can shift gradually or suddenly. Gradual shifts can result from a slow evolution in processes or inattention of management to slowly increasing costs. Inflation can also lead to a gradual shift in the relation, but the analyst can handle inflation separately, as described in Section 7.6(d). Sudden shifts can occur when the firm introduces a new technique or process.

Statistical techniques can detect shifts. One technique explicitly includes a *time* term (e.g., January is 1, February is 2, . . . , January of the next year is 13) as an independent variable. If the resulting coefficient is significant, this suggests that some time trend exists. Typically, the change is not caused by time itself but by some underlying cause that moves with time. Whether the analyst must identify this underlying cause or whether use of the time variable itself is sufficient depends on the requirements of the analysis. Including a time variable usually does not substitute for making inflation adjustments.

Analysts sometimes plot the residuals in time sequence to find cost shifts. A pattern in the residuals may indicate some cost shifts. If the analyst suspects a cost shift at a particular date, either of two alternatives can help. First, with sufficient available data, one can perform the analysis using only the data before (or after) the shift. Second, an indicator variable (set to 1 for every period after the shift and to zero for every period before the shift) can pick up the effect of the cost shift. (See Section 7.6(e) on indicator (dummy) variables.)

(vii) Tests of Reasonableness. The analyst needs to confirm any cost estimate, statistical or otherwise, as reasonable. The reasonableness tests include the following:

- Compare the results of applying more than one estimating method. If the results are reasonable, the methods should yield approximately the same results, or one should have good reasons for a discrepancy.
- Compare the results to reality. For example, compare estimated costs at historical volumes to historical costs. Compare results at an assumed but-for volume with historical results (at some other date) at roughly the same volume. The results need not be the same, but differences should be reconcilable.
- Compare the results to independent cost estimates. For example, the company under study may have made cost forecasts as part of a business plan before the alleged liability acts occurred. Alternative industry statistics may provide a useful baseline.
- Consider the intrinsic reasonableness of the results. Do costs increase with volume? Are costs appropriately behaved compared to changes in production capacity?
- Finally, apply a test called *interocular inspection.* This is a tongue-in-cheek name for a real test, in which you stare at the results until the meaning hits you between the eyes. In other words, consider whether your results make sense.

(k) Other Statistical Techniques. The statistical discussions in this chapter focus on regression analysis. The analyst may use other statistical techniques in cost estimation. We mention some briefly; consult a statistician for further information.

(i) Time Series, ARIMA Models. Time series analysis refers to the analysis of any data sequenced over time. The time order provides an essential element of the analysis. Regression (using time as an independent variable) offers one part of time series analysis. Another approach to time series analysis is called ARIMA (autoregressive integrated moving average) modeling, sometimes termed *Box-Jenkins analysis.* This approach searches for recurring patterns in past history to forecast the future.

(ii) Survey Research for Comparable Entities. At times, the analyst needs information on averages for an industry. A survey of the industry, of customers, or of the public may be appropriate to measure some factor. Many rules and methods exist for conducting a statistically valid survey.

(iii) Statistical Sampling for Attributes. Statistical sampling estimates a characteristic of a population without observing every item. Use it when the cost of a complete enumeration becomes prohibitive. For example, if you need to establish the average

size of an invoic
vide usable resu

7.7 CONCLUSI
ing the tools of c
dustrial enginee
cost. If one appli
invalid. In cost e

NOTES

1. For an introdu
Managerial Accour
TX: The Dryden P
2. Two editors of
sis identified 25 p
chapter, identified
3. For a more com
Using Regression,
29); or Schroeder,
Guide, 1986 (Series
published by Sage

size of an invoice, sampling a statistically valid selection of invoices will likely provide usable results. The audit process frequently uses sampling.

7.7 CONCLUSION. Cost estimation occurs often in litigation analysis. By employing the tools of cost analysis, including cost accounting, statistics, economics, and industrial engineering, the analyst can arrive at accurate and defensible estimates of cost. If one applies the tools by rote or without requisite thought, the results may be invalid. In cost estimation, success requires careful thought and careful analysis.

NOTES

1. For an introduction to cost accounting, see M. W. Maher, C. P. Stickney, and R. L. Weil, *Managerial Accounting: An Introduction to Concepts, Methods, and Uses,* 6th ed. (Fort Worth, TX: The Dryden Press, 2000), or another cost accounting text.

2. Two editors of the *Handbook* once opposed each other in a litigation where account analysis identified 25 percent of costs as variable, but regression analysis, described later in this chapter, identified variable costs as 70 percent of the total.

3. For a more complete but still simple discussion of regression, see Achen, *Interpreting and Using Regression,* 1982 (Series on Quantitative Applications in the Social Sciences, Number 29); or Schroeder, Sjoquist, and Stephan, *Understanding Regression Analysis: An Introductory Guide,* 1986 (Series on Quantitative Applications in the Social Sciences, Number 57). Both are published by Sage Publications, Beverly Hills, CA.