APPENDIX E

Table of Contents

Paragraph *Page*

E-000 Graphic and Regression Analysis Techniques

E-001 Scope of Appendix E1

E-100 Section 1 --- Correlation and Use of Scatter Diagrams

E-101 Introduction E1
E-102 Audit Objective..... E1
E-103 Terminology E2
E-104 The Scatter Diagram..... E4
E-104.1 Purpose E4
E-104.2 Construction – Scatter Diagram..... E5
Figure E-1-1 Close Positive Linear Correlation..... E5
E-104.3 Interpretation E5
Figure E-1-2 Close Negative Linear Correlation..... E5
Figure E-1-3 Close Curvilinear Correlation..... E6
Figures E-1-4 through E-1-7 Correlation Examples..... E7
Figure E-1-8 Little or no Correlation – Lack of association between X and Y..... E8
Figure E-1-9 Reduced Correlation E8
E-104.4 Selection of Appropriate Scales E9
Figure E-1-10 and E-1-11 Choices of Scales for Scatter Diagrams E9
E-104.5 Use and Limitations..... E9
Figure E-1-12 Freehand Estimation of the Dependent (Y) Value from a Known Independent (X) Value E11

E-200 Section 2 --- Correlation and Regression Analysis

E-201 Introduction E12
E-202 Simple Linear Regression Analysis..... E12
E-202.1 Definition and Concept E12
E-202.2 Example of Regression..... E13
Figure E-2-1 Graph of Simple Linear Regression Line E14
E-203 Curvilinear Regression Analysis E14
E-204 Multiple Regression Analysis..... E15
E-205 Correlation Analysis..... E17

Paragraph *Page*

E-205.1 Interpretation of Coefficient of Determination E17
Figure E-2-2 Regression and Partitioning of Deviations from the Mean, Explained and Unexplained Sum of Squares E18
E-205.2 Determining the Existence of Correlation E19
Table E-2-1 Table of Significant Values of the Coefficient of Determination (R^2) E23
E-205.3 Evaluating the Extent of Correlation E23
E-205.4 Adjusted R^2 ..... E24
E-206 Adjustment for Economic Factors in Application of Regression to Wages E24
E-207 Other Considerations in Using Regression Analysis E25

E-300 Section 3 --- Time Series Charts

E-301 Introduction E28
E-302 Definition and Concept E28
Figure E-3-1 Number of Indirect Employees by Quarters E28
E-302.1 Comparison of Concepts E29
E-302.2 Application E29
E-303 Chance and Unusual Variations E30
E-304 Secular Trends E30
Figure E-3-2 Monthly Values vs. Monthly Average E32
E-305 Seasonal Variations E32
E-306 Business Economic Cycle E33
E-307 The Moving Average E33
E-307.1 Description E33
E-307.2 Construction – Moving Average E34
Figure E-3-3 Three-Month Moving Average E35
E-307.3 General E35
E-308 Adjustment of Data E36
E-309 Length of the Analysis Period E37

E-400 Section 4 --- Audit Applications of Graphic and Computational Techniques

E-401 Introduction E38
E-402 Identification of Departures from Historical Cost Patterns E38
E-402.1 General E38
E-402.2 Compilation of Data E38

Paragraph *Page*

E-402.3 Analyzing the Data E39
E-403 Evaluation of Proposed Overhead Costs E40
E-404 Evaluation of Allocation Bases E41
E-405 Evaluation of Proposed Labor Rates E41
E-405.1 General E41
E-405.2 Illustration of the Use of Analysis in the Evaluation of Labor Rates E42
Table E-4-1 Average Monthly Direct Labor Hourly Rates and Number of Direct Workers E43
Figure E-4-1 Average Monthly Direct Labor Rates and Number of Direct workers E44
E-405.3 Alternative Methods E45
E-406 Evaluating Standards E45

APPENDIX E

E-000 Graphic and Regression Analysis Techniques

E-001 Scope of Appendix

a. This appendix presents guidance on the use of graphs and correlation and regression analysis to help analyze contract costs and associated production data. The focus is on special cost evaluation problems in contract auditing.

b. The analytical procedures discussed in this appendix are not by themselves means of evaluating costs or cost estimates. When these analytical procedures are used under the appropriate conditions, in conjunction with other auditing procedures, they are valuable tools that can help establish an empirical basis upon which sound conclusions about the reasonableness and acceptability of contractor cost statements can be formulated.

c. Analytical procedures can be used as risk assessment procedures to identify areas for further evaluation and may also be performed as part of the substantive procedures that the auditor designs based on the risk assessment to obtain sufficient and appropriate audit evidence. Analytical procedures and inquiry alone are not sufficient to support the high level of assurance provided in examination engagements. Auditors should generally perform tests of details for all significant/sensitive accounts or other areas identified in the risk assessment for further evaluation. However, it may be appropriate in limited cases to perform only analytical procedures for a specific area identified for further evaluation based on the risk assessment. In those cases, the auditor should clearly document the rationale for concluding that analytical procedures alone result in sufficient evidence for that particular account or other audit area.

E-100 Section 1 --- Correlation and Use of Scatter Diagrams

E-101 Introduction

This section presents information on the objectives of correlation techniques and the preparation of scatter diagrams.

E-102 Audit Objective

a. The audit objective of using Correlation and Regression Analysis is to evaluate both direct and indirect costs to determine the relationships and interdependencies that may or may not exist between two or more variables. The audit concerns include determining, describing, and estimating the way in which a particular cost or group of costs is influenced by changes in other factors.

b. A principal audit use of these techniques is to identify departures from historical cost patterns. The graphic and computation techniques described later provide a ready means to focus attention on which costs are deviating from historical trends and require special examination. By identifying which cost factors have significantly varied from historical trends, the auditor can focus on ways to further improve internal controls in those areas as well as plan future audits.

c. A very important use of these techniques is to predict future costs. For example, the techniques may be used to evaluate proposed costs in forwarding pricing proposal or forward pricing rate audits. Such applications establish a mathematical relationship between a cost that is to be predicted and one or more additional variables for which predictions are already available. For example, if a contractor's overhead costs are related to direct labor hours, this information can be used to predict overhead rates. Such predictions assume relationships observed in the past will continue in the future. They are most reliable when they are within

the relevant range of the historical data. If a change in condition is known or anticipated, such as operating efficiencies, change in plant location, expanded or compressed vacation periods, change in the number of plant work shifts, the termination or addition of product lines, etc., that will affect the overhead costs being projected, then historical data will likely require adjustment to reflect these changes.

d. Another important use of these techniques is to evaluate how effectively various, alternative direct and indirect allocation bases distributed costs to contract cost objectives in Cost Accounting Standards audits.

E-103 Terminology

a. *Adjusted* R^2. Like R^2, the *Adjusted* R^2 is a descriptive measure of how well the regression fits the values of the dependent variable. The shortcoming of R^2 is that it cannot help but increase as independent variables are added to a regression, spectacularly so as the number of independent variables nears the number of data points. This doesn't happen with the *Adjusted* R^2, which factors in the difference between the number of data points and the number of independent variables, a critical consideration when the data points are few. If an independent variable is added, R^2 will always increase but *Adjusted* R^2 may not. Neither the R^2nor the *Adjusted* R^2 is a statistic that can be used to assess whether the apparent fit is anything more than a chance occurrence, something that is addressed using the computed *F* statistic and its associated level of confidence.

b. Coefficient. A coefficient is a number or quantity generally placed before and multiplying another quantity. For example, the number 3 is a coefficient in the mathematical expression 3x.

c. Coefficient of Determination ($R^{2)}$. The Coefficient of Determination or the R^2 is a number between 0 and 1, and it measures the degree to which changes in the dependent variable can be estimated by changes in the independent variable(s). The R^2 can also be expressed as a percentage. The coefficient of determination is a useful measure of the quality of a regression, but it does have limitations. A primary goal in developing a regression analysis is to predict future values of the dependent variable *y* values, given future values of the independent variable(s). For this purpose high R^2 values are sought based on the expectation that a regression line explaining more of the variation in historical *y* values will yield correspondingly more reliable predictions of future *y* values. Paradoxically, larger data sets having more observation points generally produce more reliable regression line estimates than smaller ones, yet the R^2 value for a given regression model (equation) seldom improves, and almost always declines, as the number of data points is increased.

d. Comparison Assurance. The comparison assurance (the level of confidence associated with the *F* statistic) is a statistical assessment of the entire regression equation that can be used to measure its reliability for making predictions. In terms of ability to predict the dependent variable values, a comparison is made between the equation and an alternative. In EZ-Quant, the alternative is the average value of the dependent variable. The required level of comparison assurance is a matter of audit judgment. As a guideline, comparison assurance should normally exceed 90 percent.

e. Confidence. Confidence or achieved confidence is the complement of significance. Confidence represents the maximum level of confidence at which the regression or individual regression coefficients could be tested and still pass the test.

f. Correlation. Two variables are said to be correlated when there is a measurable tendency for a change in one to be accompanied by a change in the other. By definition, if a variable (e.g., direct labor hours) is considered to cause changes in the other, it is called the independent (or *x*) variable. The other variable (e.g., indirect labor hours) is called the dependent (or *y*) variable. The nature of this relationship may vary from one of a high degree of dependence, where a change in the independent variable directly causes a change in the other, to one of merely coincidental association. It is important to recognize that changes in the dependent variable may or may not have a direct correlation to the independent variable.

Our audit objective is to quantitatively identify variables that have strong associative relationships to each other, which are based on logical rational relationships that can be used to project costs.

g. Correlation Analysis. Correlation analysis evaluates the reliability of the estimated equation from a regression analysis, and it evaluates how well the equation fits the historical data used in the regression.

h. Degrees of Freedom. The degrees of freedom is a statistic representing the number of values that a variable is free to assume. This statistic is importantly used in computing significance and confidence intervals. The degrees of freedom is used in the calculation of the widely used t statistic, where it is decremented by one for each independent variable added to the regression analysis. It is also used in conjunction with the *F* statistic, which has two sets of degrees of freedom associated with it.

i. Dependent Variable. The dependent variable is associated with one or more other independent variables. In a regression equation projection, it is the variable being predicted or estimated by the independent variable.

j. *F* Test. The *F* test establishes whether one can confidently assert the regression analysis demonstrated evidence of at least some correlation between the dependent variable, represented by y, and the independent variable(s). The *F* test protects the auditor from the risk of concluding correlation exists when, in fact, it does not. The *F* test is not a demanding test, and failure of it is good reason to abandon using the specific independent regression variables being tested; and to instead evaluate the use of other independent variables or other alternatives. The *F* test can be performed by comparing EZ-Quant's reported comparison assurance (itself derived from the *F* statistic and its two sets of degrees of freedom) to a confidence level goal established in the audit's planning phase.

k. *F* Statistic. The *F* statistic is the ratio of the variance (not variation) explained by the regression to the variance of that which is not explained. Both variances refer to the dependent variable. Each of these variances has its own degrees of freedom. The *F* statistic is used compute comparison assurance and to perform the *F* test.

l. Inclusion Assurance. Inclusion assurance represents the assurance or confidence that the regression equation is a better predictor with (than without) that particular variable. Each variable's inclusion assurance assumes all other independent variables included in the regression remain there. The required level of inclusion assurance is a matter of auditor judgment. A high inclusion assurance enables the auditor to be reasonably sure the particular variable is worth retaining when other independent variables are present. An inclusion assurance of less than 50 percent is clearly poor. The closer the inclusion assurance is to 100 percent, the better.

m. Independent Variable. The independent variable is a variable whose movement causes some degree of change in a dependent variable. The independent variable is used in the regression equation to predict or estimate the dependent variable.

n. Multicollinearity. Multicollinearity is a condition whereby two or more independent variables included in a regression are materially correlated with each other. Taken together, the affected independent variables can explain much about the dependent variable, but their individual inclusion assurances can be poor, making it unclear which if any should be discarded. (In cases where there is a very high degree of multicollinearity, because the independent variables are especially closely related, the result can be extremely adversely impacted (i.e., they can become very unreliable).

o. Outliers. Outliers represent data points that are prominently out of pattern in a scatter diagram of all data points within a relevant historical range.

p. Regression Analysis. Regression analysis estimates the coefficients for an equation which expresses the presumed functional relationship between dependent and independent variables.

(1) Simple Linear Regression Analysis. A simple linear regression analysis is a statistical procedure that involves the use of one dependent and one independent variable. Simple li-

near regression fits a straight line through a set of data points in such a way that it minimizes the sum of the squared vertical distances between the data points and the fitted line.
(2) Multiple Regression Analysis. Multiple regression analysis is a statistical procedure that attempts to assess the relationship between a dependent variable and two or more independent variables. Multiple regression analysis is used when it is clear from reasoning or observation that the dependent variable is substantially influenced by more than one independent variable.

q. Relevant Range. The relevant range represents the historical range of data, or not excessively far outside of it. It is the range over which contractor operations normally occur.

r. Scatter Diagram. The scatter diagram, or scattergram, is a preliminary graphical analysis technique in EZ-Quant that displays data to help establish whether there is a relationship between dependent and independent variables. A scatter diagram may immediately show there is no closely correlated dependent-independent variable relationship.

s. Significance. In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. The amount of evidence required to accept that an event is unlikely to have occurred by chance is the significance level or the critical *p* factor. In other words, the significance (*p*) is the minimum significance level at which the regression analysis could be tested and it could still be demonstrated that correlation exists.

t. Slope. The slope of the regression line is represented by the *b* coefficient in the $y = a + bx$ linear regression equation. The slope is important because the *bx* term represents the change in the dependent variable *y* based on a change in the independent variable(s).

u. *t*-statistic. The *t*-statistic is a measure of the degree which each independent variable has a statistically discernable relationship with the dependent variable at a specific confidence level, (i.e., that the independent variable is a sufficient "explainer" of the movement of the dependent variable at a given confidence level.

v. Sum of Squares Error (SSE). The SSE is a measure of the variation (not variance) of *y* that is not explained by the regression equation. SSE is the sum of the squared differences between the observed values of y and the calculated value of *y*. The SSE represents the random variation of the observations around the regression line.

w. Sum of Squares Regression (SSR). The SSR is a measure of the variation of y that is explained by the regression equation. SSR is the sum of the squared differences between the calculated value of y and the mean of *y*.

x. Sum of Squares Total (SST). The SST is a measure of the total variation of *y*. The SST = SSR + SSE.

y. Variable. A variable can represent either an individual item or all the items in a class and it can assume different values. For example, a contractor's direct labor-hours, average labor rates, and expense levels are examples of variables that can assume different values at different points in time. A variable can also represent a characteristic and have different values for different members, such as a wage rate or period of service for individual employees.

z. *y*-intercept. The *y*-intercept represents the theoretical value of y when *x* equals 0.

E-104 The Scatter Diagram

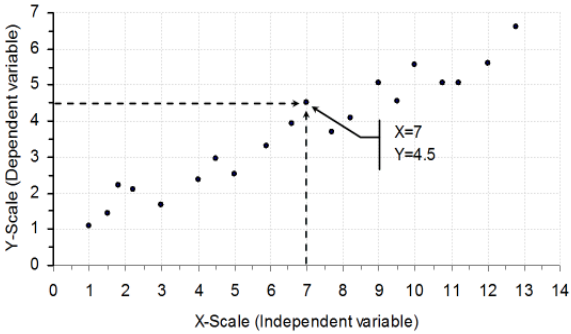
E-104.1 Purpose

The scatter diagram, sometimes called a scattergram, is a preliminary graphical analysis technique included in EZ-Quant that displays data to help establish whether there is a relationship between dependent and independent variables. A scattergram can be quickly prepared by entering data into EZ-Quant and then examining the resulting graph. A scattergram can also be prepared manually or using other software. Examining a scattergram is a helpful and revealing first step, as it may immediately show the lack of a closely correlated dependent-independent variable relationship.

E-104.2 Construction – Scatter Diagram

A scatter diagram, by convention, is plotted using the *x* axis to show the values of the independent variable and the *y* axis to show the values of the dependent variable, as shown in Figure E-1-1.

Figure E-1-1
Close Positive Linear Correlation

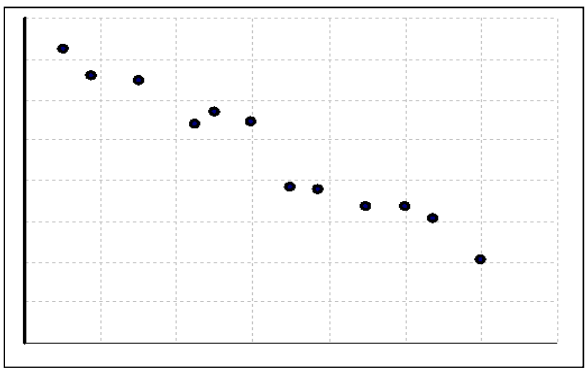


E-104.3 Interpretation

a. Direction, Position, and Linearity of Pattern.

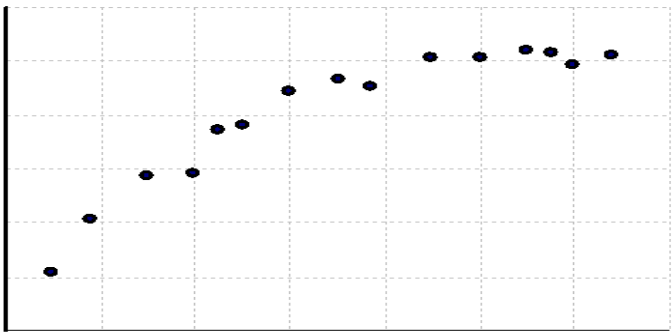
(1) If there is correlation between the two variables being plotted, the points of a scatter diagram form a pattern having a discernable direction in relation to the axes of the graph. As illustrated in Figure E-1-1, when the general path of the pattern is from the lower left to the upper right of the graph then the dependent *y* values increase correspondingly with increases in the independent *x* values. In this situation there is said to be a directly related or positive relationship. When the path is from the upper left to the lower right (Figure E-1-2), the dependent *y* values decrease as the independent *x* values increase, and the correlation is said to be inversely related or a negative relationship.

Figure E-1-2
Close Negative Linear Correlation



(2) The general path of the scatter of points might also indicate whether the correlation is linear or curved (nonlinear). If the central path through the pattern of points from the lowest to the highest *x* value is approximately a straight line as shown in Figures E-1-1 and E-1-2, the correlation is linear and there is said to be a constant or straight-line relationship between the variables. If the central path through the pattern is curved, as illustrated in Figure E-1-3, the correlation is curvilinear and a nonlinear relationship is said to exist between the variables. It is important to distinguish between linear and nonlinear relationships because different computational analysis techniques are required. Substantial errors may result from assuming a straight-line or linear relationship when the true relationship is curvilinear. An important reason for constructing and evaluating a scatter diagram prior to examining the computational analysis is to determine, from the pattern of the points, the type of analysis that should be used. If a curvilinear relationship exists, it should be evident in the scattergram. However, the determination of a functional relationship, linear vs. nonlinear should not be made exclusively from a scatter diagram, especially if only a few plot points are available. Visual evidence in this respect is not exclusively compelling by itself. The auditor should also consider what type of functional relationship between the dependent and independent variables exists and whether it makes sense and is logically supportable.

Figure E-1-3
Close Curvilinear Correlation



b. Degree of correlation.

(1) In unlikely but illustrative cases, all the points might fall on a slanted straight line or smooth curve, as in Figures E-1-4 and E-1-5. In these situations there is perfect correlation, and the value of one variable may be visually inferred with substantial accuracy from the value of the other variable. When this line is straight a unit change in one independent variable is expected to result in a fixed change in the dependent variable, within a specific relevant range. When the plotted points fall in a line that is either horizontal or vertical, as in Figures E-1-6 and E-1-7, there is no correlation between the two sets of data. There is no correlation because only one value for one of the variables (e.g., *x*) is associated with all the values of the other variable (e.g., *y*). Figure E-1-6 might represent a scatter diagram of fixed expenses, such as straight line depreciation, per accounting period, Figure E-1-7 might represent the dependent variable direct labor hours that have been historically required to produce one part number.

Figures E-1-4 through E-1-7
Correlation Examples

Figure E-1-4
Perfect Positive Linear Correlation

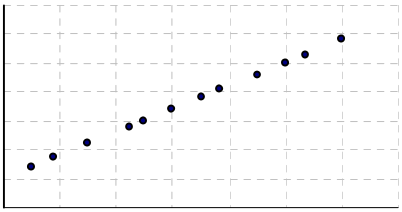


Figure E-1-5
Perfect Curvilinear Correlation

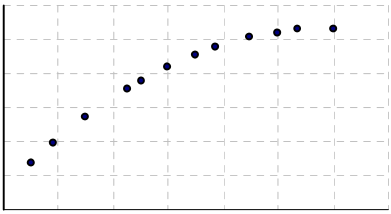


Figure E-1-6
No Correlation - One value of Y associated with all values of X

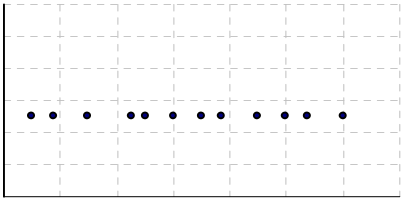
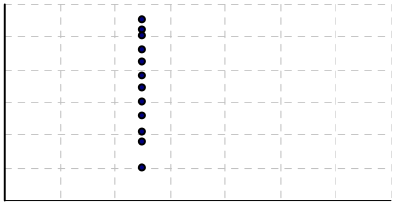
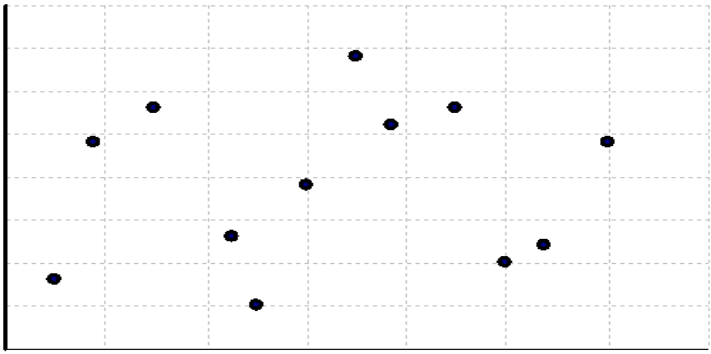


Figure E-1-7
No Correlation - All values of Y associated with one value of X



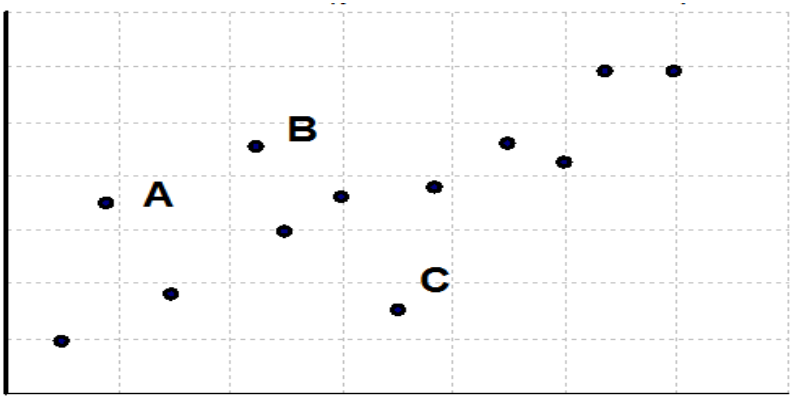
(2) Ordinarily, points do not fall in a line but tend to form a band-like pattern as in Figure E-1-1. The width of this pattern (the scatter of the points about a trend line), together with its general direction from the lower-left to the upper-right, is a visual indication of the degree of correlation, or the closeness of fit, of the relationship between the two series of data. The degree of correlation affects the confidence interval that can be placed on the historical validity of the relationship. A very narrow sloping pattern, closely concentrated around a straight line, indicates a high degree of correlation. The wider the pattern around the trend line or the more the pattern's trend line approaches either a horizontal or vertical orientation, the lower the degree of correlation. If the pattern formed by the points is very widely disbursed about a trend line, as shown in Figure E-1-8, there is little correlation present. This lack of a trend line pattern indicates there is little tendency for the dependent variable to change in relationship to the independent variable.

Figure E-1-8
Little or no Correlation – Lack of association between X and Y



(3) Outliers, data points prominently out of pattern, commonly appear in a scattergram of points, as illustrated in Figure E-1-9. While most of the plotted points are located within an area forming a cohesive generalized pattern, points A, B, and C lie outside this area. The values represented by these out-of-pattern points have the effect of reducing the apparent correlation between the two variables. Aside from correlation concerns, outliers may represent potential audit leads and the possibility that unusual circumstances might have caused the out-of-pattern plot points. The cause of such circumstances should be determined through further examination. If the outliers do not result from either error or unusual circumstances, and they are instead characteristic of the data being analyzed, they should be included in the computations. However, if examination indicates the outlying data points are the result of atypical, non recurring expenditures, then these data points should be properly adjusted before the data.

Figure E-1-9
Reduced Correlation



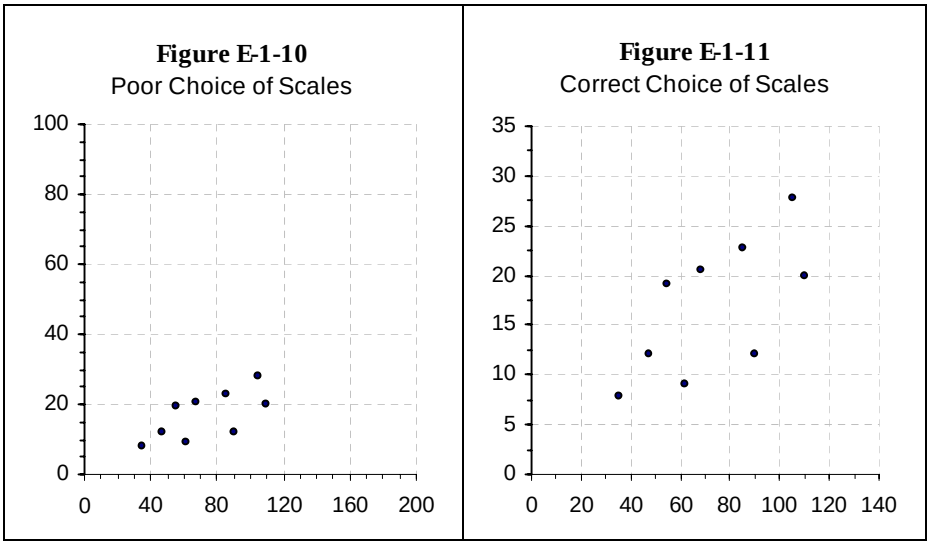
E-104.4 Selection of Appropriate Scales

a. Proper preparation and use of scatter diagrams can help the auditor reach reasonable conclusions about the pattern of plot points, and guide further analysis, of the extent of correlation between the variables. Some considerations are discussed below.

b. Ordinarily, arithmetic scales should normally be used for both the *x* and *y* axes when drawing scatter diagrams. However, in some instances the general pattern of plot points may appear to be curved when arithmetic scales are used. Therefore, a logarithmic scale on one or both axes might yield a straight line pattern, thereby suggesting the form of the correlation between *x* and *y* is curvilinear.

c. Though the scales selected must have a wide enough range to accommodate the lowest and highest values of the data to be plotted, it is generally good practice to configure the scales so the plot points span a substantial portion of the graph area. EZ-Quant will perform this graphing process automatically. Plotting data on most of a graph's available area facilitates visual evaluation of the relationship between the *x* and *y* variables and the degree of variation. Alternatively, compressing data points into a small space may not convey a useful image of the relationship between the variables. An example of a poor choice of scales is illustrated in Figure E-1-10. The restriction of the plot points to a small rectangular area of the graph results in a diagram that can be difficult to interpret. Figure E-1-11, where the same data points have been plotted using a smaller scale for both the *x* and *y* axes, provides a more precise picture of the degree of scatter of the data points.

Figure E-1-10 and E-1-11
Choices of Scales for Scatter Diagrams



E-104.5 Use and Limitations

a. Examining a scatter diagram, without evaluating EZ-Quant's associated computational analysis may be adequate when the auditor's purpose is to identify current departures from historical patterns, as it can provide a quick visual way to assess whether two variables appear to be related. However, EZ-Quant produces the correlation evaluation

statistics by default, and the auditor is encouraged to examine these statistics in order to mathematically assess the degree of correlation.

b. Whether a scattergram is prepared manually or by using EZ-Quant, it generally enables a visual estimate of: (1) the line of best fit and the (2) projected values for the dependent y variable. For example, using the data below Figure E-1-12 is produced showing a simple scatter diagram of the *x* and *y* values.

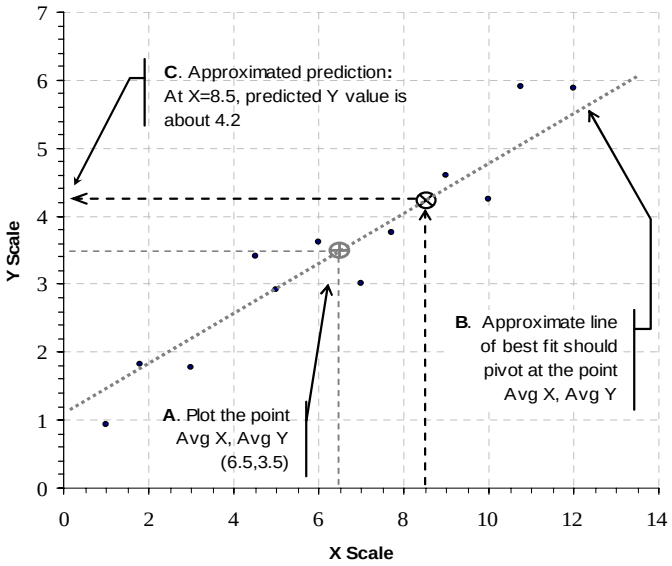
Data Points for Scatter Diagram Example

obs	*x*	*y*
1	1.0	0.94
2	1.8	1.81
3	3.0	1.78
4	4.5	3.40
5	5.0	2.93
6	6.0	3.61
7	7.0	3.00
8	7.7	3.76
9	9.0	4.59
10	10.0	4.25
11	10.8	5.91
12	12.0	5.87
Average	6.5	3.50

The average *x* and *y* values were placed on the scatter diagram as another plot point. A straight line, pivoted on this added point, is then drawn through the scatter pattern in a manner that fits all the plot points as closely as possible. An approximate prediction of the value of *y* for a given value of *x* can then be developed as follows. Using a given *x* value of 8.5 on the horizontal scale, you will reach the approximate line of best fit when you move vertically. From this point move horizontally, right to left, to the y-scale and read the indicated value. The estimated projected value of the dependent y variable is 4.2 in this example.

c. A scatter diagram alone might provide sufficient information to make a preliminary determination of whether a dependent variable displays a relationship to an independent variable, but its usefulness is limited by its lack of quantitative precision. If a scattergram is drawn manually, instead of using EZ-Quant, it is unlikely any two individuals would manually draw exactly the same trend line. Therefore, for most analytical purposes, a line (or curve in some cases) of best fit should be developed using computer software.

Figure E-1-12
Freehand Estimation of the Dependent (Y) Value from a Known Independent (X) Value



.

E-200 Section 2 --- Correlation and Regression Analysis

E-201 Introduction

This section briefly discusses how linear regression can be used to assess the relationship between a dependent variable and one or more independent variables. The mathematical computations of the regression procedure are beyond the scope of this appendix, and those wishing to explore and understand the underlying computations should refer to a standard statistics textbook.

E-202 Simple Linear Regression Analysis

E-202.1 Definition and Concept

a. A simple linear regression analysis involves the use of one dependent and one independent variable. As previously discussed, it is possible to fit a straight line, and in some cases a curvilinear line, to a set of data. The underlying rationale for the variables selected should be both logical and explainable. The calculated line of best fit should closely match the actual graphed data points. Fortunately, many direct and indirect cost relationships of interest to the contract auditor tend to follow a straight line within the relevant range. The relevant range represents the range of the historical data, or not excessively far outside of the historical data over which contractor operations normally occur. The form of the relationship (linear of curvilinear) should make sense, and it should be conceptually supportable. This section is limited to fitting a straight line to data.

b. If two variables x and y are associated in a linear or straight-line manner, the equation expressing the relationship is commonly presented as:

$$y = a + bx$$

where a and b are coefficients estimated by regression computations. The coefficient "a" represents the y-intercept, or the value of y when x is zero. The coefficient b represents the slope of the line, or the change in the dependent variable y associated with a unit change in the independent variable x. When b is positive, the line slopes upward from lower-left to upward-right. When b is negative, the line slopes downward from upper-left to downward-right.

c. For approximation purposes a straight line can be visually fitted to a scattergram of data, as shown in Figure E-1-12. However, beyond an exploratory role, the line of best fit should be mathematically computed using a least squares regression estimation of the coefficients.

d. Least-squares estimation of the coefficients computes a line of best fit (in the case of a single x variable) or slopes of best fit for each x variable (in the case of multiple x variables). The best fit is achieved using a mathematical process which minimizes the squared vertical differences between the actual values of the dependent variable and the fitted line. Figures E-2-1 and E-2-2 show simple linear regression lines fitted to data points. Figure E-2-2 is helpful as a visual reference for the following characteristics of such a line:

(1) The sum of the squares of the vertical distances between the regression line and the points is less than for any other straight line. The sum of the squares is minimized, and this process is sometimes termed, "least squares" analysis.

(2) The sum of the vertical distances above the line equals the sum of the distances below the line.

(3) The line passes through the point corresponding to the mean of the observed values of the x variable, and the mean of the observed values of the y variable.

e. In addition to the values of a and b, regression computations produce various measures describing how well the line of best fit explains the values of the dependent variable.

The most commonly recognized measure is the Coefficient of Determination, or the R^2. The R^2 is a number between 0 and 1, and it measures the degree to which changes in the dependent variable can be estimated by changes in the independent variable.

E-202.2 Example of Regression

a. The following table shows quarterly totals for an overhead account and an allocation base.

Example Data: Overhead and Base

Year	Qtr	Mfg Overhead ($000)	Direct Labor ($000)
1	1	839	1,203
	2	845	1,304
	3	713	872
	4	860	1,104
2	1	873	1,176
	2	912	1,258
	3	736	983
	4	886	1,283
3	1	962	1,576
	2	878	1,372
	3	820	956
	4	780	957
Average		842	1,170

In a regression analysis of manufacturing overhead expenses (the dependent variable) is related to direct labor expenditures (the independent variable). The relationship can be expressed in the following mathematical expression.

$y = a + bx$ or

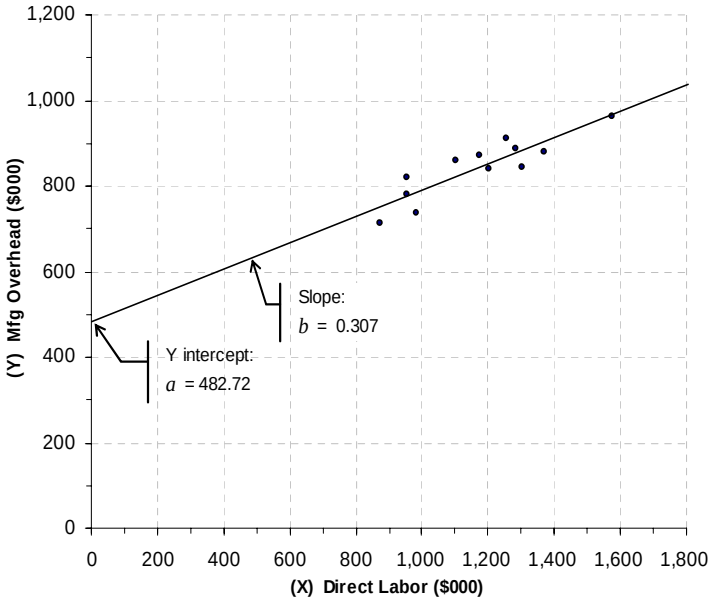
Manufacturing Overhead = a + b(*Direct Labor*)

The estimated coefficient values are as follows:

Coefficient		Value
a	Y Intercept	482.72451
b	Slope	0.3069856

The line of best fit is shown graphically in Figure E-2-1.

Figure E-2-1
Graph of Simple Linear Regression Line



E-203 Curvilinear Regression Analysis

a. The same least-squares estimation process used to compute a straight line of best fit can also be used to fit curvilinear equations to the data. Curvilinear equations have nonlinear form when plotted on an arithmetic scale, but they can be converted to linear form by transformation of the values of the dependent variable *y*, the independent variable *x*, or both to non-linear or non-arithmetic scales. There are a number of curvilinear equations, and a great number of non-linear equations, but only a few of them have been found to have application in contract cost audits.

b. Two curvilinear equations that have been useful in contract auditing are the power and exponential curves. The power curve ($y = a\ x^b$) is the form used for improvement curves. The exponential curve ($y = a\ bx$) can be expressed as $y = a(1+r)x$, where $b=1+r$.

c. The auditor should use the equation which has been judgmentally determined to have a logical relationship between the variables. These equations are linear in their logarithmic forms when they are graphed on log or semi-log paper, but are curves when graphed on standard arithmetic graph paper as a scattergram. The process of mathematically fitting this line to data is referred to as curvilinear regression analysis.

d. Though few curve types are typically used in contract audits, in special situations there are many types of nonlinear equations that might be used. Some types cannot be transformed to linear form, and nonlinear least squares curve estimation processes are needed for curve estimation. The auditor should seek technical assistance in dealing with such applications.

e. The auditor should be able to support the use of a nonlinear equation for estimating the relationship between variables. The auditor should be able to explain why the specific type of estimated relationship is a reasonable representation of the actual interaction between the dependent and independent variables. This logical support is particularly important when the range of independent variable values is narrow, in which case most curves

will themselves appear to be nearly linear within that range and more clearly nonlinear only outside the range. Caution is advised concerning predictions from the curve or line using independent variable values significantly outside the data's relevant range. As values move increasingly further outside the relevant range, less reliable predictions are generated.

f. Assistance with curvilinear and nonlinear regression applications is available from the Technical Audit Services Division.

E-204 Multiple Regression Analysis

a. Multiple regression analysis is a statistical procedure that attempts to assess the relationship between a dependent variable and two or more independent variables. Multiple regression analysis is used when it is clear from reasoning or observation that the dependent variable is substantially influenced by more than one independent variable. Multiple linear regression analysis should be used cautiously; increasing the number of independent variables will always increase the R^2 and thus can appear to improve a regression when in fact there may be no actual improvement. Adding an independent variable that is poorly correlated with the dependent variable cannot be expected to improve the regression's usefulness for predicting dependent variable values, though R^2 would increase to some degree. The same is true for adding an independent variable that is highly correlated with one or more independent variables already included in the regression.

b. Multiple regression analysis uses the least-squares method to simultaneously estimate the individual effect, or coefficient, of each independent variable on the dependent variable. Under virtually no circumstances will the coefficient of a specific independent variable be the same in a multiple regression as it would have been if it were the only independent variable in a simple linear regression analysis. Multiple regression analysis uses the following general equation:

$$y = a + b_{1x1} + b_{2x2} + \cdots + b_{J\,xJ}$$

where y is the dependent variable, x1 is the first independent variable, x_2 is the second, and so on to the Jth independent variable. The letters a, b_1, b_2 . . b_J. refer to coefficients whose values, as a group, minimize the sum of the squares of differences between (1) the actual data values of y and (2) the corresponding values calculated from the equation using the independent variable data values. As with simple linear regression, multiple regression yields the coefficient of determination (R^2) that measures how much of the variation in the dependent variable has been explained by the independent variables. Statistical information for the each of the estimated coefficients (a, $b1$, $b2$ . . bJ) is also reported. Most important is the confidence level (or 100-significance level), which shows that the coefficient passes a t-test of significance up to that level of confidence. The t-test of significance, which is based on the reported t value and its degrees of freedom (number of data points less the number of regression coefficients). This test determines whether the estimated coefficient b_i is sufficiently statistically significant that it likely represents an actual relationship as opposed to one that might just appear to be significant, without a quantitative measurement. The confidence level is thus a statistical assurance at a stated level that a relationship actually exists. Whether this achieved confidence is sufficient is a matter of audit judgment. Additionally, EZ-Quant calculates the F statistic and its maximum confidence level in a significance test of the entire regression. In multiple regression, only in rare circumstances will the estimated effect of a specific independent variable (e.g., $y = a + b_{1x1} + b_{2x2}$) be the same as if it were the only independent variable in a simple regression. The relationship between the dependent variable and each of the independent variables cannot be shown on a single graph, except in the case of two independent variables being plotted with three dimensional graphics.

c. Only independent variables that are in use and are actively changing in relationship to varying dependent variables should be considered for inclusion in a regression analysis. For example, manufacturing overhead is driven not only by a direct manufacturing labor

hours component of the production process, but also can be driven by other non-labor components of the manufacturing process such as the number of active robotic milling machine hours. Also, consider for example, a relatively fixed or semi-variable independent variable such as floor space in a one-product manufacturing environment where the floor space variable has not significantly changed over many years. Adding the independent variable floor space in that environment would not likely improve the accuracy of the regression equation because it is a relatively unchanging static variable. However, if the amount of production floor space was both available data-wise, and it changed significantly in relation to different products actually being built then it may be helpful in explaining changes in the dependent variable. Alternatively, the inclusion of production floor space as an independent variable might be appropriate in some circumstances because these numbers are constantly changing in order to support different size job lots.

d. Testing using various additional independent variables is also useful in an exploratory role when the regression of *y* on a single independent variable is found to have merit but where a substantial share of the movement in the *y* values still remains unexplained. A possible remedy is to include one or more additional independent variables ("explainers") in the regression analysis. However, any additional independent variables must have a logical relationship in explaining changes in the dependent variable. To be of any explanatory benefit, a variable being considered for inclusion cannot be highly correlated with one or more independent variables already in place. For example, adding labor hours and then the direct labor dollars, which are based on labor hours, is not appropriate because these independent variables are essentially the same.

e. In the case of multiple linear regression analysis there is often some correlation among two or more of the independent variables, a condition known as multicollinearity. In varying degrees it is present among most variables used in multiple linear regression applications, including those in contract auditing. For example, a regression might logically include both direct labor and material costs as independent variables. In such a case there would be a minimal expectation that some degree of correlation between them might exist. In cases of minimal multicollinearity the regression's line of best fit to the dependent variable is not significantly affected. However, multicollinearity can significantly and adversely affect the regression's line of best fit because its primary effect is to increase the variability of the individual coefficients of potentially every independent variable. In fact, multicollinearity can cause coefficients to vary radically, unexpectedly, and produce seemingly valid but ineffective results. This occurs because when multicollinearity is substantial it becomes increasingly difficult to reliably estimate the value of the coefficients of the independent variables. The results produce larger standard errors of the coefficients (a measure of their variability) and correspondingly lower levels of confidence. This greater variability means that the coefficient values are more likely to be materially overstated or understated. The coefficient values can also change substantially when another independent variable is added or when the number of observations is slightly changed. The coefficient variability induced by multicollinearity can cause difficulty in separating the net effect of individual independent variables upon the dependent variable. Such a regression analysis might have a high level of confidence (per its *F* statistic) as well as a high R^2 value, but its individual coefficients might have low confidence levels (per their *t* statistics), which may be easily overlooked. Multicollinearity can best be minimized by careful choice of the set of independent variables included in the regression. Considering a variable about to be included in a regression, the greater its correlation with one or more independent variables already included, the less it can add to what is already explained about the values of the dependent variable. The auditor should avoid independent variables which merely redefine the same costs or are overlapping subsets of the same costs that are already used as independent variables. For example, consider a regression with material costs as an independent variable. Total material costs typically include: raw materials, purchased parts, standard hardware, and subcontracts/outside production. It is inappropriate to include total material costs as well as one or more of the above four

subcategories of material costs as independent variables. For example, using both total material costs and the subcategory of raw material costs as independent variables in a regression analysis is inappropriate because the same dollars are included in both independent variables. Likewise, using total material costs and all four subcategories of material costs is also an overlapping definition of the same material costs.

E-205 Correlation Analysis

A distinction is made here between regression analysis and correlation analysis. Regression analysis is concerned with estimating the coefficients for an equation which expresses the presumed functional relationship between dependent and independent variables. Correlation analysis is concerned with evaluating the reliability of the estimated equation and how well it fits the historical data used in the regression. The following discussion demonstrates how regression results can be used in correlation analysis.

Example Data: Overhead and Two Independent Variables

Year	Qtr	Mfg Overhead ($000)	Direct Labor ($000)	Robotic Milling Machine Hours
1	1	839	1,203	870
	2	845	1,304	880
	3	713	872	875
	4	860	1,104	1,250
2	1	873	1,176	1,155
	2	912	1,258	1,275
	3	736	983	1,075
	4	886	1,283	1,005
3	1	962	1,576	1,278
	2	878	1,372	1,367
	3	820	956	1,340
	4	780	957	1,255
Average		842	1,170	1,135

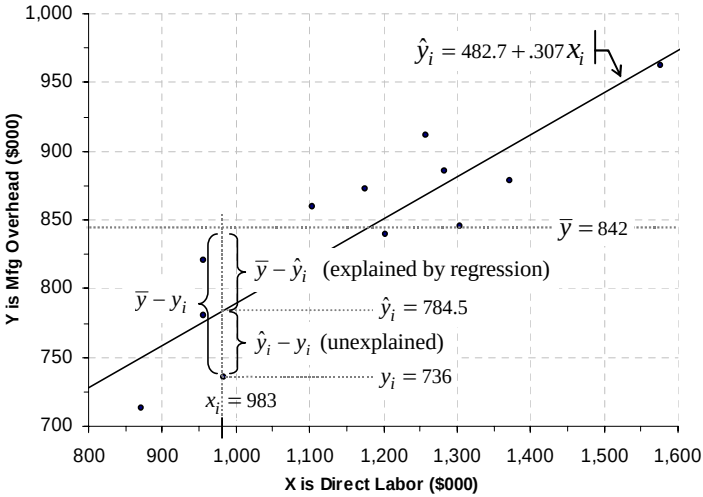
E-205.1 Interpretation of Coefficient of Determination

a. The most widely used measure to evaluate how well an equation fits the data is the Coefficient of Determination, which is given the variable name R^2. This coefficient shows, in either a percent or decimal format, how much of the variability in the dependent variable from its mean (i.e., average) has been attributed by the regression analysis to the independent variable(s). Stated less formally, it shows how much of the change in the dependent variable from one data point to another can be "explained" by changes in the independent variable(s). As a percent, it ranges in value from 0 to 100. Thus, R^2 is a descriptive measure of how closely the dependent variable changes in relationship to the independent variable(s).

b. Computation of the Coefficient of Determination is shown in Figure E-2-2. The difference ($\bar{y} - y_i$) between each value of the dependent variable and the average of all y values is divided into two components: (1) the change in the dependent variable which is explained by the independent regression variables (SSR or Sum of the Squares Regression) and (2) the change which is not explained (SSE or Sum of the Squares Error). Figure

E-2-2 graphically displays the individually plotted actual data points (y_i). The explained portion of the change in the dependent variable represents the difference between the mean and the calculated value on the line of best fit or $\overline{y} - \hat{y}_i$, where $\hat{y}_i$ is the regression-predicted value. The remaining unexplained portion of the change is $\hat{y}_i - y_i$ and it represents the difference between the calculated value on the line of best fit and the actual value.

Figure E-2-2
Regression and Partitioning of Deviations from the Mean, Explained and Unexplained Sum of Squares



As shown below, the (1) total (unpartitioned) deviations $\overline{y} - y_i$ are squared and summed to form the total sum of squares *SST*, (2) explained deviations $\overline{y} - \hat{y}_i$ are squared and summed to form the regression sum of squares *SSR*, and (3) unexplained deviations $\hat{y}_i - y_i$ are squared and summed to form the error sum of squares SSE. The regression procedure itself determines the coefficient values (intercept and slope) that minimize the value of SSE, maximize the value of *SSR*, and thereby maximize the value of R^2 = *SSR/SST*.

	Partitioned Sum of Squares, Mfg Overhead Regressed on Direct Labor		
	SSR (Regression, explained)	$\sum(\overline{y} - \hat{y}_i)^2$ = 43,777	$R^2 = SSR/SST$ $= 1 - SSE/SST$ $= .7812$
	SSE (Error, unexplained)	$\sum(\hat{y}_i - y_i)^2$ = 12,263	
	SST (Sum of the Squares Total)	$\sum(\overline{y} - y_i)^2$ = 56,040	

$\hat{y}$	Regression - predicted value
$\overline{y}$	Average of all y values
yi	Individual data points

c. If all of the observed values of *y* were to fall on the regression line, R^2 would equal 1.0 (decimal format), representing perfect correlation. However, perfect correlation will essen-

tially never occur, except when the number of observations equals the number of coefficients in the regression equation. For example, if there are only two data point observations, a straight line is produced in the case of simple (one independent variable) linear regression. In this two-point case the R^2 value of 1.0 is meaningless as an indicator of correlation because even totally unrelated variables will achieve perfect correlation. On the other hand, R^2 will equal zero if there is no better fit to the historical data than the average of the *y* values. This result, representing a horizontal line at the average value of *y* for every value of *x*, is also very unlikely. Be aware that the values of even unrelated variables will typically display at least some minor measure of correlation. Thus, additional independent variables should not be included in a regression analysis unless each one contributes to a significantly better estimation of the dependent variable.

d. The coefficient of determination is a useful measure of the quality of a regression, but it does have limitations. A primary goal in developing a regression analysis is to predict future values of the dependent variable *y* values, given future values of the independent variable(s). For this purpose high R^2 values are sought based on the expectation that a regression line explaining more of the variation in historical *y* values will yield correspondingly more reliable predictions of future *y* values. Paradoxically, larger data sets having more observation points generally produce more reliable regression line estimates than smaller ones, yet the R^2 value for a given regression model (equation) seldom improves, and almost always declines, as the number of data points is increased.

e. The principal problem encountered in the use of R^2 stems from the fact that if the number of observations is small, regression analysis can produce a high coefficient of determination even if the variables are unrelated. That is, with a small data set, a high R^2 value and the appearance of fit can occur by chance alone and thus cannot by itself be interpreted as conclusive evidence of a significant statistical correlation. On the other hand, the larger the number of observations used in a regression, the more remote is the possibility of a high coefficient of determination occurring only by chance. Whether the number of data points is small or large, the regression should be tested to determine if there is statistical support for an assertion that the variables are actually correlated. The coefficient of determination can then be regarded as a measure of the extent that changes in the *y* values can be explained by changes in the *x* values.

E-205.2 Determining the Existence of Correlation

a. Regression estimation software reports several statistics used to test regression results. Among these is the *F* statistic. The *F* statistic is used to perform the *F* test, which establishes whether one can confidently assert the regression analysis demonstrated evidence of at least some correlation between dependent variable y and the independent variable(s). With respect to an *F* test of significance, the table below presents certain information about the regression shown in Figure E-2-2 above.

R^2	F Statistic	Degrees of freedom (df_1 ,df_2)	Significance (p)	Confidence (1-p)•100 (Comparison Assurance)
0.781	35.699	1, 10	0.00014	99.986

The *F* statistic is computed from some of the same information as is R^2, but the computation also factors in the degrees of freedom. The degrees of freedom represent the number of values in the final calculation of a statistic that are free to vary. Using terms given in E-205.1(b) above, $F = (SSR/df_1)/(SSE/df_2)$. The first degree of freedom (df_1) represents the number of regression coefficients (i.e., a and b or two minus one). The second degree of freedom figure (df_2) represents the number of data points (12 in this case) minus the number of coefficients (2). The significance *p*, which depends on both the reported *F* value and

its associated degrees of freedom, is the minimum significance level (smaller is better) at which the regression could be tested and still pass as evidencing a degree of correlation. Confidence or achieved confidence is the complement of significance. "Confidence" represents the maximum level of confidence at which the regression could be tested and still pass the test. In this example, the regression passes the *F* test at the 95% confidence level but not at the 99.999% confidence level. EZ-Quant reports this confidence as the comparison assurance, reflecting the implication that the regression is being compared to the mean historical value of y as an alternative predictor of future values of *y*. However, failure of the *F* test does not imply that the mean *y* value is a good alternative. Its variance, $SST/(df_1 + df_2)$, could be so high as to prohibit its use as a predictor.

b. The *F* test protects the auditor from the risk of concluding correlation exists when, in fact, it does not. It is not a particularly demanding test, and failure of it is good reason to abandon using the specific independent regression variables being tested and to evaluate the use of other independent variables or seek some other alternative. Essentially, the *F* test asks whether any of the estimated slope coefficients (*b* in simple regression and bi in multiple regression, one for each independent variable included in the regression) is significantly different from zero. Even if correlation is relatively weak a regression analysis model will often pass the *F* test; therefore, it is reasonable to require the level of confidence demanded be high. Further, a high achieved level of confidence evidences only the existence of correlation, not its extent, and not how fully it explains the dependent variable.

c. The simplest way to implement the *F* test is to set a confidence level goal and compare it to the level achieved by the regression analysis (i.e., EZ-Quant's reported comparison assurance). Failure occurs when the achieved level fails to match or exceed the specified goal.

d. The *F* test can be performed manually by comparing the reported R^2 value to Table E-2-1, which includes *F* values at the 95% and 90% confidence levels. A regression analysis model passes the *F* test at the one of the selected level of confidence when the reported R^2 equals or exceeds the value presented in the table, for the respective number of data points and number of regression coefficients used in the analysis. For example, a simple linear regression (with two coefficients, *a* and *b*) estimated from five data points with a reported achieved R^2 value of .753 passes the *F* test at the 90% confidence level (critical value .649) but fails the F test at the higher 95% confidence level (critical value .771). As a different example, in the case of a multiple regression analysis that has three independent variables, $y = a + b_1x_1 + b_2x_2 + b_3x_3$, such a model has four coefficients (a, b_1, b_2, and b_3). If there were 16 observations, the appropriate R^2 value from Table E-2-1 for testing the significance of the regression model at the 95 percent confidence level would be .466. If the R^2 value obtained exceeded .466, the auditor would conclude that there is a 95 percent probability, or more, that the three independent variables (x_1, x_2, and x_3), as a group, have a significant influence on the dependent variable y.

e. In addition to the *F* statistic being used to evaluate the overall significance of the regression analysis, the EZ-Quant software also evaluates each of the individual coefficients and reports the inclusion assurance, or confidence, (= 100 - level of significance) derived from the associated reported t statistic for each. A low inclusion assurance (or high significance level) indicates that there is not a statistically discernable relationship between the independent variable and the dependent variable. When the inclusion assurance is relatively low, the auditor should consider removing that variable from the regression. Also, a low inclusion assurance may indicate multicollinearity, which is present when two or more independent variables are highly correlated with each other.

f. For the regression shown in Figure E-2-2 above, the following table summarizes the information reported for the two estimated coefficients (*a* and *b*).

Coefficients		Value	t	df	Significance (p)	Confidence (1-p)•100 (Comparison Assurance)
a	Intercept	482.72451	7.917	10	0.00001	99.999
b	Direct Labor	0.3069856	5.975	10	0.00014	99.986

Through its role in determining the significance (and thus the inclusion assurance), the *t* statistic is used to test whether each coefficient differs significantly from zero. In the case of a simple regression (one independent variable), the *t* test of the coefficient for the single independent variable is equivalent to the *F* test for the entire regression.

g. The unique role of the t statistic (and its associated levels of significance *p* and confidence or inclusion assurance (1-*p*)) occurs in multiple regression analysis, which is demonstrated here with the addition of another independent variable. The additional independent variable, robotic milling machines hours, was previously presented in the data table in E-205, above. With the addition of this second independent variable the *F* test summary below shows the R^2 value has increased from .781 to .851. Beware, the R^2 value will almost always show some improvement as explanatory variables are added to a regression, though the improvement may not be large enough to be statistically meaningful. Do not simply assume that just because the R^2 value has numerically increased with the addition of another independent variable that the resulting multiple regression analysis is better than a simple linear regression analysis having a seemingly lower R^2 value.

No. of Independent Variables	R2	F Statistic	Degrees of freedom (df_1 ,df_2)	Significance (α)	Confidence (1-α)•100 (Comparison Assurance)
1	0.781	35.699	1, 10	0.00014	99.986
2	0.851	25.672	2, 9	0.00019	99.981

The following table summarizes information, including the t statistic, reported for the three estimated coefficients (*a*, b_1 and b_2) of the expanded regression.

Coefficients		Value	t statistic	df	p	Confidence (1-p)•100 (Inclusion Assurance)
a	Intercept	385.39322	5.413	9	0.00043	99.957
b_1	Direct Labor	0.2913199	6.422	9	0.00012	99.988
b_2	Robotic Milling Machine Hours	0.1018705	2.051	9	0.07055	92.945

Here, the *t* statistic is used to test whether each coefficient differs significantly from zero. Rather than being a test of the magnitude of the coefficient value, the *t* test for an individual coefficient is a test of the possibility that: (1) there is no relationship between the dependent variable and the specific independent variable, given that the other independent variables (if any) are present in the regression analysis, and (2) only by chance may the R^2 value have improved because of the addition of this coefficient. In other words, it tests whether the associated independent variable contributes any statistically detectable unique "explanation" of the dependent variable's movement given the presence other independent

variables in the regression. The significance (sometimes reported as the *p* value), is the minimum significance level (smaller is better) at which the coefficient could pass a *t* test for being non-zero. Confidence (or achieved confidence), the complement of significance, computed as shown above if not reported, is the maximum level of confidence at which the coefficient could be tested and pass the significance test. EZ-Quant reports this confidence as inclusion assurance because the *t* test is used to determine if the associated independent variable warrants inclusion in the regression analysis. In the table above, the achieved confidence level of 92.945 percent (*p* value of .07055) indicates that at a confidence level of 90 percent, robotic milling machine hours warrants inclusion in the regression, but fails to do so at a confidence level of 95 percent. Despite the fact that adding the robotic milling machine hours variable raised the R^2 value from .781 to .851, the case for its inclusion is not overwhelming at these benchmark confidence levels.

h. It should be clearly understood that simply because a regression analysis passes the *F* test, it does not follow that the regression equation will provide the best possible estimates of future *y* values, or even that the estimates will be reasonably close. It merely means a statistical analysis of the historical data identifies the probability the equation will provide better estimates than could be obtained by merely averaging the historical values of *y*.

i. The auditor should also bear in mind the test considers only the mathematically expressed (or modeled) relationship between the dependent and independent variable(s). It does not take into account the audit significance of the relationship. The logic of the relationship is critically important. If the regression equation expresses a relationship which can be strongly defended by logical arguments, and only a few observations are available, some significance can be attached to the analysis even if the test does not provide conclusive statistical evidence the values are correlated. However, a special effort should be made in such cases to determine why the correlation is not close and whether the factors causing the observed variations or any other unusual conditions are likely to be present and materially affect the costs during the forecast period.

Table E-2-1
Table of Significant Values of the Coefficient of Determination (R^2)

Number of Data Points (n)	Confidence Level: 95%					Confidence Level: 90%				
	Number of Coefficients (k)					Number of Coefficients (k)				
	2	3	4	5	6	2	3	4	5	6
3	.994					.976				
4	.902	.997				.810	.990			
5	.771	.950	.998			.649	.900	.994		
6	.658	.864	.966	.999		.532	.785	.932	.996	
7	.569	.776	.903	.975	.999	.448	.684	.844	.949	.997
8	.499	.698	.832	.924	.980	.386	.602	.759	.877	.959
9	.444	.632	.764	.865	.938	.339	.536	.685	.804	.898
10	.399	.575	.704	.806	.887	.302	.482	.622	.738	.835
11	.362	.527	.651	.751	.835	.272	.438	.568	.680	.775
12	.332	.486	.604	.702	.785	.247	.401	.523	.628	.721
13	.306	.451	.563	.657	.739	.227	.369	.484	.584	.673
14	.283	.420	.527	.618	.697	.209	.342	.450	.545	.630
15	.264	.393	.495	.582	.659	.194	.319	.420	.510	.592
16	.247	.369	.466	.550	.624	.181	.298	.394	.480	.558
17	.232	.348	.440	.521	.593	.170	.280	.371	.453	.527
18	.219	.329	.417	.494	.564	.160	.264	.351	.428	.499
19	.208	.312	.397	.471	.538	.151	.250	.332	.406	.474
20	.197	.297	.378	.449	.514	.143	.237	.316	.386	.452
21	.187	.283	.361	.429	.492	.136	.226	.301	.368	.431
22	.179	.270	.345	.411	.471	.129	.215	.287	.352	.412
23	.171	.259	.331	.394	.452	.124	.206	.275	.337	.395
24	.164	.248	.317	.379	.435	.118	.197	.263	.323	.379
25	.157	.238	.305	.364	.419	.113	.189	.253	.310	.364
26	.151	.229	.294	.351	.404	.109	.181	.243	.298	.350
27	.145	.221	.283	.339	.390	.105	.175	.234	.287	.338
28	.140	.213	.273	.327	.377	.101	.168	.225	.277	.326
29	.135	.206	.264	.316	.365	.097	.162	.218	.268	.315
30	.130	.199	.256	.306	.353	.094	.157	.210	.259	.305
31	.126	.193	.248	.297	.342	.091	.152	.203	.251	.295
32	.122	.187	.240	.288	.332	.088	.147	.197	.243	.286

The number of coefficients k is the number of independent variables plus one for the intercept.
Critical R^2 values were computed from critical F values as $R^2 = F\ (k\text{-}1)\ /\ (F\cdot(k\text{-}1)+n\text{-}k)$, where n is the number of observations and k is the number of coefficients.

E-205.3 Evaluating the Extent of Correlation

a. The statistical testing, F and t, is used to determine whether there is sufficient evidence to conclude correlation between a dependent variable and one or more independent variables actually exists. The tests are convenient because: (1) they can be readily implemented using statistical measures reported by EZ-Quant, and (2) they consider the degrees of freedom (df_1, df_2). These tests are well-suited to identifying which independent variables should be discarded, but beyond that they do not determine whether a regression explains enough about the dependent variable to suit audit purposes.

b. The coefficient of determination R^2 is well recognized as descriptor of how much of the total variation in the dependent variable has been explained by the regression. Equally

well recognized is that the closer R^2 is to 1.0, the better the regression equation fits the actual values of the dependent variable. A high R^2 says that much of the total variation observed in the dependent variable has been explained by the regression, and relatively little remains unexplained; however, there is no threshold value that says an achieved R^2 is high enough. The R^2 values in Table E-2-1 are not threshold values in this respect, since they show what the regression must achieve to reliably say the regression explains anything at all, not that it has explained enough.

E-205.4 Adjusted R^2

a. The coefficient of determination or R^2 represents the proportion of variation in the dependent variable *(y)* that can be explained by the independent variables *(x_1, x_2, etc.)* in a regression model. The coefficient of determination or R^2 can also be used to determine which of two different equations, each having the same dependent variable but a different independent variable, provides the better fit to historical data. The two computed R^2 amounts can be compared to each other to determine which independent variable provides a better predictor of the dependent variable. A direct comparison between the values of R^2 obtained in two different regression analyses is valid only if both analyses use (1) the same number of observations, and (2) the same number of independent variables. Otherwise, comparisons should be made using the *Adjusted* R^2, since it is not limited to these specific circumstances.

b. Like R^2, the *Adjusted* R^2 is a descriptive measure of how well the regression fits the values of the dependent variable. Neither of them is a statistic that can be used to assess whether the apparent fit is anything more than a chance occurrence, something that is addressed using the computed F statistic and its associated level of confidence. The shortcoming of R^2 is that it cannot help but increase as independent variables are added to a regression, spectacularly so as the number of independent variables nears the number of data points. This doesn't happen with the *Adjusted* R^2, which factors in the difference between the number of data points and the number of independent variables, a critical consideration when the data points are few. If an independent variable is added, R^2 will always increase but *Adjusted* R^2 may not. In any case, to judge whether improvement in R^2 or *Adjusted* R^2 is due to more than chance, one would look at the confidence (inclusion assurance in EZ-Quant) associated with the t statistic for the added independent variable. The *Adjusted* R^2 can be computed from R^2 using the formula $1-(1-R^2)*(N-1)/(N-k-1)$. When the number of observations (N) is small and the number of predictors (k) is large, there will be a much greater difference between the R^2 and *Adjusted* R^2. Conversely, when the number of observations is very large compared to the number of independent variables, the value of the R^2 and *Adjusted* R^2 will be much closer.

c. *Adjusted* R^2 values provide an objective means of choosing between two equally logical equations, whether or not they use different numbers of independent variables or different numbers of observations. The greater the number of observations used and the more substantial the differences are between the *Adjusted* R^2 values, the more reliable the comparison.

E-206 Adjustment for Economic Factors in Application of Regression to Wages

Many contract audit applications of regression analysis include variables which are affected by changes in wage and price levels. When economy-wide changes in price levels have significantly affected any of the variables during the period covered by the historical data, it is typically necessary to adjust the data to eliminate the effects of the economic changes before performing a regression analysis. For example, to measure the effect of direct labor volume on labor costs, changes in labor rates resulting from economic factors (such as cost-of-living allowances) would first be factored into the analysis by adjusting

the historical labor rates for inflation. The following simple (two-variable) least squares regression equation is applicable:

$$\grave{y}_i = a + bx_i$$

where $\grave{y}_i$ is the inflation adjusted average labor rate for time period i; x_i is the number of employees for time period *i*; and *a* and *b* are coefficients to be estimated by the regression. The individual adjusted values $\grave{y}i$ are computed as y_i/I_i. I_i, the index for time period *i*, represents the amount (dollars) in time period *i* that equals the purchasing power of a base period dollar.

(1) The most precise economic labor rate (cost-of-living) indexes are formed using data on wage-related economic changes specific to a particular contractor, considering union agreements and other labor records. However, if such agreements or records do not exist or it is too time consuming to gather the data and build the index, other economic indexes can be used to make the adjustments. Economic indexes that can be used for this purpose are discussed in DCAAP 7641.74, Use of Economic Indexes in Contract Audits. Salary and wage changes generally include both inflation (cost-of-living) as well as the effects of other general economic changes and can be obtained by using DCAA's Intranet. If labor rates are to be adjusted to account for inflation alone, the Consumer Price Index, Urban Wage Earners and Clerical Workers (CPIW) should be used. However, if labor rates are to be adjusted for all wage-related economic changes (including inflation), wage indexes should be used. DCAAP 7641.74 provides a discussion of the types of indexes available to make such adjustments. However derived or obtained, for the regression analysis purposes described above, the value of the index for period *i* (I_i) must be relative to the index value for the base period. That is, the index value for a period must be the ratio of that period's index value to the base period's index value.

(2) When adjusting an individual cost element or a homogenous grouping of similar costs, an economic index (e.g., commodity, industry, or category) that is the most appropriate disaggregate index available should be used. When adjusting a group of costs, either a composite index should be developed for the group, or the most representative disaggregate index of the group of costs should be used. The method for placing costs on a common base is discussed in DCAAP 7641.74.

(3) When the regression equation is used to predict future costs, the anticipated effects of economic changes between the base period and the prediction period(s) must be considered. Projections using the regression equation will be in the same base period dollars as the adjusted data values, so the projections will need to be adjusted for the anticipated inflation that will occur. Predictions of future wage and price levels, particularly long-range, can best be made by qualified economists. DCAAP 7641.74, provides information on the forecasting services purchased by DCAA, which includes IHS Global Insights economic indices. A list of economic indices that are readily available to all auditors can also be obtained at IHS Global Insight's web site.

E-207 Other Considerations in Using Regression Analysis

a. The equation used in a regression analysis should reasonably represent the relationship between the variables. When, for example, the auditor uses the equation $y = a + bx$, he or she should first establish that a straight line adequately describes the relationship between the two variables. This step will require: (1) scrutiny of the graphically plotted data points contained in a scatter diagram, (2) consideration of how changes in the *x* variable can logically be expected to influence the value of the y variable, and (3) a determination if the values of the coefficients *a* (i.e., the y-intercept) and *b* (i.e., the slope) in particular, generally agree with these logical expectations.

b. Changes in facilities, production methods, and accounting procedures can limit the period of time over which consistent cost data can be obtained without the need for extensive adjustment. The historical data included in a regression analysis should normally cover the entire period over which there is reasonably consistent data. However, it may be

necessary to adjust historical data for past accounting changes. When the period of analysis is short and thus the number of annual data points is small, one might instead use quarterly, monthly or perhaps weekly data in regression analyses in order to have more data points. The selection of an appropriate data interval (i.e., annual, quarterly, monthly, or weekly) for compiling historical cost data will be governed by the contractor's ability to collect data in uniform increments. It may be necessary to use annual data if extensive adjustments are required to normalize quarterly, monthly, or weekly data into equivalent units of production activity or effort. When periods of a month or less are used, it is normally necessary to adjust the data for the effective number of work weeks, days, or hours, as discussed in E-309. Shifting from longer to shorter data intervals increases the number of data points but typically decreases the value of the coefficient of determination (R^2) and the *Adjusted* R^2. There are typically more period-to-period random unexplainable variations in the dependent variable's values when shorter rather than longer data intervals are used. That is, as the period is shortened, the number of observations increases, but the variation to be explained increases as well. Although the R^2 and *Adjusted* R^2 values will likely drop as the number of observations increases, the significance of the regression analysis will not necessarily change materially unless there are only very few observations. Shortening the data interval affects how projections are made using the regression equation. For example, if the original data period was a year but was shortened to a quarter, to project the value for an upcoming year one would make four separate quarterly projections.

c. It is sometimes necessary to adjust historical costs and bases to some extent in order to restate them on an equivalent, comparable basis. This need generally increases as the relevant range expands. Other adjustments may be required to eliminate extraordinary events such as unusual or nonrecurring costs (e.g., cost accounting changes such as shifts in cost center or overhead pool components, changes in the makeup of labor categories, unique interdivisional cost transfers, the realignment or relocation of manufacturing facilities, costs incurred during interruptions in normal activity because of strikes, floods, or fires). Other events that may require adjusting the data include: cost accounting system changes such as transfers of costs recorded in one period being shifted to another period. However, data points should not arbitrarily be eliminated without a stated, logical rationale. Additionally, data points should not be eliminated in an attempt to arbitrarily influence the computed regression line.

d. Sometimes the sign of a coefficient obtained from regression analysis is the opposite from what is logically expected. Although the *a* and *b* coefficients of the regression equation work together, in producing an accurate projection of future costs generally the slope of the regression line is of greater importance than the coefficient *a* (often interpreted as a measure of fixed costs) of the regression equation. The slope of the regression line is represented by the b coefficient in the $y = a + bx$ linear regression equation. The slope is important because using the *bx* term we can compute the change in costs as *x* changes from one value, x_i, to another value x_j *as* $b(x_j-x_i)$. The coefficient *a* is irrelevant in this computation. It is this variable cost component within the dependent variable, which is generally more significant than the fixed cost component of the equation. For example, in terms of a regression line's slope, when evaluating the relationship between direct labor x and overhead costs *y*, it is normally expected that as direct labor increases overhead expenses would likewise increase, implying that the *b* coefficient should be positive. A negative *b* coefficient would indicate that increases in direct labor are accompanied by decreases in overhead. Secondly, in terms of evaluating a regression equation's y-intercept at *x*-zero, a negative value of *a* would imply a negative amount of fixed costs, which is impossible. Be aware that attempting to use linear regression analysis to accurately determine fixed costs as represented by the a coefficient will not generally produce a sufficiently precise measure of those costs. As is often the case, the relevant range of the independent variable(s) data values is so far from *x*-zero that the a value (the *y*-intercept) cannot be reliably interpreted as an accurate measure of fixed cost. Or, some of the indi-

vidual point-to-point y data values may uniquely contain accounting changes or random data anomalies that cause varying ratios of fixed or semi-variable costs in relationship to each dependent variable's value that, in turn, adversely impacts the computed line of best fit. Under these conditions, when the computed coefficients are illogical, the auditor should reexamine the logic between the associated variables to determine if the equation used expresses a valid relationship between the x and y values as well as the point-to-point consistency of the data values used. The auditor should be satisfied regarding the acceptability of the basic data used in the correlation. In some instances the data may require further adjustments as described. When the data appear to be correct but results in an illogical value such as an unexpectedly negative relationship, the auditor should consider forecasting the value of y using some other independent variable or using another method, such as the average or most recent ratio of the y values to the x values. Alternatively, in some cases a curvilinear data relationship may exist which precludes the use of a linear regression analysis model on an arithmetic scale.

e. When applying statistical methods to regression analysis, there is a fundamental assumption that the distance of each y value from the underlying regression line is independent of the other y values. If the dependent variable y values are, for example, moving averages or cumulative averages, this assumption is not met, thus undermining basis for the statistics used to test the significance of the regression.

f. The auditor should be aware of the significance of a "run of points" in the data. A run consisting of a long series of points which are all above or below the regression line may occur when the historical data are arranged chronologically or in order of increasing values of the independent variable. The existence of such runs may indicate that: (1) some factor not considered in the regression equation is influencing the dependent variable, (2) the equation being used in the analysis does not fully describe all the underlying relationships between the variables, and/or (3) the data do not satisfy the assumption of independence. Suppose, for example, that earlier values of the dependent variable are predominantly below the regression line and the later values are predominantly above it. This condition suggests that a general rise in price levels, or some other factor such as an accounting change which is not considered in the regression equation, that may have caused the dependent variable to increase over the period of time covered by the analysis. As an example of the second condition, suppose that all of the small and large values of the independent variable are below the straight line of best fit, while those for intermediate values are above the line. This condition could indicate the existence of a curvilinear relationship between the two variables.

g. Predictions made from a regression equation are most reliable when the independent variable values used for the projection are close to their corresponding average values in the data. In the case of simple linear regression, the regression line passes through the point corresponding to the average observed values of the x and y variables. Note the slope yielded by a regression analysis is an estimate of the true but unknown slope. A small difference between the true slope and the estimated slope, resulting from random fluctuations in the historical data, has an increasingly greater impact on projected values as the independent variable's value moves further from its historical average. Regarding the independent variable value used for projection, even though a regression equation's quality shows it will yield a reasonably good projection within its historical range, as its value moves increasingly further outside the range it yields less reliable predictions because random errors are magnified.

E-300 Section 3 --- Time Series Charts

E-301 Introduction

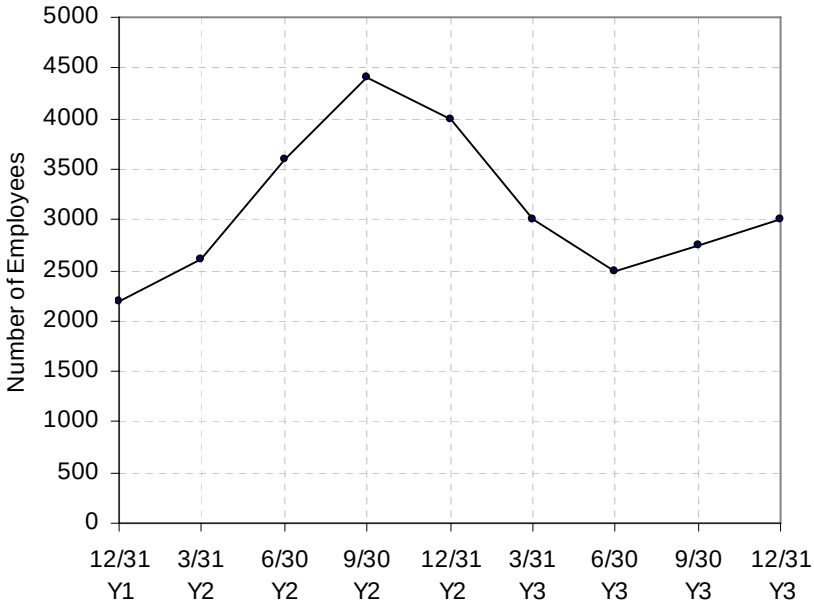
This section presents a discussion of the use of time series analysis to evaluate contract costs.

E-302 Definition and Concept

a. Time series are data which are collected and classified by successive time intervals reflect changes occurring with the passage of time. The graph of a time series is in effect a scatter diagram in which the independent variable has been replaced by a time element, such as years, months, or weeks. An illustration of indirect labor plotted as a time series is given in Figure E-3-1. As shown on the chart, successive plotted points are usually connected by a series of straight lines to show more clearly the variations in the data. This usually is not done in a scatter diagram because the plotted points are seldom in chronological order.

b. Few time series graphs approximate straight lines. A few look like rough curves, but the majority of time series graphs appear as a series of hills and valleys. For this reason, time series analysis is used to bring some order into this seemingly irregular and seemingly erratic pattern. A variety of circumstances and conditions acting simultaneously, with their effects superimposed, give time series their irregular appearance. It is customary to divide the fluctuations of a time series into the following four general classifications: (1) chance and unusual variations, (2) secular or long-term industry and company trends, (3) seasonal variations, and (4) business or economic cycles.

Figure E-3-1
Number of Indirect Employees by Quarters



E-302.1 Comparison of Concepts

a. Time series chart and trend lines have been used for years in analyzing costs and in preparing cost estimates. While in most instances these devices and techniques have been properly and successfully used, a number of cases reflect a lack of understanding of the purpose, nature, and limitations of time series analysis in the evaluation and estimation of costs. Much of the misuse appears to stem from a confusion of the devices and methods of time series analysis with those of correlation analysis. Because of the similarity between these two techniques, the problems involved in the interpretation and use of a time series from the contract auditing standpoint can probably be best visualized in contrast with those of correlation analysis.

b. In the evaluation of contract costs and cost estimates, it is frequently important to know and to measure the interrelationships between the various costs and cost factors; to know whether, with what probability, and under what conditions a change in one cost parameter will be accompanied by a change in another cost parameter. These facts about the cost parameters can only be determined by studying their interrelationships independently of all other parameters, including that of time. The discovery and measurement of these interrelationships, without reference to the sequence of the events in which these relationships occur, is the objective of correlation analysis.

c. Although few occurrences in business are exclusively functions of time, the amount of cost that will be incurred in any operation is directly dependent on the nature and extent of the contractor's facilities and the skill and experience of its workers. These capabilities are dynamic. They are constantly changing and expanding with time as new facilities are acquired and experience is gained. Any evaluation of incurred or forecasted costs must take into consideration this changing pattern of capability. Therefore, any analysis of changes in costs or cost parameters must also take these multiple sequential relationships into account. This is the function of time series analysis: to discover, analyze by type, and depict in sequence the changing values of a variable during a selected period of time.

d. Correlation analysis and time series analysis are supplemental, not alternative procedures. For example, the growth and decline of indirect expense may be related to the changes in the direct labor which may be identified by correlation analysis. But their relationship can be fully understood only if the changes in these variables are also considered in relationship to time. A time series chart displaying the costs of both direct labor and indirect expense might be used for this latter purpose; and these, in turn, could be correlated to changes in plant production capacity, such as the use of extra shifts and extra production lines.

e. A time series is a succession of periodically measured values from a sequence of individual events. A time series line, which joins the otherwise disconnected points, serves only to make fluctuations in the measured values and their sequential relationship more readily apparent. As a result, the segments of the line between the data points have no meaning in regard to intermediate values. Changes in the measured values during a period of time are associated, except in a few instances, with the changing events and not with the passage of time. The trend line of a time series does not portray a mathematical relationship whereby the value of one variable may be computed from the known or assumed value of another variable. Instead, it indicates the general direction of the variable values during the total time period being analyzed.

E-302.2 Application

a. Although the trend line is generally not used as a regression line, this does not mean that it cannot be used in evaluating cost forecasts. The trend line and the moving average line may be projected into the immediate future to the extent the future values will develop in a predictable manner based on the current operating position. However,

their usefulness diminishes rapidly the farther they are extended. If this extension is carried beyond the point where influences may be evaluated, unknown factors intervene so the connection from the current to the future operating status becomes extremely tenuous. The extended lines then lose their value, and their use in making or evaluating an estimate is hazardous. The validity of using the extended trend line or moving average line in connection with cost forecasts is closely associated with: production lead time and planning, industrial and economic conditions, and the forecaster's or evaluator's knowledge and understanding of the cost parameters.

b. A time series readily reveals limits within the historical operations, such as the highest, lowest, and normal values. If similar factors can be expected to operate in the future, these limits may be used as a guide to the audit evaluation of forecasted values. This use of time series is based entirely upon judgment, and though one might use regression analysis to estimate a time-based trend line or curve, the statistical evaluation of a correlation is not typically relevant. The statistical evaluation is irrelevant because a time series has only one variable whose values are sequentially related, and time itself is not viewed as a cause for changes in that variable.

c. Arranging data in the form of a time series facilitates comparisons between data occurring at different points in time and thereby gives meaning to observations that otherwise would signify virtually nothing. For example, it is possible to ascertain whether the data reveals a trend in the movement of the values or whether their occurrences are erratic. If there is a trend, and it disappears and reoccurs over a period of time, it is usually indicative of a cyclic operation, and it would be desirable to determine its nature and causes. A study of the interrelationships implied by coincidental troughs and/or peaks for several cost factors should lead to a better understanding of the effects of changes in production methods and policies. On the other hand, lead time is frequently indicated by changes in one cost parameter regularly preceding those of another parameter. This fact will often enable the auditor to anticipate changes before they appear as historical data.

E-303 Chance and Unusual Variations

Chance and unusual variations of time series are those changes which, being caused by unusual events such as strikes, floods, and fires, do not follow the normal pattern of business operations. To the extent that variations are caused by chance, there is little that can be done to predict the behavior of the time series. However, changes in time series which are due to unusual events may be correlated with the events, simply by inspecting the data. The primary reason for the discovery and isolation of fluctuations due to rare events is that adjustment can subsequently be made for these variations in the investigation of secular trends (i.e., long term industry trends) and seasonal patterns.

E-304 Secular Trends

The term "secular trend" is used in this appendix to refer to any long term industry or company trend as distinguished from long term variations in the business data caused by the business or economic cycle.

a. Secular trends depict a gradual and consistent pattern of long term changes in business operations. Whether a long term trend is secular or a phase of a longer term economic cycle is often difficult to determine; however, generally, it will represent a combination of both. It is for this reason that, as long as the data are comparable, it is advisable to have data covering as long a period of time as practicable.

b. In simple analyses and evaluations of time series, secular trends are usually thought of as straight lines fitted to the series indicating the gradual growth or decline of the variable. For example, if wage rates increase from year to year, we say there is an upward trend. Time series trends are usually shown as straight lines on regular or semi-logarithmic graph paper. A straight line on semi-logarithmic graph paper represents an

exponential curve. This type of curve is applicable to a variable which tends to increase or decrease by a fixed percentage from one period to the next. Few time series are actually of such a simple nature that they closely follow a straight line when plotted on either regular or semi-logarithmic graph paper. For example, national indexes of wage and price levels generally have not closely followed either a simple linear or exponential curve. For reasons of simplicity, straight line trends are the most commonly used and this discussion, also for simplicity, will be limited to their use. Nevertheless, straight lines should be used only if they reasonably represent the trend for the desired period. When a straight line does not reasonably approximate the trend, the analysis requirements for cost auditing purposes can usually be met by drawing a moving average line.

c. As in the case of a scatter diagram, a convenient means for obtaining an objectively fitted trend line to time series data is by the method of least-squares. This approach is useful in studying the behavior of historical data when there is a reasonable degree of correlation between the passage of time and the growth or decline of a variable. Such correlation is demonstrated when there are few changes in the general direction of the slope and if the approximate trend, when plotted on arithmetic paper, is a straight line. However, business data tend to result from the convergence of many variable factors whose effects cannot be easily individually isolated and measured, which result in a high degree of variability in the accumulated data. This high degree of variability may render the mathematical precision of the least-squares trend line ineffective for the analysis of time series data. In this case a rough idea of the trend is needed and that can be obtained by a freehand line based solely on judgment and sight which can be sufficiently accurate. Extreme caution should be exercised in using trend lines fitted to time series data as a forecasting technique.

d. The precision of the semi-averages method lies between the method of least-squares and the freehand drawing of a line. This method, which also may be applied to several partial averages, requires the following steps:

(1) divide the data into two or more parts; normally, two equal parts are used; however, if it would be more appropriate, especially if a very long time period is used, the data can be divided into three or more parts,

(2) calculate the individual average (mean) for each part of the data,

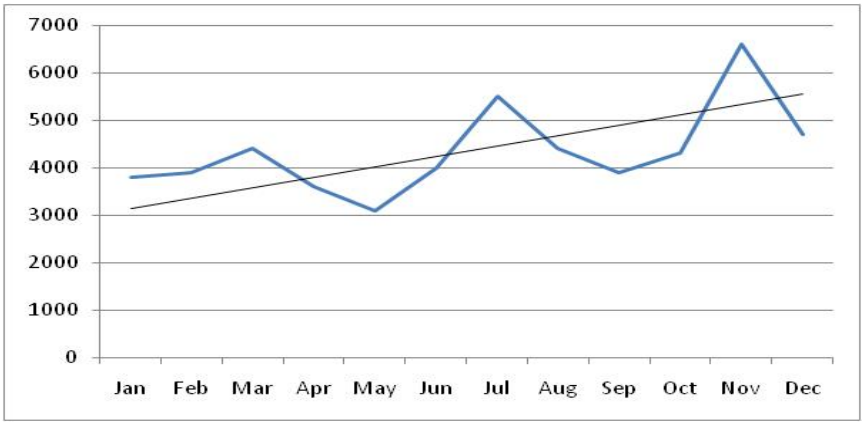
(3) plot these averages at the midpoints of each part of the data, and

(4) extend a smooth line through the points.

(a) To illustrate the method of semi-averages, consider the following data:

Month	Monthly Value	6-Month Total	Monthly Average (Plot point)
Jan	3,800	22,800	3,800
Feb	3,900		
Mar	4,400		
Apr	3,600		
May	3,100		
Jun	4,000		
Jul	5,500	29,400	4,900
Aug	4,400		
Sep	3,900		
Oct	4,300		
Nov	6,600		
Dec	4,700		

Figure E-3-2
Monthly Values vs. Monthly Average



(b) In the above table the six-month totals and the corresponding monthly averages are listed at the midpoint of each averaging period (i.e., between March and April and between September and October). The monthly values from this table are then plotted at the midpoint of each month, and the two averages (3,800 and 4,900) are plotted at the midpoints of each six-month period. The line drawn through these two points is the semi-average trend line.

E-305 Seasonal Variations

a. The type of variation in a time series easiest to understand is seasonal variation, which consists of regularly repeating patterns of moderate or short duration in the contractor's operation. Although the name of this type of variation implies a connection with the seasons of the year, it is used to indicate any kind of variation which is periodic in nature, provided it is also of short duration.

b. While as a rule, it is easy to determine the length or period of seasonal cycles, the description of their characteristics is more involved. There are a number of techniques for describing them, including some which involve rather tedious calculation. A relatively simple method of measuring seasonal patterns, which may be called the simple average method, consists of constructing a seasonal index. This is a descriptive measure that compares, by a series of ratios, the value of each month with the overall monthly average for the entire period.

Seasonal Index of Direct labor Hours

Year	Month	Direct Labor Hours	Seasonal Index
1	Jul	38,414	88.15
	Aug	38,610	88.60
	Sep	45,203	103.73
	Oct	36,734	84.30
	Nov	31,368	71.98
	Dec	41,415	95.04
2	Jan	55,556	127.49
	Feb	44,158	101.33
	Mar	34,545	79.27
	Apr	41,214	94.58
	May	68,252	156.62
	Jun	47,460	108.91
Total		522,929	
Monthly Average (522,929/12)		43,577	

The seasonal index is computed by dividing the labor hours for the month by the monthly average labor hours and multiplying by 100. For example, the seasonal index for the month of July, rounded to 88.15 in the table, is computed as follows:
38,414/43,577 x 100 = 88.15

c. The principal use of the seasonal index in contract auditing is in the evaluation of the relative level of an estimate for an interim period. To illustrate this use, assume that the seasonal labor hour index for the month of May is 156.62, as shown in the above table, and that this value is considered typical. Based on the seasonal pattern the number of labor hours for May could be expected to be approximately 156.62% of the monthly average. If the total for the year were estimated at 600,000, the average monthly estimated hours would be 50,000 (= 600,000/12) and the estimated hours for May would be 78,310 (= 50,000 x 1.5662).

E-306 Business Economic Cycle

a. The economic cycle consists of those up and down movements of a time series which extend over long periods of time and originate from and are associated with general economic conditions such as growth, inflation, and recession. That is, the economic cycle for a particular business would be those variations in a time series for that business which would remain if the trend, the seasonal variation, and the chance or unusual fluctuations were eliminated.

b. The general business economic cycle may be analyzed into a number of different types of cycles associated with various theories which endeavor to explain their occurrence. However, since most cost evaluations for contract auditing purposes cover relatively short periods of time, it is not usually feasible to apply the refined techniques necessary to isolate the business economic cycle.

E-307 The Moving Average

E-307.1 Description

When the time series trend is not a straight line and when in evaluating costs the only interest is in the general movement of the series, it is customary to study the behavior of the series by means of a "moving average." A moving average time series is an artificially

constructed series composed of overlapping averages in which the value for each period is replaced by the average (mean) value for two or more adjacent periods. In computing the second and subsequent averages, the earliest value of the preceding computation is dropped and the next new value in point of time is picked up to yield the new overlapping average. In constructing a graph of a moving average time series each average value is plotted at the midpoint of its respective averaging period. The moving average has the effect of smoothing the minor fluctuations in a time series. In fact, a perfectly regular, periodic pattern can be eliminated from the series by the use of the appropriate moving average. The moving average method, however, will yield a smooth trend only if the variations to be eliminated are essentially stable in both duration and amplitude and if the trend is basically linear. By averaging the effects of the seasonal and other short term variations, the resultant average line will primarily express the approximate general trend of the longer term variations.

E-307.2 Construction – Moving Average

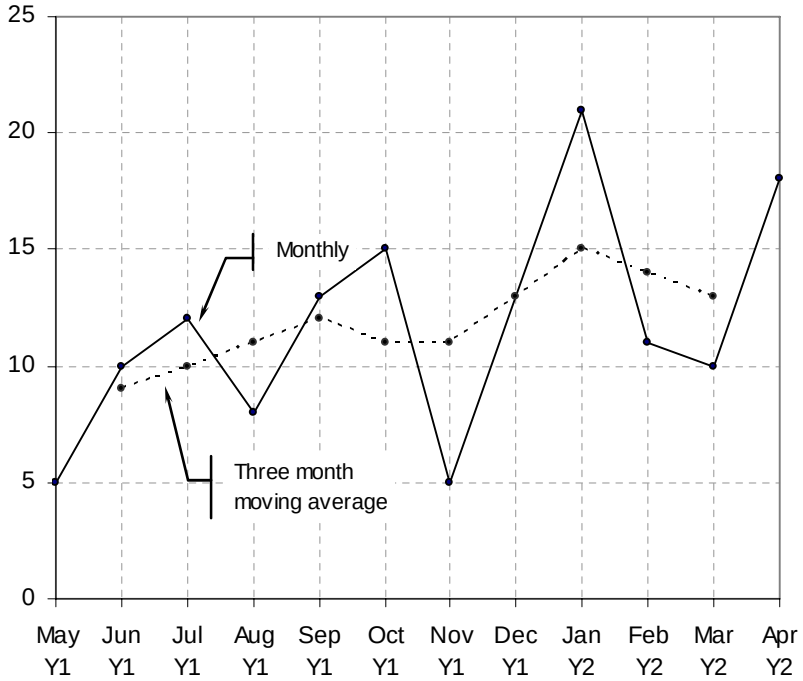
a. The first step in the construction of a moving average time series is the selection of the proper averaging period. As a general rule, it should correspond to the average length of the cycle of the variations to be eliminated. If a shorter period is selected, then influences pertinent to the average but occurring before or after the averaging period will be disregarded and, as a result, part of the cycle will remain in the moving average. On the other hand, if the period is too long, occurrences which in point of time are not pertinent to the average will be included in its computation.

b. The computation of the moving average is illustrated in the following table; a graph showing both the monthly and moving average data is given in Figure E-3-3. For ease and clarity of presentation, an averaging period of three months was selected and data for only one year has been shown. The moving total, column 4, was computed by first adding the values for the first three months and entering the amount opposite the midpoint of the period, the second month. Next a similar total is computed by adding the values for the second, third, and fourth months, and entering the total after the third month. At each subsequent shift of the moving total a new total is computed dropping the earliest of the months previously used and substituting the value for the next succeeding month. The final total (for February, March, and April of year two) is entered opposite March of year two. The moving average values, column 5, are computed by dividing the moving total by the number of terms in the average period; in this case, three.

COMPUTATION OF A THREE-MONTH MOVING AVERAGE FOR FIGURE E-3-3

Year	Month	Monthly Values	Three-Month Moving Total	Three-Month Moving Average
1	May	5		
	Jun	10	27	9
	Jul	12	30	10
	Aug	8	33	11
	Sep	13	36	12
	Oct	15	33	11
	Nov	5	33	11
	Dec	13	39	13
2	Jan	21	45	15
	Feb	11	42	14
	Mar	10	39	13
	Apr	18		

Figure E-3-3
Three-Month Moving Average



E-307.3 General

a. As shown above, both the moving total and moving average are entered at the midpoint of the averaging period. The reason for this is that an actual figure or a calculated figure, referring to a given period of time, refers equally to all portions of that period; and it is usually best to show the figure as applicable to the middle of the period. A yearly figure is referenced to the middle of the year, July first, a monthly figure to the fifteenth of the month, and a figure applying to an odd number of terms (weeks, months, or years) is referenced to the midpoint of the middle period. From this reasoning an average for an even number of terms would be located between the data for the two middle periods, and it may be so shown. However, it is usually desirable to show the average in its customary position, as applying to the middle rather than the end of a period; this requires one more computation to shift the average one half term. It may be accomplished by adding in turn each two successive moving averages and dividing by two. This latter average is then centered at the midpoint of the combined periods. This process, known as "centering the moving average," is applicable to any situation where the moving average period includes an even number of terms.

b. While the moving average will indicate the general direction and extent of a trend, it will closely portray only a linear trend. When the basic trend is curvilinear, the moving average line will follow a path which is inside the curve of data points. In effect, in a curvilinear situation the moving average values do not reflect the full swing of the trend variation, but tend to somewhat reduce it. Another limitation on the use of moving averages is

that they cannot be computed for time periods nearer to the end of the series than approximately one-half of the averaging period.

E-308 Adjustment of Data

It is essential that the data be comparable item by item throughout the analysis period. Because this is not always true of the raw data, it is frequently necessary to adjust the figures before an analysis is undertaken. Two common types of adjustments are discussed below:

a. Adjusting for Changes in Classification Methods.

Since the methods of classifying and apportioning costs may be changed from time to time, the contents of the accounts to be analyzed may also change during the analysis period. Occasionally, these changes are not significant to the analysis and may be disregarded. This, however, is not usually true, and, to the extent the changes are significant, the data accumulated before the change is not comparable to the data accumulated after the change. For evaluation of estimates, the contract auditor is interested in the analysis of only data which were developed under current procedures. If a significant change in procedures is made, the auditor would have a choice of using a shorter analysis period, or adjusting the data to place it on a comparable basis, preferably adjusting the first part of the data. Adjustment of the data is the better method if the analysis period is relatively short. In this case, any further shortening of the period to avoid adjustment of the data could result in a period so short that the full pattern of the variations might not be reflected.

b. Adjusting for Periods of Various Lengths.

(1) Most of the data which the contract auditor will use are available in monthly, weekly, or daily forms; and the auditor must recognize that the work month, work week, and even the work day are variable units of time especially when shifts of varying lengths are often worked. The shortest month is about 10% less than the longest; and the variation may be far greater if holidays or plant shut down periods are taken into consideration. It is also important to recognize that during a 52 week year there is a quarterly 4-4-5 pattern in the number of weeks per month (with 13 weeks per quarter), that will almost always necessitate that monthly data be adjusted to account for the 5-week months. Likewise, the work week may vary from one to seven days; and the work day may vary from a few hours to one containing considerable overtime or to one composed of three eight-hour shifts. There are also instances when either manufacturing or all plant operations may cease for up to three weeks at various times of each year. The solution to the problem, usually, is to put the data on a "per standard working day" basis by dividing the total for each period by the approximate number of standard working days or hours. When there is considerable overtime, more than one shift, shift of different lengths, or an abnormally low work load, the reduced efficiency of these abnormal operations will also likely affect the comparability of the data, depending on the purpose of the analysis. For example, if the object of the analysis is to determine the normal costs for the operation of a production center at normal levels of production, reduction of the data to a standard working day basis without recognition of the extra costs due to the lowered efficiency of extended operations could distort the results of the analysis.

(2) The adjustment for the number of working days is necessary for any series, like that of monthly production and cost totals, where each individual working day means a more or less proportionate increase in the data values. There are, however, two kinds of data which do not require adjustment to a working day or similar basis: those involving values as of certain dates and those stated as averages. For example, figures showing the number of workers in departments are usually based on the number on the payroll on a certain day of the month; and inventory quantities and values are also stated as of a specific day. On the other hand, monthly average labor rates, and percentages, and index

numbers are obviously values which do not depend in any systematic way on the length of the period to which they are applied.

E-309 Length of the Analysis Period

a. The length of the selected analysis period can greatly affect the outcome of a trend analysis. Although no simple rules can be given for determining the proper length, the period normally should be sufficiently long to reflect the full pattern of changes that may be pertinent to the purpose of the audit. When this is not possible and a trend analysis is prepared on the basis of a few observations, the auditor should bear in mind the resulting limitations.

b. In selecting the period for fitting a trend line to cyclical data, the series should normally start and end in about the same phase of the cycle. For example, if it is to end during a period of high costs, then it normally should also start during a period of high costs. If it starts and ends at different points in the cycle, then the slope of the trend line is apt to be steeper or shallower than conditions warrant. On the other hand, changes that suddenly reverse a trend are not uncommon; and, if there is a distinct or abrupt change in the series, it is usually best to break the series into two parts and fit separate trend lines.

E-400 Section 4 --- Audit Applications of Graphic and Computational Techniques

E-401 Introduction

This section presents illustrations of the application of graphic and computational analysis to specific issues of cost evaluation in the field of contract cost auditing.

E-402 Identification of Departures from Historical Cost Patterns

E-402.1 General

a. Both time series analysis and scatter diagrams are easily used and especially helpful devices in focusing attention on costs which deviate from historical patterns. Regression analysis can also be used for this purpose. Whether or not a specific historical trend may be applicable to a current situation can be determined by a direct comparison of past and current data. The auditor should be aware of the cost patterns which are normal to the contractor's operation, and should look for, recognize, and investigate significant departures from normal patterns. Special attention should be given to the impact of any significant changes in products or production methods.

b. The auditor's goal is to identify those accounts which have deviated significantly from experienced trends and thus require a more detailed analysis. If, for example, there are significant changes in overhead expenses, overhead rates or direct labor rates which cannot be explained by a corresponding change in the level of production, the auditor will want to know the reason. Such changes may signal lack of management control over cost levels or a change in accounting procedures. This does not mean that costs falling within the normal pattern are automatically acceptable and need not be examined. The auditor should analyze individual expenditures which are so large that they have a material impact on Government contract costs. In addition, samples of items in cost categories which account for large proportions of the costs being audited should be examined.

E-402.2 Compilation of Data

The data to be collected can be easily retrieved electronically using software such as the Agency’s SAS or other software that can retrieve data directly from the contractor’s information systems (such as Enterprise Resource Planning, i.e., SAP, J.D. Edwards, Oracle Financials, etc.) and Oracle or other types of database systems. It can be analyzed at various detail levels from the contractor's accounts and can be done on a daily, weekly, monthly, quarterly, or an annual basis. (Contact your Regional Office or the Headquarters Technical Services Division, OTS, if assistance is required. The following is a non-inclusive list of basic cost elements, accounts or groups of accounts that may be useful for analysis purposes. In addition to collecting and analyzing current data for such variables, valuable information can be gained from tracking budgeted or forecasted amounts.

a. Overhead expenses, whether by a pool’s, total cost, or a lower tier cost center, or at an account level, or a homogeneous group of accounts, such as:

(1) Indirect labor
(2) Employee benefits and payroll taxes
(3) Repairs and maintenance
(4) Marketing
(5) Bid and proposal expense
(6) Allocations
(7) Other variable overhead expenses
(8) Fixed expenses (depreciation, rent, etc.)

b. Overhead base for each pool.

c. Direct costs and other variables related to production, such as:
(1) Direct material dollars
(2) Direct labor dollars and/or hours
(3) Other direct charges (tooling, travel, etc.)
(4) Sales
(5) Cost of sales or cost input
(6) Employees work force changes, i.e., hired onboard, departed
(7) Number of employees, direct and indirect
(8) Years of experience/skill level of employees
(9) Changes in Square feet of plant space-idle facilities/expanded site operations

E-402.3 Analyzing the Data

Some examples of the types of graphs the auditor may wish to consider in identifying departures from normal cost patterns follow. The more points of data included in an analysis, the more useful the analysis will be for identifying deviations from historical patterns.

a. Time-Series Analysis.

(1) Overhead rates by month. Both the actual monthly and year-to-date rates should be plotted on a time-series graph. The current estimating rate should also be plotted for comparison. These graphs will enable the auditor to become aware of within-the-year trends and evaluate whether the current estimating rate is acceptable for the remainder of the year. If rates from prior years are also plotted this will enable the auditor to identify months which deviate from experienced trends.

(2) Relationship of estimated amounts of key variables to actual amounts. The objective of these analyses is to determine if there is any pattern of underestimation or overestimation by the contractor. For example, does the contractor consistently underestimate sales, the number of employees, or the amount of the overhead base? Does the contractor overestimate certain expenses which are not particularly susceptible to audit? Over a period of time, such patterns, if they exist, will show up and the auditor can consider them in the audit of future estimating rates. The analyses can be accomplished by plotting the estimates on graphs used to track the actual data.

b. Scatter Diagrams.

(1) Relationship of overhead dollars to the base of allocation. The total overhead dollars are plotted, preferably monthly or quarterly, on a scatter diagram on which overhead dollars are scaled on the *y* (vertical) axis and the base is scaled on the *x* (horizontal) axis. This type of graph can also be used to compare individual accounts or groups of accounts with the base. Such graphs will assist the auditor in determining when overhead costs for a period are out of pattern. In addition, it will give the auditor some insight into whether the overhead costs are dependent on the base of allocation or whether they might be controlled by other factors instead.

(2) Relationship of overhead expenses to other variables. For example, scatter diagrams relating total overhead expenses to sales, cost of sales, or direct labor will enable the auditor to evaluate whether overhead costs are increasing or decreasing consistently with changes in business volume. A scatter diagram relating indirect labor dollars, hours or employees to direct labor dollars, hours or employees will indicate whether indirect labor is increasing or decreasing disproportionately. A graph relating rent, depreciation and/or maintenance expense to robotic milling machine hours may be helpful in determining the reasonableness of these expenses. A graph relating the expenses of the payroll and/or personnel operations to the number of employees is helpful in evaluating the reasonableness of the expenses of these operations. Other potentially useful graphs are the relationship of allocated expenses to the appropriate allocation bases; the relationship of employee benefit costs to the total labor dollars; the relationship of recruitment and personnel costs to the number of employees hired; and the relationship of material handling costs to material costs.

E-403 Evaluation of Proposed Overhead Costs

a. One of the most important applications of regression analysis is analytical testing of the reasonableness of estimated overhead costs. Overhead rates vary in response to many causes but because many overhead costs are fixed or semi-fixed, the level of operations is one of the most important factors. Overhead rates generally are lowest when a plant is operating at capacity, and the rates can increase substantially as production levels are reduced. This relationship becomes extremely significant when evaluating bid proposals resulting in large changes in the level of contracting effort from low to higher volumes of production, and even more so for production level changes that will occur over many future periods of time. In these cases it is necessary to predict overhead rates many years in advance on the basis of operating levels projected for these years.

b. The simplest way of using regression analysis to evaluate proposed overhead costs is to fit a least-squares line to the historical amounts of an overhead pool (the dependent variable) and the base of allocation or some other predictable measure of the level of operations (the independent variable). A computerized example is presented in E-202.2. This procedure assumes the overhead pool consists of: (1) fixed expenses which remain approximately the same at all operating levels, (2) variable expenses which tend to be directly proportionate to the base, and (3) semi-fixed expenses which cannot be reduced below a certain level while the plant is in operation and will increase above this level in proportion to the base. The value obtained for "*a*" (i.e., the y-intercept) is loosely assumed to represent the average value of the fixed expense plus the fixed portion of semi-fixed expenses during the period covered by the historical data. However, be aware that many factors influence interpretation of the "a" (i.e., y-intercept) value, including how far the relevant operating range of data point values of the x (i.e., independent) variables is from the x-y origin. Thus, the concept of the y-intercept accurately representing fixed costs is typically only a theoretical possibility. The value obtained for "b" (i.e., the slope) represents the average ratio of the variable expenses plus the variable portion of the semi-fixed expenses to the base. Estimates of the a and b coefficients can also be obtained by judgmentally segregating fixed and variable expenses based on an evaluation of each account in the overhead pool. However, many overhead accounts consist of semi-fixed expenses and it is extremely difficult to judgmentally determine the amount of fixed and variable expense in each such account. Regression analysis may be an easier means of estimating the total amount of fixed and variable expenses in an overhead pool when the observed data points are not far from the y axis and within a relevant range, and the observed data points are tightly grouped around the trend line. Caution should be exercised in the use of the *a* value obtained from a regression analysis as the total amount of fixed expense however, in many contract audit applications, the use of simple linear regression has been found reliable to predict overhead costs. The following paragraphs describe refinements to this basic procedure which may be needed to improve predictions of overhead. The extent to which these refinements are appropriate in any given audit situation depends on the extent to which departures from the assumptions cited earlier in this paragraph can be expected to affect Government contract prices.

c. If the variables included in the regression analysis have been significantly affected by changes in major economic factors (such as inflation or wage levels), adjustments will be required. Adjustments may also be required to eliminate unusual or nonrecurring costs and to compensate for accounting system changes.

d. Some overhead expenses can often be predicted more accurately by methods other than regression analysis. For example, future depreciation expense may be obtained from depreciation schedules for existing facilities and planned acquisitions. Likewise, future lease expenses may be predicted based on existing leases. Acceptable research and development and bid and proposal expenses may be predictable based on advance agreements. Changes in union agreements, management policies, payroll tax laws, and the rate of inflation can cause significant deviations in employee

benefit costs from past patterns. Consequently, it is necessary to predict employee benefits based on a judgmental consideration of these factors for the projected employment levels.

e. In some situations management may be planning to institute changes which will result in greater operating efficiencies and lower costs. The auditor should adjust the historical data accordingly as it will not properly consider the resulting reduced costs.

f. Aside from CAS issues, some overhead expenses may vary in response to a parameter other than the base of allocation. If the parameter is an accurate predictor, then better forecasts of future related costs may be obtained by applying regression analysis using these historical amounts. In some cases, more than one parameter can be identified as influencing certain expenses. For example, maintenance costs may vary in response to both the number of machines in active use and the amount of direct labor. Multiple regression analysis might be considered as a means of forecasting such expenses.

g. The use of regression analysis assumes overhead costs will be the same in the future as in the past except for predictable changes in the independent variable(s). If a change in the basic pattern of an expense is anticipated, regression analysis applied to unadjusted historical data will obviously not provide the most useful results. The historical data can perhaps be adjusted to reflect the unchanged portions of past operations and the changed portion of operations can be independently projected, with the results then being combined, In any event, when significant changes in production capacity are planned, it generally will not be reasonable to always assume that past patterns will continue.

E-404 Evaluation of Allocation Bases

a. While perfect correlation between any overhead pool and any available base of allocation is highly unlikely, some bases of allocation will provide better correlation than others. The better the correlation, the more likely it is a base will provide an equitable distribution of overhead. One of the principal arguments that can be presented to support an audit recommendation for changing a contractor's overhead allocation base is to show in the auditor's proposed overhead pool has historically changed more closely in relationship to the proposed base recommended by the auditor than the contractor's allocation base. An objective evaluation of the relative effectiveness of alternative allocation bases can be obtained by comparing the coefficients of determination R^2obtained as a result of fitting a least-squares trend line to different pool base comparison situations. In order to maintain comparability between the coefficients, data for the same periods must be included in each analysis. The relationship between the various bases tested relative to the specific overhead pool should be tested over a sufficient period of time to establish there is historic stability. Both the conditions tending to produce stability and the causes of any material deviation should be evaluated.

b. Better predictions of some expenses in an overhead pool may be obtained by correlating them with some parameter other than the base of allocation. When this situation is encountered, consideration should be given to recommending these expenses be included in a separate pool for allocation based on this parameter if allocated costs would be significantly different.

E-405 Evaluation of Proposed Labor Rates

E-405.1 General

a. Labor rates have generally increased from one period to the next because of inflation alone, so that over time labor rates have risen for individuals with the same qualifications doing the same job. In analyzing labor rates projected for a Government contract, the auditor must be sure the types of employees whose wages are being projected are the same as those

to be assigned to the contract. It would not be proper, for example, to project engineering department labor costs based on average departmental rates, if a proposed contract will require a disproportionate amount of time by lower paid, entry level engineers. In addition to the category of work required, the auditor must consider changes in the projected level of production, (e.g., if new workers are to be hired, they will probably start at entry level wage rates that are considerably lower than the current average).

b. In the usual application of regression analysis to the prediction of labor rates, average historical rates are first adjusted to eliminate the effect of changes in wage-related economic factors. The adjusted wage rates are then correlated with the plant population or the number of labor hours. The logic underlying this procedure is that as the number of employees increases, new employees will be hired at rates below the plant average, thereby reducing the average labor rate. Conversely, when the plant population is reduced, the employees with the least seniority and experience will be released first, thereby increasing the average labor rate. This procedure has produced excellent results in many contract audit applications. However, it does have limitations. It is most appropriate when the historical data included in the regression analysis is relatively recent, so the auditor can be reasonably sure the manufacturing techniques and conditions have not changed substantially during the period of time covered by the data. It is also best when the projected plant population falls within, or close to, the range of population levels included in the data. A possible complication is that although a plant's population may vary at times, average seniority and hence labor rates may tend to rise during extended periods (say several years) of stable plant population. Under such circumstances, an evaluation based on estimates of the number of employees to be hired and separated and the average wage rates of the employees in each category may be required. In any projection of labor rates, the impact of expected changes in economic factors and employee mix must be considered along with the impact of expected changes in the number of employees.

c. Limitations in regard to the use of time series analysis are pertinent to the forecast of labor rates using multiple regression analysis. In other words, wherever time is not appropriate as an independent variable in a simple regression analysis, it is equally not appropriate when combined with another independent variable in a multiple regression analysis.

E-405.2 Illustration of the Use of Analysis in the Evaluation of Labor Rates

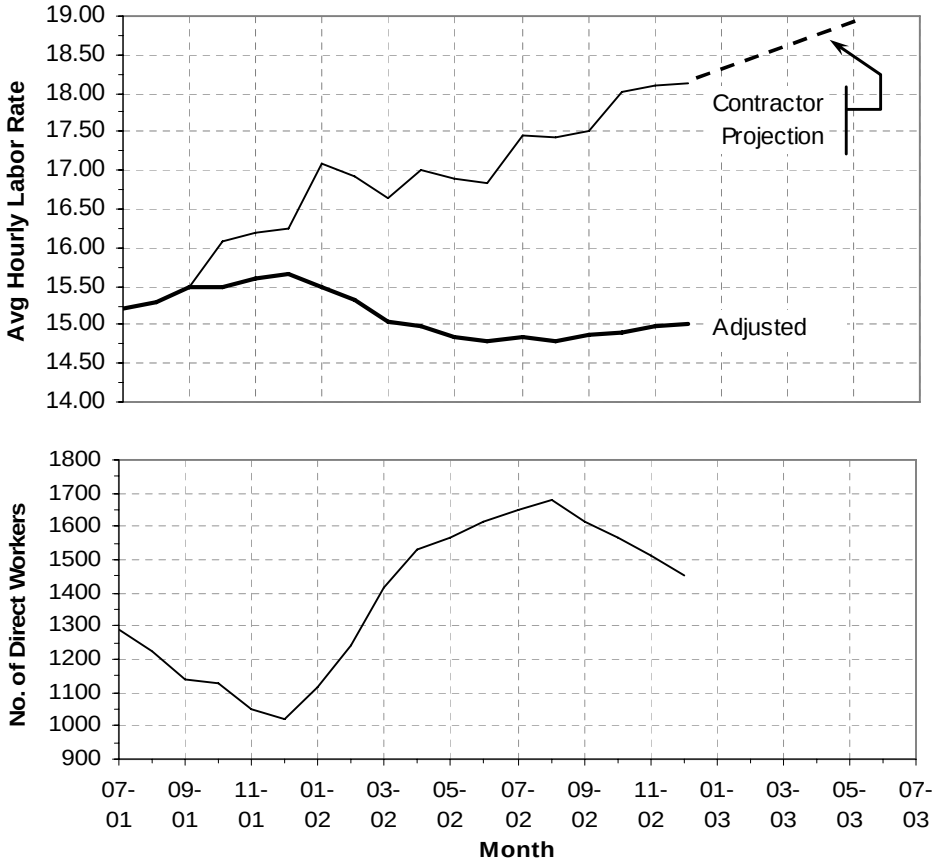
a. The time series charts in Figure E-4-1 graphically portrays data on the number and average hourly rate of a contractor's direct manufacturing employees for an 18-month period starting July 1 of year one and ending December 31 of year two. The data are shown in Table E-4-1.

Table E-4-1
Average Monthly Direct Labor Hourly Rates
and Number of Direct Workers

Year	*Month*	*Total Wages*	*Number of Direct Workers*	*Avg Hourly Labor Rate*	*Cum Rate Increase*	*Adjusted Avg Hourly Labor Rate*	*Adjusted Total Wages*
1	Jul	19,605.69	1289	15.21		15.21	19,605.69
	Aug	18,773.10	1227	15.30		15.30	18,773.10
	Sep	17,693.64	1143	15.48		15.48	17,693.64
	Oct	18,149.52	1128	16.09	0.59	15.50	17,484.00
	Nov	17,058.60	1053	16.20	0.59	15.61	16,437.33
	Dec	16,607.50	1022	16.25	0.59	15.66	16,004.52
2	Jan	19,100.70	1117	17.10	1.60	15.50	17,313.50
	Feb	21,048.48	1244	16.92	1.60	15.32	19,058.08
	Mar	23,597.97	1419	16.63	1.60	15.03	21,327.57
	Apr	26,059.32	1532	17.01	2.03	14.98	22,949.36
	May	26,400.32	1564	16.88	2.03	14.85	23,225.40
	Jun	27,214.11	1617	16.83	2.03	14.80	23,931.60
	Jul	28,843.92	1652	17.46	2.63	14.83	24,499.16
	Aug	29,300.44	1682	17.42	2.63	14.79	24,876.78
	Sep	28,243.63	1613	17.51	2.63	14.88	24,001.44
	Oct	28,255.36	1568	18.02	3.13	14.89	23,347.52
	Nov	27,400.43	1513	18.11	3.13	14.98	22,664.74
	Dec	26,339.28	1452	18.14	3.13	15.01	21,794.52

b. Two charts are shown in Figure E-4-1. The upper chart refers to hourly labor rates, exclusive of premium overtime pay. The top curve in this chart depicts the "Avg Hourly Labor Rate" series and the lower curve in this same chart depicts the "Adjusted Avg Hourly Labor Rate", both series being listed in Table E-4-1. The adjustment eliminates the effect of general pay increases granted during the period. In this example a union contract provides for quarterly cost-of-living adjustments to wage rates and a general wage increase was negotiated as part of a 2-year contract effective in January of year two. Hourly pay increases total to $3.13. The peak month of employment occurred in August of year two when there were 1,682 workers. The cumulative pay increases up to that time totaled $2.63. Adding this increase to July of year one starting rate of $15.21 would make a prospective average rate of $17.84 ($15.21 + $2.63) as of August of year 2. The average rate excluding overtime at that time was however, only $17.42. The decline of 42 cents ($17.84 - $17.42) resulted from other causes, primarily the increase in the number of direct workers from 1,289 in July of year one to 1,682 in August of year two. During the following four months when the number of workers declined to 1,452, the trend was reversed and the average wage rate increased 72 cents, of which only 50 cents was due to a cost-of-living increases. A projection line has been constructed indicating an expected increase in the average hourly labor cost over the next year of 17 cents per month. This represents a projection made by the contractor and corresponds to the average experience for the last 18 months. Two defects in the contractor's projection are apparent: (1) the projected rate of increase reflects the general wage increase granted in January of year two although only cost-of-living increases are provided under the union contract during the period covered by the projection and (2) it gives no consideration to any anticipated changes in the number of direct employees.

Figure E-4-1
Average Monthly Direct Labor Rates
and Number of Direct workers



c. It is apparent that changes in the number of employees have had an impact on adjusted hourly rates, as can be seen by comparing the curve for the number of workers and the curve for the adjusted labor rate. However, the precise relationship between these two variables cannot be determined from the graph. A ready means to estimate the relationship is to use simple linear regression, regressing total adjusted wages (the rightmost column in Table E-4-1) on the number of workers. This yields the following estimated coefficients.

Coefficients		Value	*t*	*df*	Significance (*p*)	Confidence (1-*p*)•100
a	Intercept	2352.35	36.819	16	< 0.000001	> 99.9999
b	No. of Workers	13.39423	292.89	16	< 0.000001	> 99.9999

The statistics for the entire regression are as follow.

R^2	F	Degrees of freedom (df_1,df_2)	Significance (α)	Confidence (1-α)•100
0.999814	85784.31	1, 16	< 0.000001	> 99.9999

The estimated equation is

Adjusted Total Wages = 2352.35 + 13.39423 • Number of workers

This equation could be used to estimate future values of adjusted total wages. However, dividing both sides of the equation by the number of worker yields the following equation for estimating future values of the adjusted average labor rate. That is,

Adjusted Average Labor Rate = 13.39423 + 2352.35/Number of Workers

d. The following computation illustrates how the foregoing analysis might be used in the evaluation of proposed direct labor costs. It is assumed that the contract will be performed in the second, third and fourth quarters of Year three with employment reaching a peak in the third quarter.

	Year 3		
	2nd Qtr	3rd Qtr	4th Qtr
Estimated Number of Employees (Workers)	1,500	1,650	1,450
Estimated average rates using equation developed with regression analysis (base is Jul of year 1)	$14.96	$14.82	$15.02
Estimated cumulative increases since Jul of year 1 ($3.13 in 4th Qtr of year 2, plus $.61, $.68 and $.70 for each respective Qtr thereafter)	$3.74	$4.42	$5.12
Estimated hourly labor rate	$18.70	$19.24	$20.14
Estimated direct labor hours for proposed contract	62,000	119,000	31,500
Estimated total direct labor cost for proposed contract	$1,159,400	$2,289,560	$634,410

E-405.3 Alternative Methods

a. In the preceding example, the adjustment of historical data for pay increases was simplified by the fact that the union contract provided for uniform increases for all workers. If the increases had not been uniform, a more detailed computation based on the amount of the raise granted to each class of worker and the number of workers in each class would have been necessary. When the number of workers in each class is not readily available, estimates may be based on samples of the workers.

b. At some contractor locations, union pay increases are granted to non-union workers. In such cases, the union agreement can be used to adjust the rates paid to non-union personnel. Even at some plants which are not unionized, management grants across-the-board increases to employees to compensate for changes in economic levels. Adjustment of historical data for such increases can be accomplished in the same manner as for union increases.

E-406 Evaluating Standards

a. Correlation techniques may also be used in evaluating standards, such as:
(1) material handling costs,
(2) scrap, rework, and spoilage,
(3) export packaging costs, and
(4) field service warranty or guarantee expense.

b. As indicated in Chapter 9, contractors frequently use loading factors based on such standards in developing cost estimates. Before accepting such a standard, the auditor should know the:

- relationship between the costs included in the standard and the costs to which the standard is applied, and
- causes of any material deviations from this relationship.