Regression Analysis Supplement

This supplement provides advanced coverage of regression analysis. The advanced coverage is not required for any of the exercises or problems at the end of Chapter 8. While the text provides comprehensive and in-depth coverage of regression analysis, some instructors and students can pursue an advanced coverage of regression analysis through the use of this supplement. While the exercises and problems in the text do not require this advanced coverage, those exercises and problems that deal with regression analysis could be addressed at an advanced level using the content provided in this online supplement.

This supplement uses an example to explain the development of a regression estimate and the related statistical measures. The supplement then interprets the statistical measures to assess the precision and reliability of the regression.

THE REGRESSION ESTIMATE

To illustrate the manner in which a regression estimate is obtained, we use the data in Chapter 8 of the text (see Exhibit 8.3). Recall that regression analysis finds the unique line through the data that minimizes the sum of the squares of the errors, where the error is measured as the difference between the values predicted by the regression and the actual values for the dependent variable. In this example, the dependent variable, supplies expense (Y), is estimated with a single independent variable, production level (X). The regression for the three data points is

$$Y = a + (b \times X) = \$220 + (\$0.75 \times X)$$

The intercept term, labeled *a,* and the coefficient of the independent variable, labeled *b,* are obtained from a set of calculations performed by spreadsheet and other programs and are described in basic textbooks on probability and statistics. The calculations themselves are beyond the scope of this text. Our focus is on the derivation and interpretation of the statistical measures that tell management accountants something about the reliability and precision of the regression.

STATISTICAL MEASURES

The statistical measures of the reliability and precision of the regression are derived from an analysis of the variance of the dependent variable. *Variance* is a measure of the degree to which the values of the dependent variable vary about its mean. The term *analysis of variance* is used because the regression analysis is based on a separation of the total variance of the dependent variable into error and explained components. The underlying concept is that in predicting individual values for the dependent variable, the regression is *explaining changes (i.e., variance) in the dependent variable* associated with changes in the independent variable. The variance in the dependent variable that is not explained is called the residual, or *error variance.* Thus, the regression's ability to correctly predict changes in the dependent variable is a key measure of its reliability and is measured by the ratio of explained variance to error variance. Based on the data in Exhibits 8.3 and 8.4 in the text, Exhibit S.1 shows how the variance measures are obtained.

The first two columns of Exhibit S.1 show the data for the independent (X) and dependent (Y) variables. Column (3) shows the mean of the dependent variable (YM), and column (4) shows the regression prediction (YE) for each of the points. The last three columns indicate the three variance measures. Column (5) shows the total variance, or variance of the dependent variable, measured as the difference between each data point and the mean of the dependent variable ($Y - YM$). Column (6) shows the variance explained by the regression ($YE - YM$), and column (7) shows the error variance, ($Y - YE$). The measures

EXHIBIT S.1 **Variance Components for Regression Analysis: Total Variance, Regression Variance, and Error Variance**

				Variance Components		
1 Dependent Variable *Y*	2 Independent Variable *X*	3 Mean of *Y* *(YM)*	4 Regression Prediction for *Y* *(YE)*	5 Total Variance of *Y* *(T) = (Y − YM)*	6 Regression (Explained) Variance *(R) = (YE − YM)*	7 Error Variance *(E) = (Y − YE)*
250	50	295	257.5	(45)	(37.5)	(7.5)
310	100	295	295.0	15	0.0	15.0
325	150	295	332.5	30	37.5	(7.5)

in these last three columns are squared and summed to arrive at the desired values for *total* variance, explained variance, and error variance, respectively. The sum of the error and explained variance terms equals total variance. These terms are illustrated in Exhibit S.2 and the values calculated in Exhibit S.3.

The **analysis of variance table** separates the total variance of the dependent variable into both error and explained variance components.

The three variance terms are the basic elements of the statistical analysis of the regression. This is best illustrated in the analysis of variance table in Exhibit S.3. The **analysis of variance table** separates the total variance of the dependent variable into both error and explained components. The first two columns of the table show the type and amount of variance for each of the three variance terms. The third column shows the **degrees of freedom** for each component, which represents the number of independent choices that can be made for that component. Thus, the number of degrees of freedom for the explained variance component is always equal to the number of independent variables, and the total degrees of freedom is always equal to the number of data points less 1. The error degrees of freedom equal the total less the explained degrees of freedom.

The **degrees of freedom** for each component of variance represents the number of independent choices that can be made for that component.

EXHIBIT S.2
Variance Components for Regression Analysis

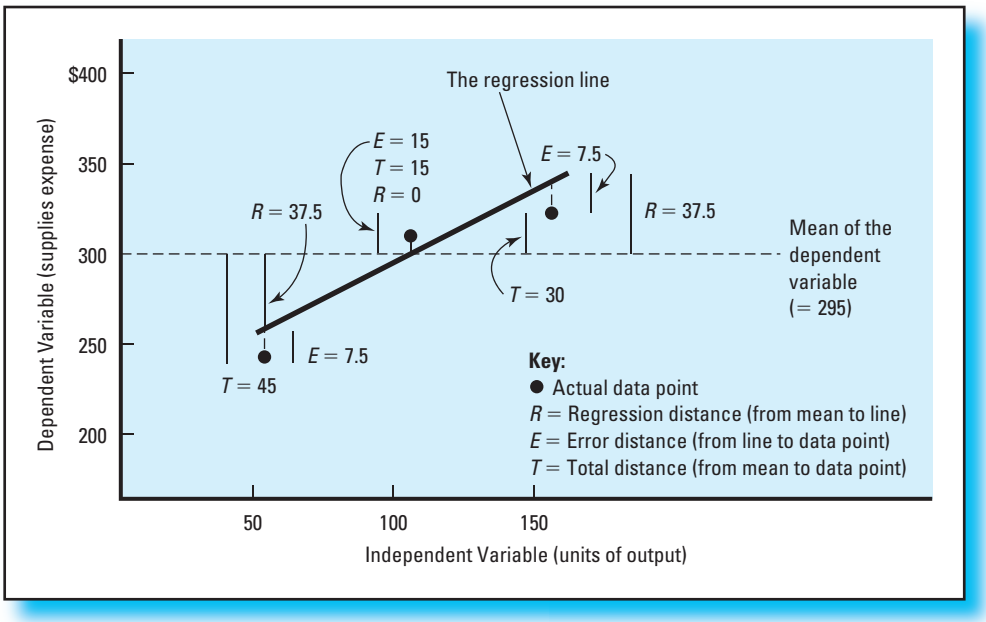


EXHIBIT S.3
Analysis of Variance Table for Regression Analysis

Source of Variance	Variance of Each Component of the Regression (also called *sum of squares*)	Degrees of Freedom	Mean Squared Variance
Regression (explained)	$37.5^2 + 0^2 + 37.5^2 =$ **2,812.5**	1	2,812.5
Error	$7.5^2 + 15^2 + 7.5^2 =$ **337.5**	1	337.5
Total	$(45)^2 + (15)^2 + (30)^2 =$ **3,150.0**	2	1,575.0

EXHIBIT S.4
Six Key Statistical Measures

Precision

1. Precision of the regression (measured by the standard error of the estimate)

Reliability

2. Goodness of fit (*R*-squared)
3. Statistical reliability (*F*-statistic)
4. Statistical reliability for each independent variable (*t*-value)
5. Reliability of precision
6. Nonindependence of errors (Durbin-Watson statistic)

Mean squared variance is the ratio of the amount of variance of a component to the number of degrees of freedom for that component.

The fourth column, **mean squared variance,** is the ratio of the amount of the variance of a component (in the second column) to the number of degrees of freedom (in the third column). For example, the mean squared error is 337.5.

The analysis of variance table serves as a useful basis to discuss the key statistical measures of the regression. Of the six principal measures in Exhibit S.4, one measure refers to the precision of the regression and five measures refer to the reliability of the regression. *Precision* refers to the ability of the regression to provide accurate estimates—how close the regression's estimates are to the unknown true value. *Reliability* refers to the confidence the user can have that the regression is valid; that is, how likely the regression is to continue to provide accurate predictions over time and for different levels of the independent variables.

Precision of the Regression

The standard error of the estimate (SE) is a useful measure of the accuracy of the regression's estimates. The standard error is interpreted as a range of values around the regression estimate such that the management accountant can be approximately 67% confident the actual value lies in this range (see Exhibits 8.7A and 8.7B in the text). An inverse relationship, and therefore a trade-off, exists between the confidence level and the width of the interval. The value of the SE for a given regression can be obtained directly from the analysis of variance table as follows:

$$\begin{aligned} \text{SE} &= \sqrt{\text{Mean square error}} \\ &= \sqrt{\frac{\text{Error variance}}{\text{Error degrees of freedom}}} \\ &= \sqrt{\frac{337.5}{1}} = 18.37 \end{aligned}$$

The precision and accuracy of the regression improve as the variance for error is reduced and as the number of error degrees of freedom increases (the error degrees of freedom increases with the number of data points and decreases with the number of independent variables).

Goodness of Fit (*R*-squared)

R-squared (also called the *coefficient of determination*) is a direct measure of the explanatory power of the regression. It measures the percent of variance in the dependent variable that can be explained by the independent variable. *R*-squared is calculated as follows, from the information in Exhibit S.3

$$\begin{aligned} R^2 &= \frac{\Sigma \text{ of squares (explained)}}{\Sigma \text{ of squares (total)}} \\ &= \frac{2{,}812.5}{3{,}150} = .892 \end{aligned}$$

The explanatory power of the regression improves as the explained sum of squares increases relative to the total sum of squares. A value close to 1 reflects a good-fitting regression with strong explanatory power. Note that *R*-squared and SE travel in opposite directions. A regression with a high *R*-squared will have a relatively small SE and vice versa.

Statistical Reliability (*F*-Statistic)

The *F*-statistic is a useful measure of the statistical reliability of the regression.

The ***F*-statistic** is a useful measure of the statistical reliability of the regression. Statistical reliability asks whether the relationship between the variables in the regression actually exists or whether the correlation between the variables is a chance relationship of the data at hand. If only a small number of data points are used, it is possible to have a relatively high *R*-squared (if the regression is a good fit to the data points), but this offers relatively little confidence that a statistical relationship exists because of the small number of data points.

The larger the *F,* the lower the risk that the regression is statistically unreliable. The determination of an acceptable *F*-value depends on the number of data points, but the required *F*-value decreases as the number of data points increases. Most regression software programs show the *F*-value and the related *p*-value. The *p*-value should be less than approximately 5%. The *F*-statistic can be obtained from the analysis of variance table as follows:

$$F = \frac{\text{Mean square variance for regression (explained)}}{\text{Mean square error}}$$

$$= \frac{2{,}812.5}{337.5} = 8.333$$

Statistical Reliability for Each of the Independent Variables (*t*-value)

The *t*-value is a measure of the reliability of each independent variable and, as such, it has an interpretation very much like that of the *F*-statistic. The *t*-value equals the ratio of the coefficient of the independent variable to the standard error of the coefficient for that independent variable. The standard error of the coefficient is not the same as the standard error of the estimate, but it is interpreted in the same way. However, the standard error of the coefficient cannot be obtained directly from the analysis of variance table. For the data in Exhibit S.1, the value of the standard error for the coefficient is .2598.[1] The *t*-value is thus

$$t = .75/.2598 = 2.8868$$

A *t*-value larger than 2.0 indicates that the independent variable is reliable at a risk level less than approximately 5% and is therefore a reliable independent variable to include in the regression. Regression software such as Excel shows the 95% confidence range for the coefficient of each of the independent variables. The range of the standard error of the estimate should be relatively small. A small range provides confidence in the accuracy of the coefficient's value.

Reliability of Precision

Nonconstant variance is the condition when the variance of the errors is not constant over the range of the independent variable.

For certain sets of data, the standard error of the estimate varies over the range of the independent variable. The variance of the errors is not constant over the range of the independent variable. This condition is referred to as **nonconstant variance.** This is the case, for example, when the relationship between the independent and dependent variables becomes less stable over time. This type of behavior is illustrated in Exhibit S.5. If there is nonconstant variance, the SE value provided by the regression is not uniformly accurate over the range of the independent variable.[2]

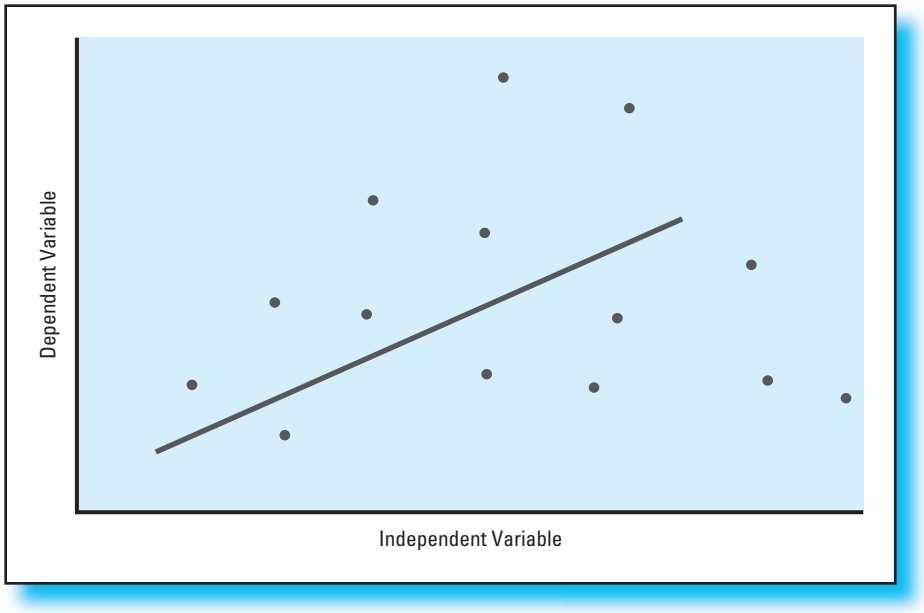
[1] The standard error of the coefficient of an independent variable is calculated as follows:

$$\text{Standard error} = \text{SE}/(\text{Std. deviation of the independent variable})$$

$$= \frac{18.37}{\sqrt{(50-100)^2 + (100-100)^2 + (150-100)^2}} = .2598$$

[2] To detect nonconstant variance, calculate the rank-order correlation between the position of the data point and the size of the error at that point. The rank-order correlation is a statistic that measures the degree to which two sets of numbers tend to have the same order, or rank. A relatively high rank-order correlation is evidence of nonconstant variance. For the data in Exhibit S.1, the Spearman rank-order correlation coefficient is .125, a relatively small correlation that indicates little evidence of nonconstant variance. The calculation of rank-order correlation is beyond the scope of this supplement but can be found in many statistics texts.

EXHIBIT S.5
Illustration of Nonconstant Variance



To fix the problem of nonconstant variance, management accountants should transform the dependent variable with the log or square root. If it does not fix the condition, management accountants should be very cautious in interpreting the SE value.

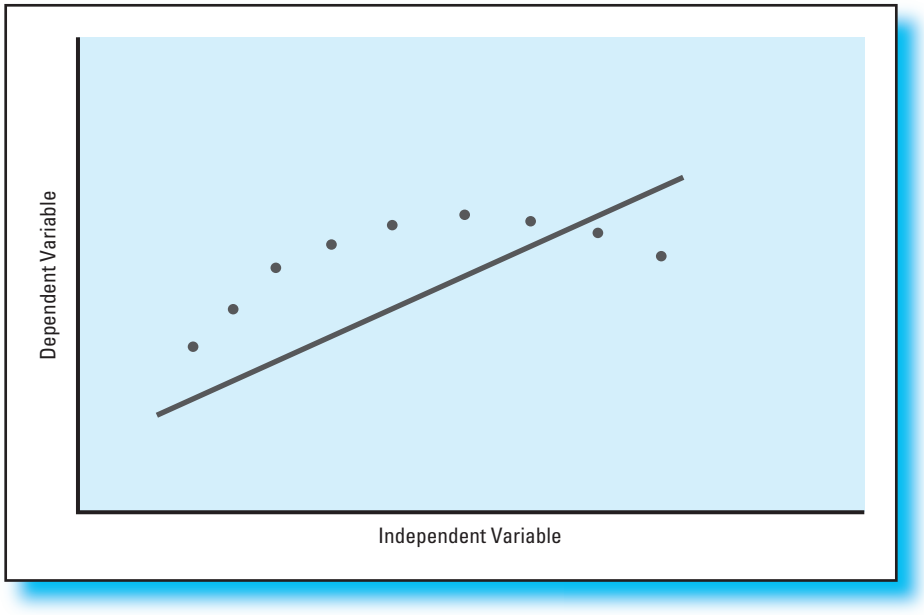
Nonindependent Errors (Durbin-Watson Statistic)

Nonindependence of errors occurs when the amount and direction of each error term is related to those around it. For example, nonindependence of errors is illustrated in Exhibit S.6—the data points are all above the regression line for small values of the independent variable and then below the regression line for large values of the independent value. Nonindependent errors usually occur when there is nonlinearity in the data, as in the illustration for Exhibit S.6. When errors are not independent, the statistical measures are unreliable and the regression predictions are biased.

The **Durbin-Watson (DW) statistic** is a measure of the extent of nonlinearity in the regression.

A common method that detects nonindependent errors is to use the **Durbin-Watson (DW) statistic.** It is calculated from the amount and change of the errors over the range of the independent variable. The DW value falls between zero and 4.0; with 20 or more data

EXHIBIT S.6
Illustration of Nonindependence of Errors



points, a value of DW between approximately 1.0 and 3.0 indicates little chance of a nonlinearity as described earlier; values less than 1.0 or greater than 3.0 should indicate the need to study the data and to choose appropriate fixes if necessary.

The problem of nonindependent errors usually can be fixed by deseasonalizing the data, using a dummy variable for seasonality, or using an index to remove the trend. Alternatively, what may be required is to convert a multiplicative relationship to an equivalent additive (that is, linear) relationship by taking the logarithm of the independent and dependent variables. The statistical measures, their indicators, and ways to fix the underlying conditions are summarized in Exhibit S.7.

EXHIBIT S.7 Summary of Statistical Measures

Measure Concerns	Statistical Measure	What Is an OK Value?*	What Is the Right Fix If Not OK?	Consequence If Not Fixed
Reliability—Goodness of fit	*R*-squared	Should be approximately .75 or better	• Add or delete independent variables • If DW is poor, could need transforms (lag, log, first differences, . . .) • Correct measurement errors in the data, for example, cutoff errors or reporting lags	• Inaccurate estimates
Statistical reliability for the regression	*F*-statistic	Depends on sample size	• Increase sample size • Other changes as suggested for reliability—goodness of fit (see above)	• Inaccurate estimates
Statistical reliability for the independent variables	*t*-value	Should be greater than 2.0	• Delete or transform the independent variable	• Inaccurate estimates
Precision of the regression	Standard error of the estimates (SE)	Should be small relative to the dependent variable	• Same considerations as for reliability—goodness of fit (see above)	• Inaccurate estimates
Reliability of precision (nonconstant variance)	Rank-order correlation	Should be small	• Square root or log transform the dependent variable • Add a dummy variable	• SE is unreliable
Reliability—Potential nonlinearity (nonindependence of errors)	Durbin-Watson statistic (DW)	Between 1.0 and 3.0	*For certain series:* • Deseasonalize • Add trend variable • Use dummy variable for shift *For nonlinear relationship:* • Log transform • Some other nonlinear transform	• Inaccurate estimates • SE is unreliable

*The values shown here are useful for a wide range of regressions. The exact values for a specific regression depend on a number of factors, including the sample size and the number of independent variables. A recent study of regression analysis applied to 20 different overhead cost accounts showed that most of the *R*-squared values fell between .83 and .93. The values for the standard error of the estimates averaged 12% of the mean of the dependent variable, with most falling between 5% and 20%. See G. R. Cluskey Jr., Mitchell H. Raiborn, and Doan T. Modianos, "Multiple-Cost Flexible Budgets and PC-Based Regression Analysis," *Journal of Cost Management,* July–August 2000, pp. 35–47. Also, see the *Parametric Estimating Handbook,* 4th edition, the International Society of Parametric Analysis, April 2008 (Figure 3.6 on p. 84) for a list of thresholds for key statistical measures (www.ispa-cost.org/ISPA_PEH_4th_ed_Final.pdf).

Key Terms

analysis of variance table, *2*
degrees of freedom, *2*
Durbin-Watson (DW) statistic, *5*
F-statistic, *4*
mean squared variance, *3*
nonconstant variance, *4*