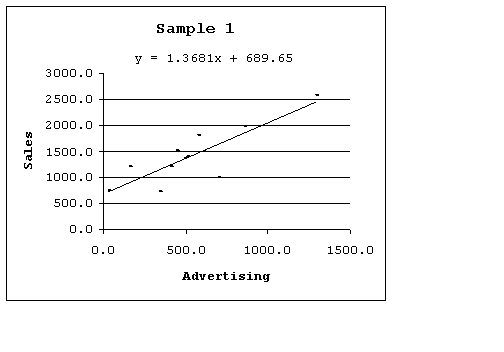
**MULTIPLE REGRESSION**

In a simple regression, the smaller the value of the standard error of the estimate, *se*, the better the predictive power of the model. A smaller *se*means that all of the confidence intervals obtained from the estimated model will be narrower. Remember, *se* is an estimate of the common standard deviation of all of the conditional distributions, namely *σ*. Also, keep in mind that *σ* measures the scatter around the regression line; or, the effect that the factors not included in the model have on the dependent variable. So, one way to improve the model’s predictive power is to reduce *se* by explicitly considering additional factors as independent variables.

Consider the simple regression model of *Sales*, *Y = A + BX + ε*, where the dependent variable, *Y*, is *Sales* and the independent variable, *X*, is *Advertising Expenditure*. If we plot the estimated line (*i.e*., ) against the actual observations (*Y*), we know that the vertical distances of the points from the line (i.e., the errors) are associated with factors omitted from the model.



It’s easy to speculate that another potential determinant of *Sales* might be the *Size of the Sales Force* (*i.e*., the number of sales people employed); a larger sales force is probably associated with higher sales. If that is true, then the points above the estimated simple regression line will tend to be more often associated with “high” sales force levels, and vice versa. This implies that using *Size of the Sales Force* as a second independent variable could improve the model’s fit. Consequently, theoretical model becomes:

*Y = A + B*1*X*1 *+ B*2*X*2+ *ε*,

where *X*1 is the *Advertising Expenditure* and *X*2 is the *Size of the Sales Force*.

Think of it this way: with only one independent variable, *Advertising Expenditure*, *X*1,the vertical distances of the points from the line are interpreted as “unexplained” variations or as errors, however, adding the *Size of the Sales Force*, *X*2, will now attribute part of these unexplained differences to a another factor, hopefully reducing the errors. This is the point of a multiple regression.

In general, the true multiple regression model with *k* independent (explanatory variables) has the form:

*Y = A + B*1*X*1 *+ B*2*X*2*. . . ..*+ *BkXk + ε*,

which is estimated from a sample set of observations. The estimated model has the form:

.

You might find it interesting to imagine how a graphical representation of a multiple regression would look. Obviously, a simple regression can be represented by a straight line drawn on a two dimensional surface, such as a sheet of paper or a blackboard. A model with two independent variables would appear as a linear plane in a three dimensional space; this would appear, with the aid of a computer simulation, as a floating table top. However, models with three or more independent variables are impossible to convey graphically because our universe has only three dimensions (excluding time) and these models essentially are “hyper planes” in four or higher dimensions.

As with a simple regression, a multiple regression should be thought of in terms of conditional distributions. In our example, we are concerned with the distribution of *Sales* for any given and fixed *X*1= advertising expenditure *and* any given and fixed *X2* = *Size of the Sales Force*. The assumptions that we made regarding simple regressions also apply to a multiple regression:

1) Each conditional distribution is normal. For example*,* the sales revenue in all cases when the advertising expenditure has been $500 and the company has employed three sales people is normally distributed. This is the *normality assumption*.

2) The variance of the dependent variable does not depend on the values of the independent variables. In other words, the variance of the conditional distribution of *Sales*, say with $500 in advertising expenditure and three sales people, is the same as the variance of the conditional distribution of *Sales* when the advertising expenditure is $600 and there are two sales people. This is the *Homoscedasticity assumption.*

3) The mean of the conditional distributions are linearly related to the independent variables *i.e*., *μY = A + B1X1 + B2X2. . . ..*+ *BkXk*.

4) Error terms are independent; meaning a large error at some observation has no bearing on the magnitude of the error at some other observation.

5) The independent variables are assumed to be deterministic in the sense that their values are assumed to be known with certainty.

The estimation of the regression coefficients *A, B*1*, B*2*, . . . , Bk*of a multiple regression by the least squares method is based on the same principle applied to a simple regression. The estimators *a, b*1*, b*2*, . . . , bk* are chosen in a manner that minimizes the sum of squared differences between the observed dependent variable (observed *Ys* in the sample) and the estimated *Ys* from the regression as . That is,  will be minimized. Of course, for the most part, we will rely on Excel to perform the estimation. We’ll focus our attention on interpreting the output.

**Example:**

Suppose an analyst suspects that the 10*-year Treasury bond* R*ate* can be predicted by the contemporaneous, prevailing overnight *federal funds rate* and the *3-month Treasury bill rate*. To test his hypothesis, he would set up his regression such that *Y =* 10*-year Treasury bond rate (response variable)* with the two independent variables *X*1 *= Federal funds rate* and *X*2 *= 3-month Treasury bill rate*. Consequently, the model to be estimated is *Y = A + B*1*X*1 *+ B*2*X*2 + ** and it will be estimated on the basis of the following sample of 16 annual observations covering the period extending from 1980 to 1995, inclusively.

|  |  |  |  |
| --- | --- | --- | --- |
| Year | *Y* | *X*1 | *X*2 |
| 1980 | 11.43 | 13.35 | 11.39 |
| 1981 | 13.92 | 16.39 | 14.04 |
| 1982 | 13.01 | 12.24 | 10.6 |
| 1983 | 11.1 | 9.09 | 8.62 |
| 1984 | 12.46 | 10.23 | 9.54 |
| 1985 | 10.62 | 8.1 | 7.47 |
| 1986 | 7.67 | 6.8 | 5.97 |
| 1987 | 8.39 | 6.66 | 5.78 |
| 1988 | 8.85 | 7.57 | 6.67 |
| 1989 | 8.49 | 9.21 | 8.11 |
| 1990 | 8.55 | 8.1 | 7.5 |
| 1991 | 7.86 | 5.69 | 5.38 |
| 1992 | 7.01 | 3.52 | 3.43 |
| 1993 | 5.87 | 3.02 | 3 |
| 1994 | 7.69 | 4.21 | 4.25 |
| 1995 | 6.57 | 5.83 | 5.49 |

So, in this case, the analyst has *n* = 16 (observations) and *k* = 2 *(*numberof independent variables). Notice that although there actually are three parameters to estimate (*A, B*1 and *B*2), *k* ignores the intercept.

The following results are obtained using Excel’s regression tool.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.93918668 |  |  |  |  |  |
| R Square | 0.88207162 |  |  |  |  |  |
| Adjusted R Square | 0.8639288 |  |  |  |  |  |
| Standard Error | 0.8965961 |  |  |  |  |  |
| Observations | 16 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 78.16684427 | 39.0834221 | 48.6182013 | 9.2367E-07 |  |
| Residual | 13 | 10.45049948 | 0.80388458 |  |  |  |
| Total | 15 | 88.61734375 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 2.89591344 | 0.818926109 | 3.53623289 | 0.00365145 | 1.12673115 | 4.66509573 |
| X1 | -1.3491821 | 0.775045719 | -1.7407774 | 0.10532142 | -3.02356656 | 0.32520239 |
| X2 | 2.37600263 | 0.937405473 | 2.53465837 | 0.02490471 | 0.35086123 | 4.40114403 |

The estimated model is . The estimated standard deviation of all of the conditional distributions (*i.e*., the standard error, *se*) is .8966.

**Interpreting the Output**

The output is divided into three sections – **A.** the Regression Statistics, which provides an overview of the model’s ability to explain the variation of the dependent variable; **B.** the ANOVA (Analysis of Variance) gives detailed information regarding the separation of the variation of *Y* into the explained and unexplained components; and **C.** the Estimated Model presents the statistical performances of the individual independent variables.

1. **Regression Statistics:**

* *R-Squared: R*2 is the *coefficient of determination* and measures the proportion of variation in 10-year Treasury rates (*Y*) that is explained by the variation in the *Federal funds rate* and the *3-month Treasury bill rate*. As an equation, *R*2 can be written as: (1- SSE/SST), where SSE is the magnitude of the unexplained variation with *n-k*-1 degrees of freedom, and SST is the total variation with *n-* 1 degrees of freedom. Using these definitions we obtain

*R*2 =1 - {},

which can be interpreted as 1 – [the unexplained variation of *Y* divided by the total variation of *Y*]. As shown in the output above, the regression accounts for about 94% of the variation in the 10*-year Treasury bond* rate leaving about 6% unexplained. The unexplained variation is potentially related to omitted factors such as the state of the economy or exchange rates, etc.

* *Multiple R:* The square root of *R*2. It measures the strength of correlation between the 10*-year Treasury bond rate* with the combination of the *Federal Funds rate* and the *3-month Treasury bill rate*.
* *Adjusted R*2: An adjustment is made to the *R*2 statistic based on the number of independent variables used in the regression. It turns out that the more independent variables you use, regardless of their true relationship to the dependent variable, the higher the *R*2. However, each additional variable causes a degree of freedom [i.e., (*n –k* -1)] to be lost. Consequently, an adjustment is made to the *R*2 that reflects the “cost” of additional independent variables in lost degrees of freedom. The adjustment is given by 1 *–* {(*n-*1)*/*(*n-k-*1)}(1*-R*2)*.* Notice a higher *k* causes a greater downward adjustment to *R*2*.*
* *Standard Error*: *se* is the estimate of the common standard deviation,*σ*, which measures the dispersion of the errors around the conditional means.

*se* =.

1. **ANOVA:**

* *Regression:* 
  + *df* degrees of freedom = *k* (the number of independent variables)
  + SSR = the sum of squared errors captured by the regression = 
  + MSR= the mean of the regression’s squared errors = SSR*/df*
* *Residual:*
  + *df* degrees of freedom = *n –* 1 *- k*
  + SSE= the sum of squared errors left unexplained = 
  + MSE= the mean of the squared unexplained errors = SSE*/df*
* *Total:*
  + *df* degrees of freedom = *n –* 1
  + SST*=* the total sum of squared errors in the data = 
  + MST=the mean of the total squared errors = SST*/df*

Note that SST *=* SSR *+* SSE and that *df* (total) = *df* (regression) + *df* (error).

* The *F*-statistic is used to test if the combination of independent variables (*i.e*., the model) is statistically related to the dependent variable. That is, it helps us determine if the model explains the movement of the dependent variable. The test is simply based on the ratio of MSR to MSE.
* *Significance F:* is the *p*-value of the *F* statistic (we’ll cover this more fully later).

1. **Estimated Model**

Each of the independent variables is evaluated, so there will always be *k+1* lines in this part of the output; *k* lines for each of the independent variables and one for the intercept.

* *Coefficients:* The values of the estimated coefficients, in this case, *a (intercept), b*1 and *b*2, respectively.
* *Standard Errors:* A measure of the precision with which we estimate a coefficient. Each of the estimated coefficients is associated with an estimated standard error (*i.e.*, *sa*, sb1, and *sb*2) that will be used for inferences about the model – confidence intervals and hypotheses on the independent variables. As before, you may think of a frequency distribution of the estimates, say *b*1, obtained from all possible random samples of 16 observations. This frequency distribution is called the *sampling distribution* of *b*1 while the standard deviation is called the *standard error* of *b*1.
* *t-stat:* The ratio of the estimated coefficient to its standard error *e.g*., *b*1*/sb*1, and of course, it is used to test the statistical significance of the relation between the specific independent variable and the dependent variable – that is, it tests if the parameter is statistically different from zero. (see below)
* *P-value:* The *p-value* comes in very handy for quickly determining the statistical significance of a coefficient estimate as determined by the *t*-statistic. Even if the true value of the parameter *Bi* is zero, it is possible to get a positive or negative estimate due simply to sampling error. The *p-value* is the probability of observing as low (or as high) a *bi* value the one calculated, even when the true value of the parameter (*Bi*) is zero. For instance, there is 10.53% probability of obtaining -1.349 or a smaller value for *b*1 even if *B*1 were zero.
* Lower and Upper 95%: The lower and upper limits of the confidence interval for estimated coefficient. That is, there is a 95% chance that the (true) *B*2 is between a low of 0.7159 and a high of 4.4011.

**Making Inferences with the Estimated Model:**

After a model is estimated from a sample it can be used to make certain general and specific inferences about the population from which the sample observations came.

Obviously, these inferences will carry a measure of risk, the statistician’s curse, in the form of sampling errors.

1. **Inferences about the Estimated Model as a Whole.**

The question here is simply whether the estimated regression equation has any predictive power. In other words, we want to determine if there is any underlying relation between the dependent variable and o*ne, some* or *all* of the independent variables. This is important because it gets to the heart of the reason we ran the regression in the first place: to obtain better predictions about the response variable. Consequently, it is important to determine if knowledge of any combination of the independent variables in our model adds to our knowledge of the response variable. More formally this issue is addressed with the following hypothesis test:

*H*o: *B*1 *= B*2 *= . . . . =Bk* = 0

*H*1: at least one *Bi* ≠ 0.

The null hypothesis postulates that none of the independent variables has any bearing on the dependent variable, while the alternative asserts that at least one has some impact. Here, the appropriate test statistic is the *F-statistic* (*F* in the ANOVA section of the output) with *k* and *n-k-*1 degrees of freedom.

In the yield curve example, the computed *F*-stat is 48.6182; the critical *F*-value (at .05 significance level) from an *F*- table (not shown in the regression output) with *k* = 2 and *n-k-*1 = 13 degrees of freedom is 3.81. Because the computed *F*-statistic of 48.6182, which far exceeds the critical *F-*value, we reject the null hypothesis and conclude that at least one of the independent variables helps us explain the movement of the dependent variable – the model has explanatory power.

Notice that the same conclusion can be reached by simply looking at the *Significance F* on the output, which is the *p*-value of the *F*-statistic. In this case, a *Significance F* value of 9.2367-7, which is virtually 0, is much smaller than the default, acceptable level of significance of .05. Viewed from another angle, this simply means that if the null hypothesis had been true (*i.e*., there is no underlying relationship), an *F*-value as large as 48.6182 would be extremely unlikely (*i.e.*, there’s only a 9.2367 -7 probability of this occurring). Ultimately, the *F-test* allows us to conclude that either the *Federal Funds Rate* or *3-month Treasury bill* rate or both have statistically significant explanatory power with regard to the 10*-year Treasury bond* rate.

1. **Inferences about the Individual Regression Parameters, *Bi***.

Here we make inferences about the impact of *each* of the independent variables individually on the dependent variable. Because we’re examining the impact of a particular independent variable, say variable *i*, the hypothesis is formulated as:

*H*o: *Bi =* 0

*H*1: *Bi* ≠ 0,

which is a two-tail test, because we are not testing if the relationship is direct (positive) or inverse (negative). If the null is not rejected, there is no relation between the dependent and the *ith* independent variable.

In this case, the appropriate statistic has a *t* distribution so that the test statistic is

**

with (*n-k-1*) degrees of freedom. Notice that because the hypothesized value is *Bi =* 0, the observed *t* = *bi/sbi*,which is provided in the standard Excel output. As usual, if the computed *t*-value is more extreme than the critical *t*-value, we reject the null and conclude that there is statistical evidence to suggest that there is a relation between the *ith* independent variable and the dependent variable. Of course, this means that the non-zero value calculated for *bi* is not likely to be the result of random chance.

In the context of the yield curve example, suppose we wanted to determine if the *Federal Funds Rate* was a reliable predictor of the *10-year Treasury bond rate*, we would test the hypothesis:

*H*o*: B*1 *=* 0

*H*1*: B*1*≠* 0 *(once again, we are not testing if the impact is positive or negative).*

The *t-statistic* shown in the output is -1.7408 which is obtained from *b*1*/*sb1=

-1.3491821/0.775045719 (because *B*1 = 0 under Ho). This simply states that the calculated estimate, *b*1, is roughly 1.7 times its standard error. The question is: how likely is it that *b*1 is such a large negative number, if *B*1 were indeed zero? The critical *t*-value from a *t-*table (not shown) for any level of significance, ** allows us to answer the question. For instance, the *t* value for **.05 (.025 on each side of the distribution for a two-tail test) with *n – k -* 1= 13 degrees of freedom is +/- 2.160. In other words, a *t*-value as large as 2.160 or larger, and as small as -2,160 or smaller, has no more than a .05 probability if *B*1 is, in fact, zero. Because the computed *t-value* (in the output) is not sufficiently extreme, (*i.e*., it does not belong in the top 5% extreme values of the *t* distribution) we cannot reject the null hypothesis at **.05. Consequently, there is not enough evidence, at this level of significance, in the sample to allow us to conclude that there is a relation between the *Federal Funds rate* and the 10*-year Treasury bond rate*.

Once again, the same conclusion can be reached without explicitly using a *t*-table. The output shows that the *p*-value is .1053, which far exceeds .05, conventional level of significance. If, in fact, there was no relation between the 10*-year Treasury bond rate* and the *Federal Funds rate*, the chances of us getting *t*-values as small as -1.7408 is greater than 10%. Think of it this way: if we ran 1,000 regressions with 16 random observations drawn from a population in which *B*1 = 0, we could expect to get a coefficient of *b*1 = -1.349 or lower 105 times – not that unlikely, or not unlikely enough at **.05 to rule out *B*1 = 0.

Now, suppose we believed *a priori*[[1]](#footnote-1) that a high *Federal Funds Rate* cannot reasonably be a sign of a high 10*-year Treasury bond rate* (i.e., *Bi* > 0 is not reasonable), we would then perform a one-tail test. The test now would be formulated as:

*H*o: *B*1 *=* 0

*H*1: *B*1 < 0.

Notice that the negative sign for the computed *b*1 suggests an inverse relation between the 10*-year Treasury bond rate* and the *Federal Funds rate*, but we want to determine if the evidence is strong enough to beat the standard of proof at the .05 level of significance. The critical *t*-value with 13 degrees of freedom (from a table) is now under .10 which is +/- 1.771, and the *p*-value is half of .1053 or .052. Because of the a priori assumption about *B*1, this obviously is an easier “standard of proof,”yet the sample still does not have enough power (although close) to meet the conventional standard; the *t*-value remains less extreme than the critical *t* of +/- 1.771 (-1.74078 versus -1.1771) and the *p*-value remains above .05. Based on either of these comparisons, we cannot reject the null hypothesis.

1. **Prediction of a Specific[[2]](#footnote-2) *Y* given *X1, X2, . . . . ,Xk.***

Because we know the values of the independent variables we can make a point estimate of the dependent variable simply by plugging in values of the independent variables; the estimate of *Y* is denoted as. However, as you can imagine, due to the existence of other factors not in the model as well as sampling errors in the estimation of the regression parameters, a point estimate is not very reliable. Consequently, it is typically more appropriate to create a confidence interval, which is simply an approximation for the set of values in which the actual *Y* value likely resides.

An *approximate* confidence interval for *Y* can be obtained as

,

where α is 1 - (the relevant confidence interval) and *tα* is the corresponding *t*-value with *n-k-1* degrees of freedom. This is approximate because it simply ignores the sampling errors in estimating the parameters (it equates *se* with *sp*).

In the yield curve example, we might want to predict the 10*-year Treasury bond rate* given that the observed *Federal Funds rate* is 9% and the 3-month observed *Treasury bill rate* is 6.67%. So, what is our best estimate of the 10*-year Treasury bond Rate*? From the regression output we can determine that the point estimate is

= 2.8959 - 1.3492 (.09) + 2.376 (.0667) = 6.6012%. An approximate 95% confidence interval can be stated using the *t*-value for α = .05 with 13 degrees of freedom which is 2.160 (from a table). So, the confidence interval extends from a low of (.066012 - 2.160 \* 0.8965 = .046647) to a high of (.066012 + 2.160 \* 0.8965 = .085376). This is an approximate interval because it ignores the sampling errors in the estimation of *A*, *B*1 and *B*2 by *a*, *b*1and *b*2, respectively. More advanced computer packages have the capability of calculating the exact confidence intervals for various combinations of independent variable levels.

1. **The Problem of Multi-Colinearity.**

One problem that may afflict a multiple regression analysis is multi-colinearity. This condition exists when some or all of the independent variables are highly related (co-linear) to one another. In the yield curve example, we would have a problem if the *Federal Funds rate* was very highly correlated with the *3-month Treasury bill rate*. This would be a problem because the two variables together really behave as one variable; the second independent variable does not contribute any additional independent information useful for predicting the 10*-year Treasury bond rate* – it simply duplicates the information embedded in the other variable.

The good news is that multi-colinearity does not cause the regression to become totally useless, but rather makes the interpretation of the results less straightforward. When multi-colinearity exists:

* The regression will still predict okay if the *F*-statistic indicates that the regression is statistically significant. The model may determine that the independent variables collectively explain the variability in *Y*, but may not be able to decide which independent variable(s) is (are) actually contributing to the predictive power. This leads to the next point:
* The *t*-statistics for the highly correlated variables may turn out to be insignificant even though the regression as a whole is significant.
* The estimated (*bk*) coefficients may turn out to have the wrong (a priori) sign.
* The model often can be improved by simply determining the highly correlated variables and dropping one from the regression.

1. *A priori* is a term that refers to a belief that exists prior to and independent of experience. Contrast this with *a posteriori*, which is based upon actual observation or upon experimental data*:* an a posteriori argument that derives the theory from the evidence. –Dictionary.com. [↑](#footnote-ref-1)
2. **We will not provide an expression for the confidence interval for the mean of the conditional distributions, because the standard error, *s* is difficult to compute for a multiple regression. [↑](#footnote-ref-2)