Module 5: Multiple Regression Analysis

Tom Ilvento,
University of Delaware, College of Agriculture and Natural Resources,
Food and Resource Economics

In the last module we looked at the regression model with a single independent variable helping to explain a single dependent variable. We called this a bi-variate or simple regression because there is only a single independent variable. While this is a good place to start, the world is often more complex than a single explanation or driver for our model. Multiple regression allows us to have many independent variables in a model and examine how each one uniquely helps to explain or predict a single dependent variable.

This module will introduce multiple as an extension of bi-variate regression. The output of the regression model will be very similar, except there will be several coefficients to examine. In addition, the interpretation of the regression coefficients will change in multiple regression. Now we will look at the effect of an independent variable on a dependent variable while controlling for the other independent variables in the model. The notion of statistical control is a very important feature in regression and it makes it a powerful tool when used properly.

BASICS OF MULTIPLE REGRESSION

In review, we said that regression fits a linear function to a set of data. It requires that we have a single dependent variable that we are trying to model, explain, or understand. The dependent variable must be a quantitative variable (preferably measured on a continuous level). Regression estimates a set of coefficients that represent the effect of a single variable or a set of independent variables on the dependent variable. The independent variables can be measured on a qualitative level (as in categorical variables represented by dummy variables), an ordinal level, or at a continuous level.

Key Objectives

- Understand how the regression model changes with the introduction of several independent variables, including how to interpret the coefficients
- Understand the assumptions underlying regression
- Understand how to use regression to represent a nonlinear function or relationship
- Understand how dummy variables are interpreted in multiple regression

In this Module We Will:

- Run regression using Excel with many independent variables
- Look at and interpret the assumptions in regression
- Estimate a nonlinear function using regression and a trend analysis with seasons represented by dummy variables

In multiple regression we have more than one independent variable in the model. This allows us to fit a more sophisticated model with several variables that help explain a dependent variable. For example, catalog sales may be a function of more than just a person's salary. Factors such as a person's age, whether they own their home, or the number of catalogs sent to the person may also help explain sales.

Multiple regression allows us to include additional variables in the model and estimate their effects on the dependent variable as well. In multiple regression we still estimate a linear equation which can be used for prediction. For a case with three independent variables we estimate:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3}$$

Once we add additional variables into the regression model, several things change, while much remains the same. First let's look at what remains the same. The output of regression will look relatively the same. We will generate a R Square for our model that will be interpreted as the proportion of the variability in Y accounted for by all the independent variables in the model. The model will include a single measure of the standard error which reflects an assumption of constant variance of Y for all levels of the independent variables in the model.

The ANOVA table will look similar in the multiple regression. The Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

will be decomposed into a part explained by our model (Sum of Squares Regression) and a part unexplained (Sum of Squares Error or Residual). These will be divided by their respective degrees of freedom. The SSR will be divided by its degrees of freedom, which is the number of independent variables in the model (denoted as k). Once we divide the sum of squares by the degrees of freedom they are noted as Mean Squared Regrerssion (MSR) and Mean Squared Error (MSE).

The F-test will still be the ratio of the two variances - MSR divided by MSE - but the interpretation of the statistical test follows a null hypothesis that none of the variables in the model are significantly different from zero. In other words, an F-test with a significance level less than .05 indicates that at least one of the variables in the model helps to explain the dependent variable.

Much of the output in multiple regression remains the same - R Square; the ANOVA Table with the decomposition of the sums of squares and the F-test; and the table of the coefficients.

The F-test in multiple regression tests to see if at least one of the independent variables significantly contributes to helping to understand the dependent variable.

The remainder of the output will also look similar, except their will be an estimated regression coefficient (slope coefficient) for each independent variable in the model, along with a standard error, t-test, p-value, and confidence interval.

The most important change in the multiple regression is that the coefficients for the independent variables will have a different interpretation. This change reflects that the multivariate relationship of an independent variable with the dependent variable won't necessarily be the same as the bivariate relationship. If we regress Y on X_1 and generate an estimated slope coefficient, b_1, we interpret this coefficient as the effect of X_1 on Y. However, if there is another independent variable in the model, the coefficient is likely to change (represented by b_1*). The reason for this change is that in multiple regression, coefficients are estimated holding constant at the other variables in the model.

The slope coefficient for X_1 is now the change in Y for a unit change in X_1, holding all other independent variables in the model constant. We take into account the other independent variables when estimating the impact of X_1 by incorporating the covariance of X_1 with all the other independent variables in the model. The way that multiple regression solves for the coefficients involves matrix algebra and simultaneous equation solving, topics beyond the scope of this module. Excel and other software packages will do this for us and we need not worry about the actual equations.

We do need to have a handle on why the coefficients are different, and their interpretation. The reason that the coefficients will change is that often the independent variables are correlated with each other, and this correlation may also be shared with the dependent variable. For example, one's age and salary are often related, and they both are related to catalog sales. Multiple regression takes into account the covariance of the independent variables when estimating the slope coefficients, attempting to estimate the unique effect of each variable, independent of the other variables in the model.

Compared to the bivariate regression, controlling for the other independent variables may:

- Increase the strength of the relationship between an independent variable (X_1) and the dependent variable (Y)
- Decrease the strength of the relationship
- Reverse the sign (e.g., from positive to negative)
- Leave it relatively unchanged

The biggest change with multiple regression is that we estimate several slope coefficients at the same time.

The interpretation of these coefficients is the effect of a unit change in the independent variable on the dependent variable, holding constant all other independent variables in the model.

Controlling for other variables in the model will likely change the bivariate relationship between an independent variables and the dependent variable - the relationship could be strengthened, weakened, change sign, or remain relatively unchanged in the multivariate model.

In the special case where X_1 is uncorrelated with the other independent variables in the model (i.e., it is independent of the other Xs in the model), the bivariate regression estimate of the b_1 will equal the multivariate regression estimate of b_1*. Otherwise we should expect some change in the multiple regression estimate.

The ability to estimate the affect of an independent variable (X_1) independent of the other independent variables in the model is a very powerful and compelling feature of regression. It allows us to use "statistical control" as opposed to control via an experimental design. If we cannot use a design to random assign subjects to different treatments, the next best alternative is to estimate effects while controlling for other variables in the model.

The ability to estimate the affect of an independent variable (X_1) independent of the other independent variables in the model is a very powerful and compelling feature of regression. It allows for "statistical control."

For example, if we are interested in understanding the relationship of education on income, we can't random assign different levels of education to men and women; different age groups; or different racial groups, and then observe the resulting income. We have to observe education and income as it is distributed in the sample. Regression allows us to estimate the effect of education while controlling for other factors such as gender, age, or race. This allows us to make a better estimate of the effect of education devoid of the other independent variables in the model. Hence it's popularity in the business applications, the social sciences, medicine, and nutrition.

However, statistical control in regression comes at a price. We can never be certain that we are controlling for all the relevant variables in the model - we may be leaving something out that is unmeasured or not included in the data collection. Leaving out important variables from the model may is called a specification error and may lead to biased results.

Regression analysis does come with certain requirements and assumptions in order to effectively run the models and to make statistical inferences.

However, regression analysis is fairly robust - small departures from some of the assumptions do not lead to serious consequences.

If there is high correlation between X_1 and the other independent variables we may have a problem estimating our coefficients . This is called Collinearity - when X_1 highly correlated with one other independent variable - or Multi-Collinearity - when X_1 is highly correlated with a set of independent variables. If there is too much collinearity, it means we can't estimate the affect of X1 very well, and our estimate will be unstable and poor. Extreme collinearity means the regression can't be estimated at all! More will be discussed about this under the assumptions of the regression model.

ASSUMPTIONS UNDERLYING REGRESSION

Regression, like most statistical techniques, has a set of underlying assumptions that are expected to be in place if we are to have confidence in estimating a model. Some of the assumptions are required to make an estimate, even if the only goal is to describe a set of data. Other assumptions are required if we are going to make an inference from a sample to a population.

Requirements of Regression

- Y is measured as a continuous level variable – not a dichotomy or ordinal measurement
- The independent variables can be continuous, dichotomies, or ordinal
- The independent variables are not highly correlated with each other
- The number of independent variables is 1 less than the sample size, n (preferably n is far greater than the number of independent variables)
- We have the same number of observations for each variable – any missing values for any variable in the regression removes that case from the analysis

Assumptions About the Error Term

Assumptions about the error term in regression are very important for statistical inference - making statements from a sample to a larger population.

We noted in Module 4 that the error term in regression provides and estimate of the standard error of the model and helps in making inferences, including testing the regression coefficients. To properly use regression analysis there are a number of criteria about the error term in the model that we must be able to reasonably assume are true. If we cannot believe these assumptions are reasonable in our model, the results may be biased or no longer have minimum variance. The following are some of the assumptions about the error term in the regression model.

- **The mean of the Probability Distribution of the Error term is zero.**

  This is true by design of our estimator of Least Squares, but it also reflects the notion that we don't expect the error terms to be mostly positive or negative (over or underestimate the regression line), but centered around the regression line.

- **The Probability Distribution of Error Has Constant Variance = σ^2.**

  This implies that we assume a constant variance for Y across all the levels of the independent variables. This is called homoscedasticity and it enables us to pool information from all the data to make a single estimate of the variance. Data that does not show constant error variance is called heteroscedasticity and must be corrected by a data transformation or *Weighted Least Squares*.

- **The Probability Distribution of the Error term is distributed as a normal distribution.**

  This assumption follows statistical theory of the sampling distribution of the regression coefficient and is a reasonable assumption as our sample size gets larger and larger. This enables us to make an inference from a sample to a population, much like we did for the mean.

- **The errors terms are Independent of each other and with the independent variables in the model.**

  This means the error terms are uncorrelated with each other or with any of the independent variables in the model. Correlated error terms sometimes occur in time series data and is known as auto-correlation. If there is correlation among the error terms of with error terms and the independent variables it usually implies that our model is mis-specified. Another way to view this problem is that there is still pattern left to explain in the data by including a lagged variable in time series, or a nonlinear form in the case of correlation with an independent variable.

A MULTIPLE REGRESSION EXAMPLE

The following example uses a data set for apartment building sales in a city in Minnesota. The value of the apartment building (PRICE) is seen as a function of:

1. The number of apartments in the building (NUMBER)

2. The age of the apartment building (AGE)

3. The lot size that the building is on (LOTSIZE)

4. The number of parking spaces (PARK)

5. The total area is square footage (AREA)

The local real estate commission collected a random sample of 25 apartment buildings to estimate a model of value. The sample size is relatively small, but we will still be able to estimate a multiple regression with five independent variables.

The descriptive statistics for the variables in the model are given in Table 1. The average sale price of the apartments was $290,574, but there is considerable variation in the value of apartments. The coefficient of variation is 73% indicating substantial variation. We want to see if the high variability in PRICE is a function of the independent variables. The other variables also show a lot of variability, and in most cases the mean is larger than the median indicating outliers and skew to the data. The exception is AGE, where there is one low value pulling the mean below the median.

Table 1. Descriptive Statistics for the MN Apartment Sales Data

	PRICE	*NUMBER*	*AGE*	*LOTSIZE*	*PARK*	*AREA*
Mean	290573.52	12.16	52.92	8554.12	2.52	11423.40
Standard Error	42305.83	2.52	5.18	839.86	0.99	2003.87
Median	268000.00	8.00	62.00	7425.00	0.00	7881.00
Mode	#N/A	4.00	82.00	#N/A	0.00	#N/A
Standard Deviation	211529.15	12.58	25.89	4199.30	4.93	10019.35
Sample Variance	44744581138.09	158.31	670.49	17634110.11	24.34	100387322.33
Kurtosis	2.80	10.04	-1.40	2.33	6.28	2.19
Skewness	1.61	2.84	-0.48	1.52	2.44	1.71
Range	870700.00	58.00	72.00	16635.00	20.00	36408.00
Minimum	79300.00	4.00	10.00	4365.00	0.00	3040.00
Maximum	950000.00	62.00	82.00	21000.00	20.00	39448.00
Sum	7264338	304	1323	213853	63	285585
Count	25	25	25	25	25	25

Table 2. The Correlation Matrix for the MN Apartment Sales Data

	PRICE	*NUMBER*	*AGE*	*LOTSIZE*	*PARK*	*AREA*
PRICE	1.000					
NUMBER	0.923	1.000				
AGE	-0.114	-0.014	1.000			
LOTSIZE	0.742	0.800	-0.191	1.000		
PARK	0.225	0.224	-0.363	0.167	1.000	
AREA	0.968	0.878	0.027	0.673	0.089	1.000

The Correlation Matrix in Table 2 provides some insights into which of the independent variables are related to PRICE. The highest correlation with PRICE is for AREA (.968) and NUMBER (.923). Both correlations are very high and positive - as the total area increases, or the number of apartments increase, so does the price. LOTSIZE is also positively correlated with PRICE at .743. On the other hand, the number of parking spaces is only positively correlated at .225 and the age of the building, while negative as might be expected, is only correlated at -.114.

We will focus on two bivariate relationships before estimating a multiple regression model SALES and AREA and SALES and AGE. We will look at the scatter plot for each one and the bivariate regression. Figure 1 shows the scatter plot for PRICE and AREA. The relationship is strong, positive and linear. The plot also shows the regression line and R Square for a bi-variate regression. The model says that every square foot of space is roughly worth $20 toward the price of the apartment building. The fit of the model, with only one variable, is very good with 94 percent of the variability of PRICE explained by knowing the area of the apartments. Clearly, area is an important variable to understand the price of the apartment building.

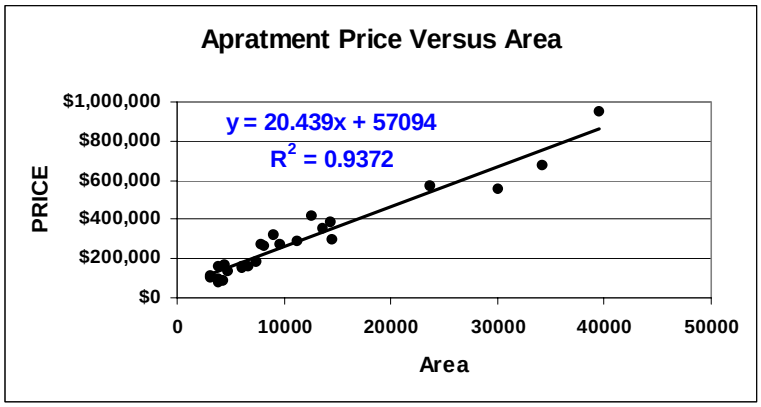

Figure 1. Scatter Plot of PRICE Versus AGE for MN Apartments Data

A useful strategy in any analysis is to start simple - i.e., bi-variate relationships - and them move to more sophisticated multi-variate models.

This helps you to see how relationships can change once we control for other variables in the model.

Let's look at one other bi-variate relationship - between PRICE and AGE. The correlation is -.114, indicating that as the age of building increases, the price decreases. However, the relationship is very weak. The scatter plot in Figure 2 shows that the relationship is weak.

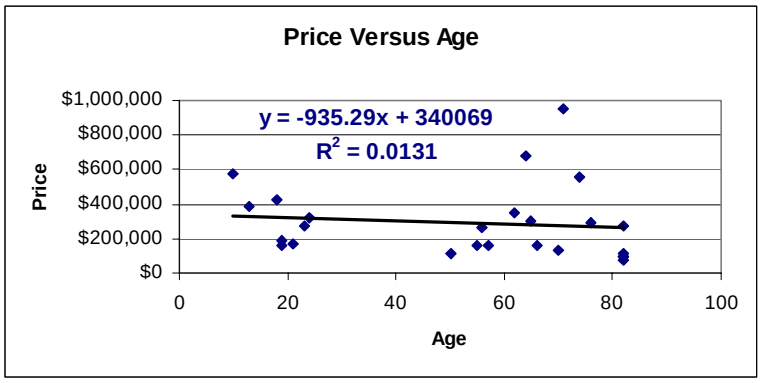

Figure 2. Scatter Plot of PRICE Versus AGE for MN Apartments Data

If we look at the hypothesis test for the coefficient for AGE (full Excel output not shown), we find the following information about the coefficient, standard error, t-test, and p-value for the regression of PRICE on AGE.

	Coeff.	*Std Error*	*t Stat*	*P-value*
Intercept	340068.922	99308.865	3.424	0.002
AGE	-935.287	1692.169	-0.553	0.586

The model estimates that for every year old the building is, the price *decreases* by $935. However, the t-statistic is only -,553 with a p-value of .586. While the model shows a negative relationship, based on a hypothesis test for a sample, I can't be sure if the real value is any different from zero. The R Square of .01 confirms that there is little going on in the data. The steps of a formal hypothesis test are given below.

Null Hypothesis: H_0: $\beta_{AGE} = 0$

Alternative: H_a: $\beta_{AGE} \neq 0$ two-tailed test

Test Statistic: t* = (-935.29 - 0)/1692.17 = -.55

p-Value: p = .59

Conclusion: Cannot Reject H_0: $\beta_{AGE} = 0$

In most cases we don't do the formal steps in a hypothesis test for a regression coefficient - we simply look for t-value greater than 2, or a p-value less than .05 in order to conclude the coefficient is different from zero.

Now let's shift to a multiple regression and see what changes (see Table 3). We will include five independent variables in the model - NUMBER, AGE, LOTSIZE, PARK, and AREA. The Excel output is given below. The R Square for te model is now .98 and the Adjusted R Square is close, .975. One way to tell if the fit has really improved over a simpler model (say for a model with only AREA) is to see what percent of the variability yet to be explained is accounted for by the new model. The model for AREA alone had an R Square of .9372. The new model has a R Square of .98. The increase is only .0428, but this is 68 percent of the remaining .0628 left to explain (1 - .9372). This is a significant improvement in the fit.

The overall F-test for the model (based on a Null Hypothesis that all the coefficients for the independent variables are zero) is 190.84 with a p-value of .00. This means that something useful is going on in our model and at least one of the independent variables has a coefficient different from zero. We can now turn to the individual coefficients for our variables to see the direction of the relationship (the sign of the coefficient), the magnitude of the relationship (the size of the coefficient) and to note which ones show a significant difference from zero (the p-value).

Table 3. Regression Output of PRICE on Five Independent Variables

Regression Statistics	
Multiple R	0.990
R Square	0.980
Adjusted R Square	0.975
Standard Error	33217.938
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	5	1052904750916.02	210580950183.21	190.84	0.00
Residual	19	20965196398.22	1103431389.38		
Total	24	1073869947314.24			

	Coef	*Std Error*	*t Stat*	*P-value*
Intercept	92787.87	28688.252	3.234	0.004
NUMBER	4140.42	1489.789	2.779	0.012
AGE	-853.18	298.766	-2.856	0.010
LOTSIZE	0.96	2.869	0.335	0.741
PARKING	2695.87	1576.742	1.710	0.104
AREA	15.54	1.462	10.631	0.000

As might be expected, only AGE shows a negative relationship with PRICE - as the apartment building gets older the PRICE decreases. The other signs are positive as might be expected; more of each of the independent variables should increase the price. If I focus on the p-values and use a criterion of .05 as a significance level, I can see that the coefficients for LOTSIZE and PARK are not significantly different from zero. In other words, I do not have information from this sample to safely conclude that the coefficients for these variables are any different from zero. I should note that this test is based on *holding constant the other independent variables in the model*.

The predictive equation from our model is

$$\hat{Y} = \$92{,}788 + \$4{,}140(\text{NUMBER}) - \$853(\text{AGE}) + \$1(\text{LOTSIZE}) + \$2{,}696(\text{PARK}) + \$16(\text{AREA})$$

Solving the regression equation for Y given a set of values for the independent variables allows us to make a prediction from our model.

We want to find the predicted value of an apartment building with 20 apartments; 50 years old; 2,000 sq ft of lot size; 20 parking spaces; and 22,000 sq ft of area, our model says the value is:

Predicted Value = \$92,788 + \$4,140(20) - \$853(50) + \$1(2000) +\$2,696(20) + \$16(22,000)

= \$540,858

Next, let's focus on the coefficients for AREA and AGE and compare them to the bi-variate relationship. The coefficient for AGE in the multiple regression model is -853 compared with -935 for the simple regression. However, now our model says that we have evidence that the coefficient for AGE is significantly different from zero (p-value = .01). Once we controlled for the other variables in the model, we can safely say that the age of the apartment has an influence on its value, and that this relationship is negative. Controlling for other variables in the model enabled us to make a better estimate of the effect of age on the value of an apartment.

The coefficient for AREA has also decreased in the multiple regression. Once we control for the other variables in the model (NUMBER, AGE, LOTSIZE, and PARK), a unit change in AREA results in a 15.4 unit change in price (down from 20.4 in the bi-variate regression). We still have strong evidence that AREA is positively related to PRICE, but now our model says the effect is smaller once we control for the other variables in the model.

REPRESENTING NONLINEAR RELATIONSHIPS WITH REGRESSION

We noted earlier that regression is used to estimate linear relationships and that the regression model is linear in the parameters. However, it can be used to represent nonlinear relationships in the data. There are many nonlinear functions that can be estimated with regression. The following are but two examples, using a polynomial function and a log function.

We can represent nonlinear relationships with regression under certain circumstances.

Polynomial Regression. A polynomial expression is an equation that is linear in the parameters that has an independent variable that is raised to a power included in the model. The equation takes the following general form (where p reflects the order).

$$Y = a + b_1 X + b_2 X^2 + b_3 X^3 + \ldots + b_p X^p$$

A second order model has X and X^2; a third order model has X, X^2 and X^3, and so forth. Figures 3 and 4 show a second and third order polynomial, respectively. The second order relationship shows a single inflection point where the relationship changes. In the example in Figure 3, the function increases to a point and then the negative squared term begins to turn the relationship downward. Using calculus, we can solve for the inflection point - in this case it is at 28.6. In figure 4 we have a third order polynomial so there are two inflection points.

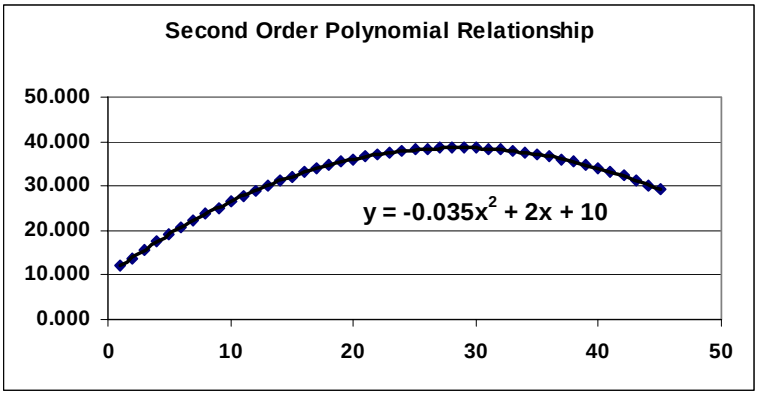

Figure 3. Second Order Polynomial Relationship

All lower order terms must be in the equation in order for the polynomial function to be represented, but the statistical test should focus on the highest order term in the model. In essence, the test is whether the highest term in the model is significantly different from zero. This tells us if the term is needed in the model, and thus whether the polynomials fits the data.

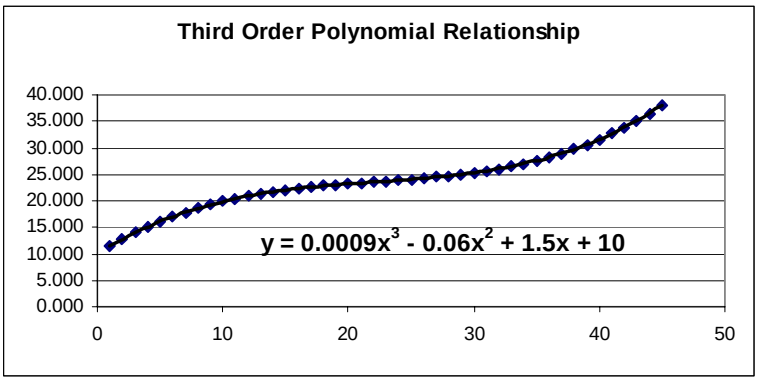

Figure 4. Third Order Polynomial Relationship

A polynomial relationship with one of the independent variables in the model also can be fitted while including additional variables in the model. For example, The relationship between age and income can be thought of as a polynomial - as age increases income does as well, but at some point the relationship diminishes and may even turn down as people reach retirement age. We could fit a model with AGE and AGE^2 while also including other variables in the model that influence income, such as gender and education.

Let's look at an example of a polynomial relationship. The following example is from sales and investments in U.S. manufacturing from 1954 to 1999. The data values are in millions of dollars. The graph of the relationship between investment and sales is given in Figure 5. The relationship looks linear, and the fitted line and R Square show a very good fit (R^2 = .976). However, there does appear to be a slight curve to the relationship. If you look closely, the points on the line at the end tend to be above the regression line while the points in the middle tend to be below it. This is an indication of a nonlinear relationship.

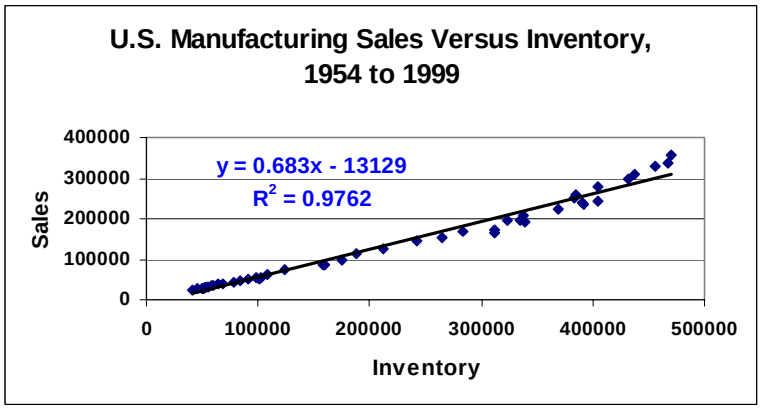

Figure 5. Scatter Plot of Sales Versus Inventory with a Linear Trendline

One way to see if there is a curve (besides the scatter plot) is to plot the residuals from the linear regression line. The best plot is the standardized residuals (with a mean of zero and a standard deviation of 1) against the Predicted values of Y. This plot should look like a random scatter of points and show no tendencies or patter to the plot. Figure 6 shows the residual plot (Excel can help create this), and there does appear to be a pattern to the plot. Values in the middle tend to have negative residuals and values at the top tend to have positive residuals. This is a sign of a nonlinear relationship.

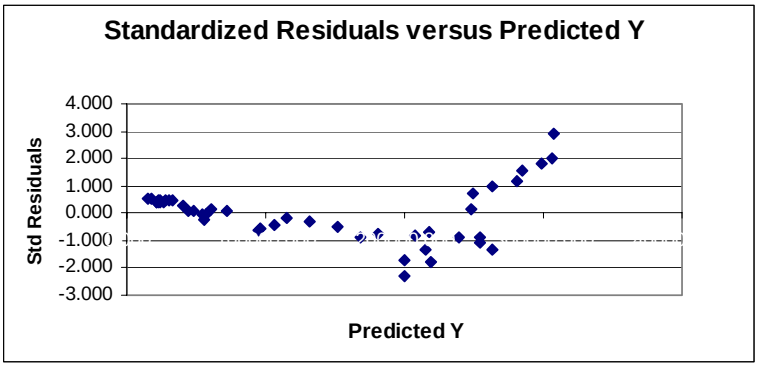

Figure 6. Scatter Plot of Standardized Residuals and Predicted Y Showing a Non-Random Pattern

To test to see if there is a nonlinear relationship, I calculated a squared inventory variable and included it in the model along with inventory. The regression output is given in Table 4. We will focus on two key things - R Square and the significance test for the squared inventory term. The R Square for the model increase from .976 to .991, a near perfect fit. This represents 63 percent of what was left to explain from the first order model, a significant increase.

The coefficient for INVENT SQ is very small, but this is not unusual for a squared term in the model. We are squaring very large numbers and the coefficient reflects this. The key issue is whether the coefficient for INVENT SQ is significantly different from zero. This requires a t-test or a check on the p-value. The formal hypothesis test is:

Null Hypothesis: H_0: $\beta_{ISQ} = 0$

Alternative: H_a: $\beta_{ISQ} \neq 0$ two-tailed test

Test Statistic: t* = 8.527

p-Value: p = .000 or p < .001

Conclusion: Reject H_0: $\beta_{ISQ} = 0$

We can determine if a higher order polynomial relationship exists by:

- examining a scatter plot to see if we can observe a curve
- looking for a pattern in a plot of the standardized residual versus the predicted Y
- running a regression model with the higher order term and testing to see if it is significantly different from zero.

Table 4. Regression Output for Polynomial Regression of SALES on INVENT and INVENT SQ

Regression Statistics	
Multiple R	0.996
R Square	0.991
Adjusted R Square	0.991
Standard Error	9954.802
Observations	46

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	478111171319.26	239055585659.63	2412.31	0.00
Residual	43	4261217706.85	99098086.21		
Total	45	482372389026.11			

	Coeff.	*Std Error*	*t Stat*	*P-value*
Intercept	17109.815	4408.482	3.881	0.000
INVENT	0.264	0.050	5.282	0.000
INVENT SQ	8.78566E-07	0.000	8.527	0.000

Based on our test we can conclude that the INVENT SQ term is significantly different from zero and thus is important in the model. This validates our use of the second order polynomial model to improve our fit, and the result is in agreement with the increase in R Square. One final point should be emphasized. Once we include a higher order term in the model and find it significant, we rarely focus on the test for the lower order term(s). The lower order terms must be included in the model to make sense of the polynomial relationship.

Logarithm Transformation - The Multiplicative or Constant Elasticity Model. One transformation of a multiplicative or constant elasticity model is the use of a log transformation. The multiplicative model is in the form:

$$Y = aX_1^{b1} X_2^{b2} \ldots X_k^{bk}$$

Taking logs of the variables can transform a nonlinear relationship in one that is linear in the parameters - but the data are transformed and are now in log units.

Taking the natural log of each variable of this type of a function results in a linear transformation of the form:

$$Y = \ln(a) + b_1 \ln(X_1) + b_2 \ln(X_2) + \ldots + b_k \ln(X_k)$$

This is also called the constant elasticity model because we interpret the coefficients as the percentage change in Y for a 1 percent change in X. The definition of an elasticity is the percentage change in Y with a one percent change in an explanatory variable, denoted as X. In most cases

the elasticity is dependent on the level of X and will change over the range of X. However, in this type of model the elasticity is constant and does not change over the range of X - it is constant at the level of the estimated coefficient. You would only use this model if you believe the elasticity is a constant.

We will look at one example of a multiplicative model. The data for this example is operating costs at 10 branches of an insurance company (COSTS) as a function of the number of policies for home insurance (HOME) and automobile insurance (AUTO). We start with the belief that operating costs increase by a constant percentage in relation to percentage increases in the respective policies.

This type of model requires us to transform all the variables by taking the natural logarithm. This is the function =LN(value) in Excel. In order to take a log of a value it must be greater than zero (it can't be zero or negative). These constraints are not issues in this data. Once I convert the data, I use the new variables in a regression using Excel. The output is given in Table 5.

The overall fit of the model is very good with an R Square of .998. However, we only have 10 observations in the data and R Square is sensitive to the number of observations relative to the number of independent variables. The overall F-test is highly significant, as are both of the independent variables in the model. All the p-values are < .001, so we can safely conclude that the

Table 5. Regression Output for Multiplicative Model for Insurance Costs

Regression Statistics	
Multiple R	0.999
R Square	0.998
Adjusted R Square	0.997
Standard Error	0.023
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	2	1.910	0.955	1736.692	0.000
Residual	7	0.004	0.001		
Total	9	1.914			

	Coeff.	*Std Error*	*t Stat*	*P-value*
Intercept	5.339	0.127	41.973	0.000
LNHOME	0.583	0.013	44.683	0.000
LNAUTO	0.409	0.019	21.902	0.000

coefficients are significantly different from 0. The interpretation of the coefficients is now slightly different to reflect the constant elasticity. In this case we say that when Home Owner policies increase by 1 percent, the operating costs increase by .583 percent, and a one percent the percent increase in Auto policies leads to a .409 percent increase in operating costs.

Be careful making model comparisons when we transform the data - the transformed data cannot always be directly compared to the original form of the data.

It is difficult to validate the logarithm transformation for the multiplicative model. While the fit is very good, a similar model of the original data also yielded an excellent fit (R^2 = .997, data not shown). In the end it comes down to whether you believe the assumptions of constant elasticities truly fits your situation. That is as much an issue about your intuition and experience (or more formally about economic theory) than about comparing models.

In fact, once you transform the data, you cannot directly compare the logarithmic model to a model of untransformed data - it becomes a comparison of apples and oranges. However, you can take the anti-log of a predicted value from your model to transform the result back to real terms.

Example of a Multiple Regression with Trend and Dummy Variables. Data over time often show cyclical or seasonal trends. These trends can be modeled by the use of a dummy variable. For example, four quarters can be represented by three dummy variables, coded zero and one to reflect which quarter is being represented. The following is an example of a multiple regression using a linear trend and seasonal dummy variables. Special emphasis in this example will be given to interpreting the dummy variables when another variable (TREND) is in the model.

The example shows quarterly data for Coca-Cola sales (given in millions of dollars) from 1986 through 2001. This example was taken directly from Data Analysis for Managers, written by Albright, Winston, and Zappe (Brooks/Cole, 2004; ISBN: 0-534-38366-1). A plot of the data (Figure 7) shows a strong linear trend, but there are regular fluctuations that show seasonal variations. A closer look shows that sales of Coca-Cola are higher in the second and third quarters when the weather is hotter. The linear trend is represented on the graph and R^2 for the model is relatively high - .88. However, we are going to see if we can improve the model by including seasonal dummy variables.

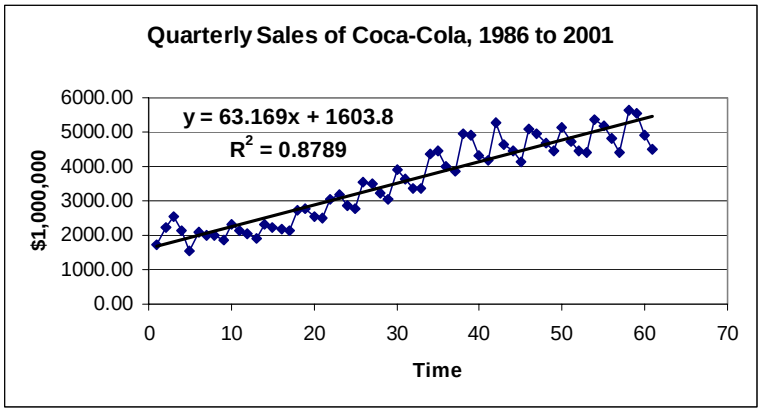

Figure 7. Coca-Cola Sales Showing a Linear Trend and Seasonal Variations

Seasonal variations in data over time can be modeled using dummy variables to represent the quarters.

Since there are four quarters in each year (k=4), we need to represent quarters with three dummy variables (k-1=3). It is somewhat arbitrary which quarter we leave out of the model to serve as the reference category. Since I calculated the dummy variables in order, it was easiest to leave out the fourth quarter sales in the model. As a result, Q4 is the reference category.

The regression analysis is given in Table 6 below. We can note that R^2 for the model increased to .941, reflecting that the inclusion of the quarter dummy variables has increased the fit of the model over the simple trend

Table 6. Regression of Coca-Cola Sales on Trend and Quarters

Regression Statistics	
Multiple R	0.970
R Square	0.941
Adjusted R Square	0.937
Standard Error	299.780
Observations	61

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	4	80821435.3	20205358.8	224.8	0.0000
Residual	56	5032605.1	89867.9		
Total	60	85854040.4			

	Coef	*Std Error*	*t Stat*	*P-value*
Intercept	1435.85	104.24	13.77	0.000
Trend	63.58	2.18	29.14	0.000
Q1	-231.12	107.76	-2.14	0.036
Q2	521.94	109.55	4.76	0.000
Q3	355.48	109.49	3.25	0.002

analysis. If we look at the coefficients we can see that there is still a significant trend in the model (t-stat = 29.14 and $p < .001$).

All of the quarter dummy variables are significant at the .05 level, indicating that we can conclude that sales in Q1, Q2, and Q3 are all significantly different on average from Q4 (the reference category). Specifically, we can see that Q1 has lower sales that Q4 because the coefficient for Q1 is negative (-231.12). This means that sales in Q1 tends to be $231.12 million less than Q4, holding constant the time period. In contrast, sales in Q2 and Q3 are significantly higher than Q4. If we think about it, these results make sense. Sales of soft drink should be related to the time of year with higher on average sales in warmer quarters.

Two keys in assessing the impact of including dummy variables in a multiple regression is whether the coefficients are significantly different from zero, and whether their inclusion in the model increases R^2 in a meaningful way.

As you can see from this model, the interpretation of dummy variables is relatively the same in a multiple regression as in it in simple regression, except the regression estimate of the coefficients takes into account the other independent variables in the model. With dummy variables , we need to interpret the coefficients in relation to the reference category.

The test of the coefficients tell us if the mean of one category is significantly different from the reference category, and the overall F-test or the increase in R^2 tells us if the overall variable (in this case quarters, represented by three dummy variables) is important in the model. In this example, including quarters in the model through dummy variables increased R^2 from .879 to .941. This is an increase to R^2 of .062, which represents over 51 percent of the amount left to explain over the first model (.512 = .062/[1-.879]). Clearly, adding a seasonal component into our model through dummy variables has helped explain an additional source of variability in sales. This will aid in making forecasts into the future.

CONCLUSIONS

Multiple regression involves having more than one independent variable in the model. This allows us to estimate more sophisticated models with many explanatory variables influencing a single dependent variable. Multiple regression also allows us to estimate the relationship of each independent variable to the dependent variable *while controlling for* the effects of the other independent variables in the model. This is a very powerful feature of regression because we can estimate the unique effect of each variable. Multiple regression still uses the property of Least Squares to estimate the regression function, but it does this while taking into account the covariances among the independent variables, through the use of matrix algebra.

Multiple regression requires that the dependent variable is measured on a continuous level, but it has great flexibility for the independent variables. They can be measured as continuous, ordinal, or categorical as represented by dummy variables. There are other requirements in multiple regression - equal number of observations for each variable in the model; limits on the number of independent variables in the model; independent variables cannot be too highly correlated with each other; and assumptions about the error terms. Overall, regression is fairly robust and violations of some of the assumptions about error terms may cause minor problems or they can be managed by transformations of the data or modeling strategies. These issues are advanced topics in regression.