Kellogg School of Management
NORTHWESTERN UNIVERSITY
TECHNICAL NOTE

DAVID DRANOVE **KEL636**

Practical Regression: Regression Basics

This is one in a series of notes entitled "Practical Regression." These notes supplement the theoretical content of most statistics texts with practical advice on solving real world empirical problems through regression analysis.

Modern statistical software makes it easy to perform regression. With data in hand, it is tempting to dive in feet first and start running regressions. Before you get carried away, it is useful to review some basic facts.

What Is a Regression Equation?

The Basics

A regression equation expresses a relationship between one or more *predictor variables* (also known as *X variables*, *regressors*, or *right hand side (RHS) variables*) and a *dependent variable* (also known as the *Y variable*, *left hand side (LHS) variable*, or *outcome*). We usually work with additive, linear equations such as equation (1):

$$Y = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k + \varepsilon \quad (1)$$

where ε is a normally distributed error term with mean 0. In an additive, linear equation, each term on the RHS is added together to get Y.

Suppose the number of observations is *n* and the number of variables (excluding the constant) is *k*. We can rewrite (1) as $Y = X\beta + \varepsilon$, where *y* is an ($n \times 1$) vector, X is an $n \times (k+1)$ matrix, and β and ε are both ($n \times 1$) vectors:

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots x_{k1} \\ 1 & x_{12} & \ldots x_{k2} \\ & & \\ 1 & x_{1n} & \ldots x_{kn} \end{bmatrix} \bullet \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_n \end{bmatrix}$$

Do not worry if you do not know matrix algebra. The above equation is just a shorthand way of saying that if you have n observations, the value of y_i for observation i is obtained from equation (1), using the appropriate values of the x_i variables for that observation and adding the appropriate error ε for that observation.

To figure out the predicted values for Y given values for the X variables, we use this formula:

$$E[Y] = E[\mathbf{X\beta} + \boldsymbol{\varepsilon}] = E[\mathbf{X\beta}] + E[\boldsymbol{\varepsilon}] = \mathbf{X\beta}$$

where $E[\cdot]$ represents the expected value and $E[\varepsilon] = 0$ by assumption. In other words, the predicted value for Y is obtained by summing up the βX's. For this reason, I will often drop the error term in this note, especially when showing you how to interpret regression coefficients.

Reminder: The regression equations you are used to working with are linear equations.

Through regression analysis, you can take several observations of the Y and X variables and use them to estimate the values of the coefficients in equation (1): β_0, β_1, and so on. The exact formula used to compute the coefficients can be found in any statistics textbook. The purpose of this note is not to teach you such formulae but to teach you how to interpret the resulting coefficients.

Estimating these coefficients will enable you to:

- Predict the value of Y for any set of values of all the X variables (e.g., predict sales if your price is $100 and you spend $10 million on advertising)
- Test hypotheses about the effect of a particular X variable on the value of Y, holding all other variables constant (e.g., test whether a dollar of advertising increases sales by at least a dollar, holding all other factors constant. Note that this is an example of a "single-tailed" hypothesis test.)
- Estimate the effect of changing a particular X variable on the value of Y, holding all other variables constant (e.g., predict the sales increase that would result if you increased advertising expenditures by $1 million, holding everything else fixed)

The most common technique used to estimate coefficients is *ordinary least squares (OLS)* regression, but it is not the only one. We will use OLS extensively in the Practical Regression series, but we will also use other methods for estimating the β's when they are more appropriate.

Manipulating Regression Equations

You can manipulate regression equations exactly like you would a simple algebraic equation. This is because regression equations are algebraic equations. Here is a simple example:

Let S = sales, P = price, and A = advertising expenditures. Suppose that you posit a simple relationship between S, P, and A:

(2) $$S = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \varepsilon$$

With historical data on sales, prices, and advertising, you could use OLS to estimate the coefficients in equation (2). Suppose you do this and obtain $\beta_0 = 5$, $\beta_1 = -0.2$, and $\beta_2 = 0.5$. Simple algebra tells us how to interpret these coefficients.

- Given particular values for P and A, you can predict sales by plugging into the equation $S = 5 – (0.2 \cdot P) + (0.5 \cdot A)$. For example, if price = 5 and advertising = 1, the best estimate of sales is $5 – 1 + 0.5 = 4.5$.
- If price were to increase to 6, sales would be predicted to decrease by 0.2 units (this is the value of the coefficient $\beta 1$).
- If your regression is unbiased, then these are your best estimates of how price and advertising affect sales. This does not necessarily mean that your estimates will be highly accurate, only that if you made similar predictions many times, you would get them right on average. Your confidence in any one prediction may be high or low, depending on the statistical significance of the predictors.

Special Predictor Variables

Equation (2) featured two very simple predictor variables. Both were *continuous* (in principle, they could take on infinitely many values). Neither was *exponentiated* or *interacted* with other variables. Sometimes you will want to use a more sophisticated model. This section covers some types of "special" variables you may want to include in your regression.

CATEGORICAL (DUMMY) PREDICTOR VARIABLES

A categorical variable takes on a finite number of values. The most common categorical variable is dichotomous or binary, which means it takes on only two values. Such variables are often called *dummy variables* or *indicator variables*. Usually a dummy variable equals 1 if some condition is met and 0 if that condition is not met.

For example, suppose that we obtain sales, prices, and advertising levels for 52 weeks. We wonder whether sales are higher in the summer than in the rest of the year, controlling for any seasonal differences in price and advertising. In this case, the condition is "Is it summer?" To determine if satisfying this condition matters, we construct a dummy variable, likely called "summer." We could let summer = 1 for the 13 observations that correspond to summer weeks, and let summer = 0 for the remaining 39 observations. (Note you might want to play around with other definitions of "summer" for robustness.)

We then estimate the following regression equation:

$$S = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \beta_3 \cdot \text{summer} + \varepsilon \quad (3)$$

Equation (3) is just another algebraic expression, so it is easy to interpret the results. In particular, *the coefficient β_3 indicates the amount that sales increase when the value of summer increases by one unit* (i.e., increases from 0 to 1), and everything else remains constant.

Let's try to be a bit clearer. The variable summer takes on one of two values. It equals 1 for the summer weeks and 0 for all other weeks. Thus, the coefficient β_3 tells us how much more is sold in summer when compared with all the other weeks.

Note that if summer = 1, then equation (3) could be rewritten:

(3a) $$S = (\beta_0 + \beta_3) + \beta_1 \cdot P + \beta_2 \cdot A$$

Written this way, we immediately see that the coefficient β_3 simply shifts the intercept of the regression line. (For convenience, we will henceforth drop the ε error term.)

There are many dichotomous categorical variables, such as sex (male/female), age (over 65/under 65), and ownership status of firm (publicly traded/private). You can determine the effect of any of these variables by including an appropriate dummy variable in the regression.

You may already be wondering what happens if your categorical variable can take on more than two values. Seasonality is a good example—we divided the year into two periods (summer/non-summer), but it is more natural to divide the year into four seasons. We do this by creating a dummy variable for each season (e.g., let spring = 1 if the week falls during the spring and let spring = 0 otherwise).

Suppose you have created four dummy variables. You might be tempted to estimate equation (4) using OLS:

(4) $$S = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \beta_3 \cdot \text{summer} + \beta_4 \cdot \text{spring} + \beta_5 \cdot \text{winter} + \beta_6 \cdot \text{fall}$$

For technical reasons that you need not worry about, OLS cannot perform this estimation.[1]

To estimate a regression that includes categorical variables, it is necessary to omit one category from the RHS. For example, equation (3) included the "summer" category but omitted the "all other seasons" category. To estimate the effects of all four seasons, we must estimate an equation like equation (5) but omit one of the seasons:

(5) $$S = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \beta_3 \cdot \text{summer} + \beta_4 \cdot \text{spring} + \beta_5 \cdot \text{winter}$$

Now comes the tricky part. *The coefficients β_3, β_4, and β_5 tell us how much more (or less) is sold during the corresponding season than is sold during the omitted season (in this case, fall).*

If you want to compare two of the included seasons (e.g., summer versus spring), you have two options. First, you can re-estimate the model, including fall but excluding spring. The summer coefficient would tell us how much is sold in summer relative to spring. We need not take this extra step, however. The difference between the summer and spring coefficients, $\beta_3 - \beta_4$, is the best estimate of the differential effect on sales. In fact, this difference will equal the summer coefficient in the re-estimated regression that omits spring.

As we discuss below, you will usually want to determine if a regression coefficient, or the difference between coefficients, is "meaningfully" different from zero (i.e., whether the coefficient is *statistically significant*). Your regression output will report the statistical significance of each individual coefficient, including coefficients on dummy variables. The

[1] The explanation has to do with the way that the computer estimates the intercept in regression. To estimate the intercept, the computer adds a column of 1's to the data and treats this as an additional variable. The "coefficient" on this variable is the intercept. If you include all four seasons in your regression and add the four variables together, the sum will equal one for every observation (because one and only one season will equal 1). This makes the four dummy variables perfectly collinear with the intercept. If you want to include all four seasons in the regression, you must tell the computer to estimate the regression without an intercept.

significance of the dummy variable coefficients tells you whether there is a meaningful difference between each included category and the excluded category (e.g., summer versus fall). In order to assess the significance of the difference between included categories (e.g., summer versus spring), we could re-estimate the equation three more times, omitting each season in turn. This could quickly become cumbersome. (Imagine if we were using dummy variables for states; we would have to run fifty equations.)

But there is an easier way. Let's return to equation (5). Suppose you want to know if the level of sales is different in summer versus spring. $\beta_3 - \beta_4$ is the best estimate of this difference. Stata can easily determine if $\beta_3 - \beta_4$ is statistically significant. In general, after you run your regression, you type **test var1=var2**. In this case:

```
regress sales P A summer spring winter
test summer=spring
```

There is a longer discussion of this issue at the end of this note (in the "Comparing coefficients in general" section).

General Rules for Dealing with Categorical Variables

- When working with a categorical variable, create a dummy for each category.
- If there are *n* categories, use *n*-1 dummies in the regression.
- The coefficient on a particular dummy indicates the average value of the dependent variable for the corresponding category relative to the average value of the dependent variable for the omitted category, all else equal.
- You can compare the effects of two included categories by directly comparing their coefficients. Be sure to use the proper significance test.

EXPONENTS

In a simple linear regression equation, the effect of changing X on Y is the same regardless of the initial value of X; a one-unit increase in X causes Y to increase by β_X. This is called a *linear effect*. You may have reason to believe that the real world situation you are examining is not linear. For example, you might wonder if the effect of additional advertising expenditures depends on the amount already spent. (This is the basis for the "saturation model," in which excessive spending has little to no effect on sales.)

An easy way to test for nonlinear effects is to include an exponential term in the regression:

$$S = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \beta_3 \cdot A^2 \quad (6)$$

By including A^2 in equation (6), we allow the effect of advertising on sales to vary with the level of advertising. We might predict that advertising boosts sales in general but that the benefits of advertising decline as A increases. If this is the case, then we should see $\beta_2 > 0$ (in general, advertising helps) and $\beta_3 < 0$ (advertising helps less as A increases).

Equation (6) is still an algebraic expression, so it is easy to work with the regression results. Suppose we run the regression and obtain $\beta_0 = 8$, $\beta_1 = -0.2$, $\beta_2 = 0.5$, and $\beta_3 = -0.05$. Then if $P = 5$ and $A = 4$, the best estimate of S is $8 - 1 + 2 - 0.8 = 8.2$.

INTERACTIONS

Sometimes the effect of one RHS variable *depends* on the value of another RHS variable. For example, the effect of advertising might depend on the price level—perhaps because advertising is more effective for high-priced products. If you want to determine whether the effect of one predictor depends on the value of another, you need to include an *interaction* variable.

To create an interaction variable, simply multiply the two predictors together:

$$\text{Sales} = \beta_0 + \beta_1 \cdot P + \beta_2 \cdot A + \beta_3 \cdot (A \cdot P) \quad (7)$$

The expression $A \cdot P$ is an *interaction term*. The coefficient β_3 indicates whether the impact of advertising on sales depends on the level of price. (Of course, this interaction is symmetric; it also tells us whether the effect of price on sales depends on the level of advertising.)

Equation (7) can be easily manipulated using simple algebra. Suppose we obtain $\beta_0 = 6$, $\beta_1 = -0.4$, $\beta_2 = 0.4$, $\beta_3 = 0.1$. Then if P = 5 and A = 4, the best estimate of S is 6 – 2 + 1.6 + 2 = 7.6. Later in this note, we will see how to compute the effect of a change in A on the value of S.

Measuring Regression Performance

The R^2

The best-known regression statistic is the R^2. Recall that R^2 is the percentage of the variation in Y (around its mean) that is explained by the predictor variables. Although many students emphasize R^2 when evaluating a model, good empirical researchers often make no more than a passing reference to it. In fact, most top-notch economics researchers will state flat out that they "do not care at all about R^2." To understand why, it is useful to learn (or relearn) the basics about R^2.

A PRIMER ON R^2

Think of regression as a tool for prediction. Suppose we have collected observations for *n* values of Y: (Y_1, Y_2, ..., Y_n). We would like to predict the values of additional realizations: Y_{n+1}, Y_{n+2}, and so on. Suppose at first that we have no idea what affects Y. In this case, the best we can do is come up with what we believe is the most likely value of Y and make this same prediction every time. (After all, there is no reason to vary the prediction.) Call this prediction β_0. Needless to say, the actual outcomes of Y are likely to differ from this prediction. Let e_i equal the prediction error—the difference between the actual Y and the prediction. This gives us equation (8):

$$Y_i = \beta_0 + e_i \quad (8)$$

Thus far, we have said nothing about how we came up with our prediction. A natural instinct is to select β_0 to equal the mean value of Y across all past observations. We use this approach to prediction all the time in everyday life. To predict how many games the Chicago Cubs will win next year, you could do worse than to take the average number of wins from the past few years.

To predict the score your friend will receive on his statistics test, you might use the average of his past tests.

Let's use some numbers. Suppose you own a coffee shop and you want to predict the sales of coffee beans for the upcoming weeks. Suppose, for now, that you have no "model" to help you make this prediction. You do, however, have sales data from the past four weeks:

Week	Sales
1	420 pounds
2	370 pounds
3	410 pounds
4	400 pounds

Being a practical sort, you decide to use this information to make your prediction. You forecast that weekly sales in the future will equal the average of past weekly sales. In this case, the average is 400 pounds.

Not only does this prediction have intuitive appeal, it has statistical virtue. The table below applies the prediction rule "retrospectively" against actual sales in past weeks and reports a statistical measure of prediction accuracy. The statistical measure is the sum of squared differences between the actual sales and the prediction.[2]

(1) Week	(2) Sales	(3) Prediction	(4) "Error"	(5) Squared Error
1	420	400	20	400
2	370	400	-30	900
3	410	400	10	100
4	400	400	0	0
				SSE = 1400

Columns (1) and (2) repeat the actual weekly sales data. Column (3) shows the prediction of 400. The "prediction error" is given in column (4); the squared error is in column (5). The total prediction error is the "sum of squared errors" for each week, or SSE. Remember, the SSE is computed using the historical data that you use to estimate your model. If your prediction is 400 each week, your SSE = 1400.

It turns out that if you must make the same prediction for each and every realization of (Y_1, Y_2, ... Y_n), then *you will minimize the SSE by predicting the mean.* Columns (1) through (5) of the following table echo the data in the previous table. Columns (6) through (8) show the computation of SSE if your prediction is 410 instead of 400. In this case, SSE = 1800.

[2] Squaring the error is a somewhat arbitrary choice for a scoring method, but it does make some sense, as it severely penalizes the most egregious errors. (It is also computationally simple—especially desirable when computers were slow.)

(1) Week	(2) Sales	(3) Prediction	(4) "Error"	(5) Squared Error	(6) Prediction	(7) "Error"	(8) Squared Error
1	420	400	20	400	410	10	100
2	370	400	-30	900	410	40	1600
3	410	400	10	100	410	0	0
4	400	400	0	0	410	10	100
				SSE = 1400			SSE = 1800

Do you see from week 2 that the SSE criterion severely penalizes large errors? Bear this in mind when thinking about outliers.

Of course, it is rather simple-minded to make the same prediction for every week. You may have additional insights into the market that enable you to make better predictions. For example, you might believe that coffee sales vary with the outdoor temperature, and you may have a reliable long-range weather forecast (as well as historical weather data) at your disposal. Regression analysis would allow you to determine the historical statistical relationship between temperature and coffee sales; you can then include weather forecasts when making your prediction.

WHAT ABOUT R^2?

This brings us back to R^2. Let SSE_T represent the total sum of squared errors when you set your prediction equal to the mean. Let SSE_R represent the sum of squared errors when you base your prediction on the results from the regression model. Then $R^2 = 1 - SSE_R/SSE_T$. In other words, R^2 equals the fraction of the variance in the dependent variable (i.e., variance around its mean) that is explained by the regression. Another way to put it is that R^2 measures the degree to which your regression model improves upon predicting the mean. If your regression model does a perfect job of predicting the historical data used in the regression, then $SSE_R = 0$ and $R^2 = 1$. If your regression does no better than if you had simply predicted the mean, then $SSE_R = SSE_T$ and $R^2 = 0$.

R^2 always increases when you add another RHS variable; this is why empirical researchers usually report the "adjusted R^2" or R^2_a:

$$R^2_a = 1 - [SSE_R/(n - k - 1)]/[SSE_T/(n - 1)] \quad (9)$$

where *n* is the number of observations and *k* is the number of predictor variables. If you inspect this formula, you can see that if you add predictors without materially decreasing the SSE_R, then R^2_a may decrease. (R^2_a can even be negative—not a good thing.) But don't worry about memorizing the formula, Stata and other statistical software packages will always compute R^2_a for you.

Note: From this point on, we will use the expressions R^2 and R^2_a interchangeably. Even so, you should generally report R^2_a.

A CAUTIONARY NOTE

There are several reasons to be wary of using R^2_a as a measure of regression performance.

- R^2 tells you how well a model fits historical data. If you are using your model for predictions, then you must be willing to assume that the model that generates future observations is the same as the model that generated your data. (For example, say you believe that the future influence of the weather on coffee sales will be similar to the past influence of weather on sales.) If this is not the case, then your model may have a high R^2 but may be worthless for prediction.
- On a related note, striving for a high R^2 could result in "overfitting" a model. Suppose you are interested in the effect of height on long-jump distance. Suppose the range of heights in your data is 5'8" to 6'5", and the tallest person jumps a short distance. You might get the highest R^2 by creating a separate dummy for height of 6'5", and that dummy will have a negative coefficient—but this is not useful for predictive purposes. You will not want to assume that all tall people will have short jump scores just because of this one outlier that you have "over-fitted" with its own dummy.
- You cannot compare R^2s of regressions with different dependent variables or different observations—these are apples-to-oranges comparisons. There are excellent, informative regressions with R^2s of 0.10 or less, as well as poor, uninformative regressions with R^2s of 0.9 or more. In other words, R^2 is not a reliable measure of regression performance.

One reason why R^2 is not a reliable measure of regression performance is that sometimes it is easy to get a high R^2 even though you have built a rather poor regression model. For example, suppose you are attempting to predict next year's U.S. GDP. You have historical data on GDP going back fifty years. Just for fun, you regress past GDP on U.S. population. To your surprise, you get $R^2 = 0.90$. Or suppose you are trying to explain variation in medical expenditures across hospitals. Your dependent variable is total medical expenses, and the predictor variable is number of patients. This time, you get $R^2 = 0.97$.

These are high R^2s, but hardly a cause for rejoicing—*you have not learned anything that you did not already know*. Of course GDP trends are correlated with other longitudinal trends, and of course larger hospitals have higher patient care expenses.

These simplistic models generate large R^2s because in each case the LHS variable has a key *trend* component. GDP trends upward over time. Firm costs trend upward with firm size. Strong trends beget high R^2s without providing much insight into the real world. (See the **Appendix** for a discussion of one way to think about regression performance in the presence of trending variables.)

A regression can have an R^2 of 0.10 or less and still be informative.

To better understand this claim, remember what R^2 means. It is the amount of variation in the LHS variable (around its mean) that you can explain with your RHS variable. To a certain extent, any positive adjusted R^2 is okay because it means that you have managed to explain something.

Sometimes even explaining a little bit can be valuable. For example, certain scenarios contain a powerful random element that causes a lot of unexplained variation in the LHS variable. In these cases, a low R^2 is nothing to complain about. A good example is the movement of stock prices. If you could predict even a small percentage of the movement of stock prices, you could make a fortune. (Some of the most important recent work in finance has attempted to obtain a

positive R^2_a in models explaining the movement of stock prices after accounting for CAPM. An R^2 of 0.01 or 0.02 in these scenarios has proven to be a revelation.)

Once you have worked with a data set for a while, you will get a feeling for whether a particular R^2 is achievable. In my own experience working with models predicting costs or prices, I am thrilled to get an R^2 in excess of 0.4 or 0.5 (aside from any trend). If you are modeling profits, do not expect R^2 to exceed 0.2 or 0.3. In all cases, do not be concerned if your R^2 is small. The key question is whether you have learned something valuable.

AN ALTERNATIVE APPROACH TO ASSESSING REGRESSION PERFORMANCE

Rather than relying on R^2, most statistical researchers use an entirely different criterion for evaluating regression performance by asking whether the regression allows them to confirm or refute key hypotheses. If key predictor variables turn out to be statistically significant and important, then the regression has value. For example, suppose you have been retained by a human resources consulting firm to evaluate whether the use of stock options in executive compensation affects firm performance. You run a regression and achieve a low R^2, but the regression does allow you to state with confidence that the use of stock options will typically reduce profit margins by 5 percent. This is an important insight regardless of the R^2. It is to these issues of confidence and importance that we now turn.

Note: Although we generally downplay the importance of R^2, there is one circumstance in which it is a valuable tool. In particular, *if you are comparing two regressions that have the same dependent variable and the same observations*, then you can use the R^2 as a basis for assessing regression performance.

Statistical Significance and Economic Importance

Statistical Significance

Even before they look at the regression R^2, most empirical researchers examine the coefficients on key predictors to determine if they are statistically significant. Researchers focus on statistical significance because they are often testing hypotheses. They usually test a *null hypothesis* that the predictor variable has no effect on the dependent variable. Stated another way, the null hypothesis is that the coefficient on the predictor variable equals zero. If the observed coefficient is larger in magnitude than expected due to random chance, then the null hypothesis is rejected and the researcher concludes that the predictor variable does affect the independent variable. Researchers usually report the *significance level* of the coefficient.

The significance level of a coefficient is the probability of observing such a large coefficient or larger (in absolute value) if the predictor variable actually has zero effect on the dependent variable (and therefore any observed effect is just due to random chance).

Recall from your statistics class that the significance level is determined by (a) dividing the coefficient by its standard error, and (b) comparing the resulting *t-ratio* to a statistical table. Fortunately, all regression packages do this for you. The conventional significance level for rejecting the null hypothesis is 0.05 or less. The researcher will write this as "the coefficient is significant at $p < 0.05$." The 0.05 cutoff is somewhat arbitrary. There is nothing that makes a

change in significance levels from 0.06 to 0.05 inherently more important than a change from 0.05 to 0.04. In fact, different researchers may refer to significance in a number of ways in order to capture the fact that there is no absolute cutoff for significance. The following table illustrates the ways to describe significance levels:

Significance Level	Description
$p < 0.01$	Highly significant
$p < 0.05$	Significant
$p < 0.10$	Significant at $p < 0.10$ (to denote that this is not a conventional significance level)
$0.05 < p < 0.10$	Borderline significant
$0.10 < p < 0.15$	Borderline insignificant

It is important for researchers to establish a high hurdle for hypothesis testing; otherwise, too many hypotheses would be accepted too rapidly. This is especially true *when you are testing many hypotheses because some are likely to appear significant just by chance*. If a coefficient is borderline significant, ideally the researcher should perform another experiment to obtain a more powerful test.

Managers are not researchers, of course, and should not necessarily adhere to the same standards. For one thing, managers may not have the luxury of waiting for the results of another test. Of greater importance, managers need to know not only whether a predictor variable matters but also how much it matters. To determine this, we need to examine the *importance* of the coefficient.

Economic Importance

To determine importance, follow these steps:

1. Compute a "reasonable" change in the predictor variable. You may have a particular change in mind (e.g., you are proposing to increase the advertising budget by $1 million and wish to project how this will affect sales).
    - A popular approach is to examine the effects of a one (or two) standard deviation change in the predictor variable.
    - A benefit of this approach is that it allows you to examine "typical changes" in each predictor, thus enabling an apples-to-apples comparison of different predictor variables.
2. Call this change ΔX. Compute the predicted effect of the reasonable change as follows:
    - In a simple linear model, you multiply the change by the estimated coefficient, that is, the predicted effect = $\Delta X \cdot \beta_X$.
3. If you have exponential or interaction terms, the calculations are a bit more complex, as we will now see.

COMPUTING MAGNITUDES: CALCULUS RULE

Suppose you have a regression equation with both exponents and interactions:

(10) $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 (X \cdot Z)$$

Suppose you are interested in computing $\Delta Y/\Delta X$, that is, the amount that Y changes when X changes. You may recall the following formula from calculus. If ΔX is small, then:

(11) $$\Delta Y = \Delta X \cdot \partial Y/\partial X$$

where $\partial Y/\partial X$ is the derivative of Y with respect to X.

Differentiating equation (9) with respect to X gives us:

(12) $$\partial Y/\partial X = \beta_1 + 2\beta_2 X + \beta_4 Z$$

This tells us how much a *small change* in X affects Y. Combining equations (10) and (11) gives us the following rule of thumb for computing the effect of a change in X:

(13) $$\Delta Y = (\beta_1 + 2\beta_2 X + \beta_4 Z) \cdot \Delta X$$

Note that if the exponent and interaction terms are zero, then $\beta_2 = \beta_4 = 0$, and this reduces to $\Delta Y = \Delta X \cdot \beta_1$, just as we would expect. If $\beta_2 \neq 0$ and/or $\beta_4 \neq 0$, then the exponent and/or interaction matter. In this case, the value of ΔY depends on the baseline values of X and/or Z. So you will need to pick values of X and/or Z before computing ΔY. This is also exactly what we would expect if exponents and/or interactions are important. Typically you would compute the effect of ΔX on Y by letting X and/or Z take on their mean values. But you may want to report these effects for other values.

COMPUTING MAGNITUDES: BRUTE FORCE METHODS

If your calculus skills are rusty or you are considering a large value for ΔX, you can use brute force algebra to estimate ΔY. This method works no matter what your regression looks like. Suppose the regression equation is:

(14) $$S = 5 - (0.2 \cdot P) + (2 \cdot A)$$

This is an easy one to work with. If A increases by 1 unit, S is predicted to increase by 2 units. We can thereby compute the change in S for any change in A simply by multiplying the change in A by 2.

Now suppose the regression equation is:

(15) $$S = 5 - (0.2 \cdot P) + (2 \cdot A) - (0.1 \cdot A^2)$$

This is a little harder. We need to rely on algebra to show what happens if A increases. Suppose that A increases by one unit. First compute S at some initial value for A, denoted A*. Call this S*:

(16) $$S^* = 5 - 0.2 \cdot P + 2A^* - 0.1A^{*2}$$

Now compute S** when $A^{**} = A^* + 1$:

(17) $S^{**} = 5 - 0.2 \cdot P + 2(A^* + 1) - 0.1(A^* + 1)^2 = 6.9 - 0.2 \cdot P + 1.8A^* - 0.1(A^*)^2$

We now subtract $S^{**} - S^* = 1.9 - 0.2A^*$. This is our brute force estimate of the effect of a one-unit change in A on S. For example, if $A^* = 5$, then increasing A to 6 would cause sales to increase by $1.9 - 0.2(5) = 0.9$. We can repeat this brute force calculation for an increase in A of any size, not just one unit. You should be able to follow the preceding steps to perform a brute force analysis of a regression with an interaction term.

Other Regression Issues

Sample Size

If you have ever seen a regression with thousands of observations, you will notice that virtually every predictor is significant. This will occur if the predictor really matters or if the predictor is slightly correlated with an omitted variable that really matters. What is going on is simple. Recall from basic statistics that the t-statistic used to determine significance equals the estimated coefficient divided by its standard error. You may also recall that the standard error of an estimated coefficient is an inverse function of the square root of the sample size. Thus, as the sample size increases, the standard error falls and the t-statistic increases. This makes everything seem significant.

When statisticians deal with large samples, they still rely on the R^2 and always assess the importance of the estimates, but they give a bit less credence to statistical significance. When sample sizes run into the thousands, many statisticians tighten the significance threshold to 0.01.

Comparing Coefficients in General

In the section on categorical variables, we discussed how to test whether the coefficients on two dummies (summer and spring) were different. The approach we used there is valid for comparing coefficients on any predictor variables.

For example, you might want to determine the effect of Internet and television advertising expenditures on sales. You have a measure of total sales (sales), a measure of expenditures on Internet banner ads (internet) and a measure of expenditures on television ads (tv). You not only want to know if advertising boosts sales but also if Internet ads or television ads are more effective.

You might start with a null hypothesis that Internet and television ad expenditures are equally effective, that is, $\beta_{internet} = \beta_{tv}$. If you cannot reject the null hypothesis, there is no statistical reason to suppose that Internet and television ads have different effects on sales. To test the null, Stata will construct an F-statistic that uses the variances and covariances of both coefficient estimates. All you need to do is type:

```
regress sales internet tv
test internet=tv
```

If the test result is significant, then you can reject the null hypothesis and conclude that Internet and television ad spending do not have equal effects on sales.

Note that this test is very flexible. We can test whether the effect of Internet ads is twice that of television ads by typing **test internet=2*tv**.

Appendix: Dealing with Trends (Optional Reading)

Here is one way to determine how well you have done in a regression in which the LHS variable follows a pronounced trend. Suppose you have the following model:

$$Y = \beta_0 + \beta_T T + \beta_X X$$

where T is some variable that picks up a trend (a measure of time, size, etc.). If Y and T follow the same trend, the R^2 will be high and you won't know if X is adding much predictive power. To find out if X adds to the model, do the following:

1. Regress Y on T.

2. Recover the coefficient β_T and compute $Y' = Y - \beta_T T$. We call Y' the *detrended* value of Y.

3. Regress Y' on X. The R^2 from this regression is a nice measure of how well you are doing aside from picking up the obvious trend.