Mgmt 469

Regression Basics

You have all had some training in statistics and regression analysis. Still, it is useful to review

some basic stuff.   In this note I cover the following material:

· What is a regression equation?

· How well does your overall model *explain the variation* in the dependent variable?

· What is the *statistical significance* of the coefficients on individual predictors?

· How *important* are the individual predictors?

What Is a Regression Equation?

A regression equation expresses a relationship between one or more *predictor variables*

and a *dependent variable*. The typical regression equation is usually *additive* and *linear*, like

equation (1):

(1) $$Y = B_0 + B_1 \cdot X_1 + B_2 \cdot X_2 + \varepsilon$$

We usually assume that ε (the error term) is a random variable that follows the normal

distribution and has mean 0.[1]

I will often refer to a regression of this form using the notation $Y = \underline{BX} + \varepsilon$, where

$\underline{BX} = B_0 + \sum_i (B_i \cdot X_i)$. In other words, $\underline{BX}$ represents the constant term B_0 plus the sum of each X_i

variable multiplied by its coefficient B_i. I will often drop the error term, especially when

showing you how to interpret regression coefficients. (I will occasionally drop the vector

notation (i.e., the __) if it is understood that X represents one or more variables.)

---

[1] Note that you can express X and Y in logs and still have the same basic linear structure.   It is possible to estimate nonlinear structures, such as $Y = B_0 + X_1^{B1} + B_2 X_2$.   You cannot use OLS to estimate B_1 and B_2 in such an equation.

Given this additive, linear structure, it is easy to calculate the expected (i.e., predicted) value of Y given any values of $\underline{X}$:

$$E[Y] = E[\underline{BX} + \varepsilon] = E[\underline{BX}] + E[\varepsilon] = \underline{BX}$$

where E[·] represents the expected value and $E[\varepsilon] = 0$ by assumption. In other words, the predicted value of Y is obtained by summing up the $\underline{BX}$'s. The residual is the difference between the actual Y and the predicted value $\underline{BX}$. OLS regression takes observations of the $\underline{X}$ and Y variables and estimates the $\underline{B}$ coefficients in equation (1) that minimize the sum of the squared residuals.

Regression allows you to determine to test the *direction* of a hypothesized relationship between a predictor variable X and a dependent variable Y. That is, regression can tell you whether an increase in X causes an increase or decrease in Y. But regression can tell you more than just the direction of a relationship; it can tell about the *magnitude* of the relationship. All it takes is a bit of algebra.

The Simple Algebra behind Regression Equations

Consider the following simple example of a regression equation:

Let S = Sales, P = Price, and A = Advertising expenditures.

You consider a simple regression equation to express the relationship between S, P, and A:

(2) $S = B_0 + B_1 \cdot P + B_2 \cdot A + \varepsilon,$

Given historical data on sales, prices, and advertising, you use OLS to estimate the coefficients in equation (2), obtaining $B_0 = 5$, $B_1 = -.2$ and $B_2 = .5$.

However, you can use a technique called "nonlinear least squares" which is well beyond the scope of the class.

Simple algebra dictates how to interpret these coefficients:

(a) Given particular values for P and A, you predict sales by plugging into the equation

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

For example, if Price=5, and Advertising=1, the best estimate of Sales is 5 - 1 + .5 = 4.5.

(b) If Price were to increase to 6, you would predict sales to decrease by 0.2 units (this is

the value of the coefficient B_1.) If advertising increased to 2, sales would increase by 0.5.

If your regression is unbiased (an issue we will take up in a later lecture), then these are your *best*

estimates of how price and advertising affect sales. This means that if you made similar

predictions many times, you would get them right *on average*. But your confidence about the

accuracy of any one prediction will depend on statistical significance, as we will discuss.

Key Point to Remember: In a linear model, the regression coefficient B_x gives the best estimate of

how much Y changes when there is a one unit increase in X.

Special Predictor Variables

The regression specification in equation (2) was as simple as they come. Both of the

predictor variables – price and advertising – were *continuous* (i.e., they could take on infinitely

many values). Neither predictor was *exponentiated* or *interacted with other variables*. You will

often want to use more complex specifications. This section covers several 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 – it takes on one of two values. These are often called *dummy variables* or *indicator variables*. Usually, the 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 differences in price and advertising. In this case, the "condition" is "Is it summer?" We can test whether this condition affects sales by constructing a dummy variable that we can call "Summer." We let Summer =1 for the 13 observations that correspond to summer weeks, and let Summer = 0 for the remaining 39 observations.

We estimate the following regression equation:

(3) $$S = B_0 + B_1 \cdot P + B_2 \cdot A + B_3 \cdot Summer + \varepsilon$$

In Stata, we would type:

regress S P A Summer

Equation (3) is just another algebraic expression, so it is easy to interpret the results. We apply the basic rule that the regression coefficient tells us how much the LHS changes when a RHS variable increases by one unit and everything else is constant. In the case of the categorical variable Summer, the coefficient B_3 *indicates the amount that sales increase when the value of Summer increases by one unit* (i.e., increases from 0 to 1) while P and A are constant.

To be more precise about this interpretation, remember that Summer equals 1 for the summer weeks and 0 for all other weeks. Thus, *the coefficient B_3 tells us how much more is sold*

in summer versus all the other weeks, while holding P and A constant.

Here is another way to look at it. If Summer=1, then equation (3) can be rewritten:

(3a) $$S = (B_0 + B_3) + B_1 \cdot P + B_2 \cdot A + \varepsilon$$

If Summer=0, we have:

(3b) $$S = (B_0) + B_1 \cdot P + B_2 \cdot A + \varepsilon$$

Comparing (3a) with (3b) reveals that the condition Summer=1 *shifts the intercept of the regression line* by the amount B_3. (For convenience, I will henceforth drop the ε error term.)

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

Important reminder: The coefficient on the dummy tells us how the value of the dependent variable changes when the dummy variable switches from 0 to 1.

You may wonder what happens if your categorical variable can take on more than two values. Seasonality is a good example. Earlier, we divided the year into two periods (summer/non-summer) but it is more natural to crease a dummy variable for each season. (Let Spring = 1 if the week falls during the spring and let Spring = 0 otherwise, etc.)

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

(4) $$S = B_0 + B_1 \cdot P + B_2 \cdot A + B_3 \cdot Summer + B_4 \cdot Spring + B_5 \cdot Winter + B_6 \cdot Fall$$

For technical reasons that I will describe in class, OLS cannot perform this estimation. 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, you must estimate something like equation (5), which omits Fall:

(5) $S = B_0 + B_1 \cdot P + B_2 \cdot A + B_3 \cdot Summer + B_4 \cdot Spring + B_5 \cdot Winter$

In Stata, you type:

regress S P A Summer Spring Winter

The coefficients B_1 and B_2 give the effects of price and advertising on sales while controlling for seasonality. The coefficients B_3, B_4, and B_5 measure *the average sales in the corresponding season relative to the average sales in the omitted season* (in this case, fall), while controlling for price and advertising!

The regression coefficients on the seasonality variables allow you to compare sales in each season with sales in Fall, all else equal. If you want to make other seasonality comparisons (e.g., Summer versus Spring), you have two choices:

1) Re-estimate the model, include Fall, but exclude Spring:

regress S P A Summer Fall Winter

The coefficient on Summer indicates how much is sold in Summer relative to Spring.

2) Directly compare B_3 and B_4 in equation (5). The difference (B_3 - B_4) is the best estimate of the difference in sales between Summer and Spring, all else equal.

You will get exactly the same result regardless of which choice you make.

Now you know how to measure the magnitude of the Summer versus Spring effect. You will also want to examine statistical significance. If you choose option (1), you just look at the significance of the coefficient on Summer. But this could quickly become cumbersome if you have to re-estimate the model every time you want to do a different comparison.

I recommend that you choose option (2) and stick with the original equation. **Stata** offers a very easy way to test if B_3 and B_4 are significantly different from each other. We will learn more about this test at the end of this session; all you need to know is the following syntax:

      regress sales P A Summer Spring Winter

      test Summer=Spring

If the test statistic is statistically significant (e.g., p-value < .05), then the average level of sales in Summer is significantly different than the average in Spring.


General Rules for Dealing with Categorical Variables

1) When working with a categorical variable, create a dummy for each category.

2) If there are N categories, use N-1 dummies in the regression.

3) The coefficient on a particular dummy indicates *the average value of the dependent variable for the corresponding category versus the average value of the dependent variable for the omitted category,* all else equal.

4) 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 coefficient B_x indicates how much Y increases when X increases by one unit. This effect is constant – it equals B_x regardless of the value of Y or the baseline value of X (i.e., the value of X prior to the one unit increase.) If you were to plot Y on X, you would get a straight line. Thus, this is called a *linear specification*.

**Linear Specification of Advertising on Sales**
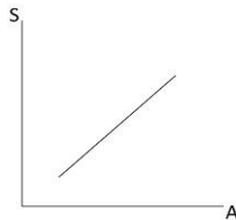
For the real world situation you are examining, you might wonder if the effect of X on Y is nonlinear. For example, you might wonder if advertising expenditures display diminishing returns (so that the plot of S on A was concave rather than linear.)

One way to allow for nonlinear effects is to include an exponential term in the regression:

(6) $$S = B_0 + B_1 \cdot P + B_2 \cdot A + B_3 \cdot A^2$$

By including A^2 in equation (6), the effect of advertising on sales may vary with the level of advertising. For example, if the benefits of advertising decline as A increases, then $B_2 > 0$ (in general, advertising helps) and $B_3 < 0$ (advertising helps less as A gets very big).

Exponential Specification of Advertising on Sales
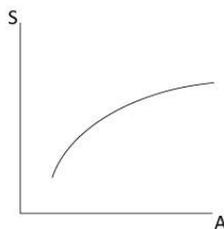
Suppose you estimate equation (6) and obtain: $B_0 = 8$, $B_1 = -.2$, $B_2 = .5$, and $B_3 = -.05$.

Then if $P = 5$ and $A = 4$, the best prediction for sales is

$$S = 8 + (-2){\cdot}5 + .5{\cdot}4 + (-.05){\cdot}4^2 = 8.2.$$

Later on in this note, I will explain how to compute the effect of a change in A on the value of S.

Interactions and Slope Dummies

Sometimes the effect of one RHS variable may *depend* on the value of another RHS variable. An *interaction term* can pick up such interdependence. For example, suppose that the effect of advertising depends on the price level, perhaps because advertising is more effective for high price products. If you want to determine if the effect of one predictor (e.g., advertising) depends on the value of another (e.g., price), you need to include an *interaction* variable.

To create a variable that captures the interaction of two predictors, simply multiply the two predictors together. For example, equation (7) includes an interaction term that captures the possibility that the effect of advertising depends on the price:

(7) $Sales = B_0 + B_1{\cdot}P + B_2{\cdot}A + B_3{\cdot}(A{\cdot}P)$

The term $A{\cdot}P$ is an *interaction term*. (Note: $A{\cdot}P$ is just the product of A and P. You use the product as a predictor in your regression.) The coefficient B_3 indicates if the impact of advertising on sales depends on the level of price. (Of course, this interaction is symmetric. It also tells us if the effect of price on sales depends on the level of advertising.)

Suppose you estimate (7) and obtain $B_0 = 6$, $B_1 = -.4$, $B_2 = .4$, $B_3 = .1$. Then if $P = 5$, and $A = 4$, the best prediction for $S = 6 - 2 + 1.6 + 2 = 7.6$.

If the effect of a variable depends on the value of a categorical variable, you again

multiply the two together. Consider the following equation:

$$Sales = B_0 + B_1 \cdot P + B_2 \cdot A + B_3 \cdot (Summer) + B_4(A \cdot Summer)$$

B_2 is the effect of A in the Winter, Spring, and Fall. The variable A·Summer is sometimes

known as a *slope dummy*. It tells us how much the slope (the effect of A) varies in the summer

versus the rest of the year. Thus, the effect of A in the summer is $B_2 + B_4$.

Regression performance

The best-known regression statistic is R^2. But what exactly is R^2? Without being too

precise at this point, think of it as the percentage of the variation in the dependent variable 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. To

understand why not, it is useful to refresh our memories about R^2.

The R^2 statistic derives from using regression as a tool for prediction. Let Y_i denote the

level of sales in a given week i. Suppose that you have collected observations for N prior weeks:

$(Y_1, Y_2, ..., Y_N)$. You would like to predict the level of sales in future weeks: Y_{N+1}, Y_{N+2}, etc.

If you have no additional information to suggest that sales might be higher in some weeks

than in others, you might as well make the same prediction for every future week.   Call this

prediction B_0. Needless to say, the actual Y_i may be higher or lower than B_0. Let e_i equal the

prediction error – the difference between the actual Y_i and the prediction B_0. This gives us

equation (8):

(8) $Y_i = B_0 + e_i$

A natural instinct is to select B_0 to equal the mean value of all previous realizations of Y. This makes sense and we do this sort of thing in everyday life. To predict the score my son will receive on his next math test, for example, I could worse than use the average of his past tests. To move us further down the road towards understanding R_2, let's use an example to see what the prediction error looks like when you let B_0 equal the mean.

In this example, you own a coffee shop and you want to predict the sales of coffee beans for the upcoming weeks. You 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. Specifically, you forecast that future weekly sales will equal the average of past weekly sales (400 pounds.)

Not only does this prediction have intuitive appeal, it has statistical virtue, provided we use the *prior weeks' data* to measure statistical performance. The next table compares the prediction of 400 pounds weekly sales against actual sales and reports a statistical measure of prediction accuracy –the sum of squared differences between the actual sales and the prediction.[2]

---

[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.

Comparing a simple prediction against the actual outcomes

(1)	(2)	(3)	(4)	(5)
Week	**Sales**	**Prediction**	**"Error"**	**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) show the prior weeks' sales data. Column (3) shows the prediction of 400.

Column (4) reports the "prediction error" and the squared error is in column (5). The total

prediction error is the "sum of squared errors" for each week, or SSE. In this case, SSE = 1400.

Now for the punch line: If you must make the same prediction for each and every realization of

Y_i, then *you will minimize the SSE by predicting the mean.*

The following table illustrates how SSE changes if your prediction is slightly different

than the mean. Columns (1)-(5) are identical to those in the previous table. Recall that

SSE=1400. Columns (6)-(8) of show the computation of SSE if your prediction is 410 instead

of 400. In this case, SSE = 1800.

Comparing SSE for Different Predictions

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Week	Sales	Prediction	"Error"	Squared Error	Prediction	Error	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

If you examine week 2 in the above table, you can see that the SSE criterion severely penalizes

large errors. Notice how the squared error for this week dwarfs the rest of the squared errors.

Bear this in mind when thinking about outliers – one very bad prediction can dominate the SSE.

It is rather simple-minded to make the same prediction for every week. If you have

additional data and some understanding of the market you might make better predictions. For

example, you might believe that coffee sales vary with the temperature, and you may have a

reliable long range forecast. Regression analysis would allow you to determine the historical

statistical relationship between temperature and coffee sales; you can include this information

when making your prediction. In all cases, OLS reports the coefficients that minimize the SSE.

What does this have to do with R^2?

If your goal is to minimize SSE, then you could predict the mean for every Y_i. But you

can reduce the SSE further by adding variables (like temperature) to your prediction model. 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 from your regression model that includes predictor

variables. We know that $SSE_R < SSE_T$. Then $\mathbf{R^2 = 1 - SSE_R/SSE_T}$.

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

selected the mean, then $SSE_R = SSE_T$ and $R^2 = 0$. Put another way, R^2 equals the fraction of the

squared variation in the dependent variable around its mean that is explained by the predictor

variables.

It seems as though a higher R^2 means more accurate predictions. But this need not be the

case. Remember that the computed value of R^2 is based on the historic observations included in

your regression, not the unmeasured observations you are trying to predict. (Only Nostradamus

could compute the R^2 for observations in the future.[3]) 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*. (E.g., 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.

There are a variety of other reasons why *a high R^2 is not always indicative of a good model*. You always increase R^2 by adding to your set of RHS variables, *even if the added variables are chosen at random*. Adding random variables may reduce the prediction error for past observations, but is likely to reduce the accuracy of predictions of future observations, which is your real concern. This is why empirical researchers usually report the "adjusted R^2" or R^2_a :

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

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!) Don't worry about the formula, Stata and other packages will always compute R^2_a for you.

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

---

[3] See http://en.wikipedia.org/wiki/Nostradamus

More on evaluating regression performance

You may be tempted to evaluate the usefulness of a regression by comparing the R^2_a to some threshold. (E.g., "The regression is worthless if $R^2 < .30$.") Or you may be tempted to compare the quality of two regression models by comparing the sizes of the R^2's. ("My mom's R^2 is bigger than your dad's R^2.") This is okay, provided that you are comparing two regressions that have *the same dependent variable and the same observations*. Under no other circumstances can you judge the performance of a regression by examining the R^2 in isolation. There are excellent, informative regressions with R^2 of .10 or less, and poor, uninformative regressions with R^2 of .90 or more!

One reason why R^2 is not a reliable measure of regression performance is that there are some situations where it is easy to get a high R^2, even with 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 and U.S. population going back for 50 years. You regress past GDP on U.S. population and obtain $R^2 = .90$. 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 $R^2 = .97$.

These are gaudy R^2s! But they are hardly a cause for rejoicing. These simplistic models generate large R^2 because, in each case, the LHS variable has a key *trend* component. GDP trends upward over time. It will be highly correlated with anything else that trends upwards over time, including population. Firm costs trend upward with firm size and will be highly correlated. Strong trends beget high R^2's. But surely you want to strive to learn more than this.

By relying on trends, you could produce a model with a high R^2 that is absolutely foolish. Suppose you have been asked by the National Collegiate Athletic Association (NCAA) to demonstrate the benefits of college athletics to the U.S. economy. You regress 50 years of data on GDP data on the number of teams that participate in the NCAA college basketball tournament. Because the NCAA tournament has grown in size over the years, from 16 teams to the present 65, you get an R^2 of .90. So what? This proves nothing.

Detrending (Optional)

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

$$Y = B_0 + B_T T + B_X X$$

where T is some variable that picks up a trend (T could be a measure of time, size, etc.). To find out whether X is informative, independent of the trend, you can do the following:

1) Regress Y on T

2) Recover the coefficient B_T, and compute Y' = Y-B_TT. Y' is 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.

Steps for dealing with a trend

1) Nail down the trend. Is there a trend in time, size, or some other variable?

2) Find a measure of that trend – year, firm size, etc.

3) Identify other variables that explain the LHS variable, aside from the trend.

4) Run your OLS regression including all the RHS variables, including the trend variable if so desired.

5) (Optional) To assess regression performance, run a regression of the detrended LHS variable on the list of variables identified in step 3.

A regression can have an R^2_a of .10 or less and still be informative.

Remember that R^2 measures the extent to which your RHS variables explain squared variation in the LHS variable around its mean. This means that any positive adjusted R^2 might be okay, because it means that you have managed to explain something! Sometimes, even explaining a little bit can be very valuable.

For example, consider trying to predict movements in the stock market.   This is subject to a powerful random component, so that there will always be a lot of unexplained variation. In this case, a low R^2 is nothing to shy away from. If you can 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 attempts to obtain a *positive R_a^2* in models explaining the movement of stock prices after accounting for CAPM. Even a R_a^2 of .01 or .02 can 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 .40 or .50 (aside from any trend). If you are

modeling profits, do not expect R^2 to exceed .20 or .30. In all cases, do not be concerned if R^2 is

small. The key question is whether you have learned something valuable.

Rather than rely on R^2, most statistical researchers use an entirely different criterion for

evaluating regression performance.   They ask 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 that a human resources

consulting firm has retained you to evaluate whether the use of executive stock options affects

firm performance. You run a regression and get a low R^2, but the regression does allow you to

say, 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.

Hypothesis testing, statistical significance and economic importance

Before they look at the R^2, most researchers examine the coefficients on key predictors to

determine if they are statistically significant. Researchers focus on significance because they are

often testing hypotheses. Researchers 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.[4] If the observed coefficient is larger in

magnitude than would be expected due to random chance, then the null hypothesis is rejected and

---

[4]The null hypothesis is the prevailing hypothesis unless evidence dictates that it is incorrect.   The standards of
scientific inquiry usually hold that the null hypothesis is that the predicted effect does not exist. This raises the
statistical hurdle for demonstrating that a hypothesis is correct.

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 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 .05 or less. The researcher will write this as "the coefficient is significant at $p < .05$" and may get excited if a coefficient just meets the .05 significance threshold. But the .05 cutoff is arbitrary! There is nothing that makes a change in significance level from .06 to .05 inherently more important than a change from .05 to .04. In fact, different researchers may refer to significance in a number of ways, to capture the fact that there is should not be an absolute cutoff for significance.

The following table illustrates some of the terminology.

Ways to Describe Significance Levels

Significance level	Description
$p < .01$	Highly significant
$p < .05$	Significant
$p < .10$	Significant at $p < .10$ (to denote that this is not a conventional significance level)
$.05 < p < .10$	Borderline significant
$.10 < p < .15$	Borderline insignificant

It is important for researchers to establish a high hurdle for hypothesis testing; we should be skeptical when evaluating hypotheses lest we accept them all too rapidly. This is especially true when you are testing many hypotheses, so that some are likely to appear significant just by chance.

Managers are not researchers 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. More importantly, managers must know the *importance* of a variable, not just its significance level.

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 avoiding an apples-to-oranges 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 $= \Delta X \cdot B_X$.

3) If you have exponential or interaction terms, follow the steps outlined below.

Computing magnitudes when you have exponents and/or interactions

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

(9) $$Y = B_0 + B_1X + B_2X^2 + B_3Z + B_4(X{\cdot}Z)$$

You are interested in computing how much Y changes when X changes by some amount ΔX.

Because X appears in several places in equation (9), you cannot simply use B_2. You have two

choices for how to proceed: the "derivative technique" or the "brute force technique." If ΔX is

small, you should use the former; if ΔX is large, you must use the latter.

Derivative Technique

You may recall the following formula from the calculus. If ΔX is small, then:

(10) $$\Delta Y = \Delta X \cdot \partial Y/\partial X,$$

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

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

(11) $$\partial Y/\partial X = B_1 + 2B_2X + B_4Z$$

This tells us how much a *small change* in X affects Y. Combining (10) and (11) gives us the

following rule of thumb for computing the effect of a small change in X:

(12) $$\mathbf{\Delta Y = (B_1 + 2B_2X + B_4Z) \cdot \Delta X}$$

Note that if the exponent and interaction terms are zero, then $B_2 = B_4 = 0$, so this reduces to

$\Delta Y = \Delta X \cdot B_1$, which is exactly what we would expect! Otherwise, you will need to know the

values of X and/or Z to determine how a change in X affects Y. This should come as no surprise,

because the estimated relationship between X and Y is nonlinear and/or depends on Z.

You will have to specify baseline values of X and Z in order to solve equation (12) and compute ΔY. The choice of baseline values is up to you. Most analysts find it useful to choose the mean or median values of X and Z. If Z is a 0/1 categorical variable, you might want to compute ΔY for Z=0 and then again for Z=1. Depending on your application, you may find it more valuable to choose other values for X and/or Z. We will discuss these options at greater length in conjunction with our hands-on projects.

Brute force technique

Suppose you are considering a large value for ΔX. In this case, equation (10) is no longer reliable and you cannot use the derivative technique. You must instead rely on "brute force" algebra to estimate ΔY. Here is an example.

Suppose the regression equation is:

$$Y = 5 - .2X + 2Z - .1Z^2$$

You want to know what will happen to Y when Z increases by 2 units. Here is the math:

1) Choose baseline X* and Z*.

2) Compute Y*:

$$Y^* = 5 - .2X^* + 2Z^* - .1Z^{*2}$$

3) Compute Y** when Z** = Z* + 2 (i.e., a two unit increase in advertising):

$$Y^{**} = 5 - .2X^* + 2(Z^*+2) - .1(Z^*+2)^2 = 8.6 - .2X^* + 1.6Z^* - .1Z^{*2}$$

4) Compute $\Delta Y = Y^{**} - Y^* = 3.6 - .4Z^*$.

For example, if Z* = 5 and X*=5, then increasing Z to 7 yields $\Delta Y = 3.6 - .4(5) = 1.6$.

You should be able to follow the preceding steps to compute the effect of any discrete change in a predictor variable. If the change in the predictor is small, the brute force and derivative techniques should give virtually identical answers. Since the derivative technique is much easier to use, that is what most people report.

*The **mfx compute** (or "marginal effects compute") command*

The **mfx compute** command in Stata reports magnitudes of coefficients in a variety of ways. The default option of mfx compute reports *the derivative of the prediction with respect to a change in each predictor variable*. In the simplest application to OLS regression, **mfx compute** is actually redundant, because it just regurgitates the regression coefficients. For example, take a look at the results from the yogurt data:

```
. regress sales1 price1 promo1

      Source |       SS       df       MS                  Number of obs =      84
-------------+------------------------------              F(  2,     81) =  106.20
       Model |  158316238     2   79158118.9              Prob > F      =  0.0000
    Residual |  60373574.2    81   745352.768             R-squared     =  0.7239
-------------+------------------------------              Adj R-squared =  0.7171
       Total |  218689812    83   2634817.01              Root MSE      =  863.34


      sales1 |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      price1 |  -103869.9   15451.71     -6.72   0.000    -134613.9   -73125.84
      promo1 |    3696.35   535.0383      6.91   0.000     2631.791    4760.908
       _cons |   11806.41   1269.885      9.30   0.000     9279.738    14333.08


. mfx compute

Marginal effects after regress
      y  = Fitted values (predict)
         =   3567.0238

 variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
----------+-------------------------------------------------------------------
   price1 |  -103869.9       15452   -6.72   0.000  -134155  -73585.1   .081019
  promo1* |    3696.35      535.04    6.91   0.000   2647.69   4745.01   .047619
-------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

The first set of results is the basic regression. The second set shows you how y changes as each X variable changes. In this simple OLS case, the results from **mfx compute** exactly equal the regression coefficients.

23

But **mfx compute** has many virtues that make it worth exploring. You may recall from

your microeconomics class that it is possible to compute the elasticity of demand along any point

on a linear demand curve. (The formula is $\varepsilon = (\Delta Y/\Delta X)/(X/Y)$, where ε is the elasticity, $\Delta Y/\Delta X$

is the regression coefficient, and X and Y are measured at their means.) Rather than do this by

brute force, you can use **mfx compute**, **eyex** to compute the elasticity for you:

```
. regress sales1 price1 promo1

      Source |       SS       df       MS                  Number of obs =      84
-------------+------------------------------               F( 2,    81) =  106.20
       Model |  158316238        2  79158118.9             Prob > F      =  0.0000
    Residual |  60373574.2       81  745352.768            R-squared     =  0.7239
-------------+------------------------------               Adj R-squared =  0.7171
       Total |  218689812       83  2634817.01             Root MSE      =  863.34

      sales1 |     Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      price1 |  -103869.9   15451.71    -6.72   0.000    -134613.9   -73125.84
      promo1 |   3696.35   535.0383     6.91   0.000     2631.791    4760.908
       _cons |  11806.41   1269.885     9.30   0.000     9279.738    14333.08

. mfx compute,eyex

Elasticities after regress
      y  = Fitted values (predict)
         =   3567.0238

variable |      ey/ex   Std. Err.       z    P>|z|  [    95% C.I.    ]      X
---------+--------------------------------------------------------------------
  price1 |  -2.359223     .35645    -6.62   0.000   -3.05784   -1.6606   .081019
  promo1 |   .0493455     .00726     6.80   0.000    .035115    .063576   .047619
```

The estimated own-price elasticity of demand is –2.359. Note that this is computed at the mean

value of price, which in this example is .081. You may recall that the elasticity changes as you

move up and down a linear demand curve. Consult the Stata interactive manual to see how to

use **mfx compute** to compute the elasticity at other prices. (Type **help mfx compute**.) Because

promo1 is a dummy variable, there is no natural elasticity interpretation.

The **mfx compute** command can be used after any regression command, and really shines

when used after commands such as **logit** and **poisson,** for which the derivatives do not equal the

regression coefficients. We will discuss these at length later on in the course.

An important caveat: The **mfx compute** command fails to account for interactions and exponents.   If you have them, you will need to use the derivative or brute force techniques described above.

Other regression issues

Sample size

If you ever have seen a regression with thousands of observations, you will notice that virtually every predictor is significant. This will occur if the predictor really does matter, of course, but it can also occur if the predictor does not matter but is slightly correlated with an omitted variable that does.

What is going on is simple. Recall that the t-statistic equals the estimated coefficient divided by its standard error. Recall as well that the standard error of an estimated coefficient is inversely proportionally to 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 they 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 decrease the significance threshold to .01*.

More on comparing coefficients

One often wants to know if two coefficients are different from each other. We already encountered this issue in the seasonality example. To take another example, suppose you want to know whether spending on Internet or television advertising has a larger effect on sales. You have a measure of total sales (Sales), a measure of expenditures on Internet ads (Internet) and a measure of expenditures on television ads (TV). You run a regression of Sales on Internet and TV and obtain the coefficients $B_{Internet}$ and B_{TV}. The coefficients are not exactly equal, of course, and you understand that any observed difference might be due to random chance. Here is how you can tell if they are significantly different from each other.

Your null hypothesis is that Internet and TV ad expenditures are equally effective; i.e., $B_{internet} = B_{tv}$. (Alternatively, your null hypothesis is $B_{internet} - B_{tv} = 0$.) You wish to know if you can reject this null hypothesis. To perform this test, you require information about the variances of $B_{internet}$ and B_{tv} as well as their covariance. Fortunately, Stata stores this information after performing a regression and can easily perform the desired test:

regress Sales Internet TV

test Internet=TV

Stata reports the significance level of an F-test to determine if the coefficients are identical. If the test result is significant, then you can reject the null hypothesis and conclude that Internet and TV ad spending do not have equal effects on sales. Note that this test is very flexible. For example, you can test whether the effect of Internet ads is twice that of TV by typing

test Internet=2*TV