**Mgmt 469**

**Ten Steps towards Convincing Empirical Research**

Most of the course centers on regression analysis and related tools. The focus on regression is entirely appropriate. Most interesting strategy questions involve *relational analysis*; i.e., the exploration of how two or more factors relate to each other. The most important tool for relational analysis is *regression*.

Regression encompasses almost every major form of data analysis. Here is a short list of analytic tools that are closely related to regression:

· Correlation (a simple analysis of the relationship between two variables)
· Comparisons of means across two different populations (basically a regression with "dummy variables")
· ANOVA (the decomposition of the sources of variation in a variable)
· Conjoint Analysis (a tool used in marketing research)

Regression analysis has its own language. Here are some terms to remember:

· The variable on the left hand side of the regression is called the "dependent variable," the "Y variable," or the "LHS variable."

· The variables on the right hand side of a regression are called the "predictors," "independent variables," "X variables," or "RHS variables".

**The Gold Standard for Regression: Convincing Results**

The highest praise that one can give to an empirical research project is to say that the findings are *convincing.*

Convincing research has at least three characteristics:

1) The reported effects (i.e. the regression coefficients) are *unbiased*. There should be no reason to suspect that the reported effects are systematically larger or smaller than the real world effects under investigation. Unbiasedness depends on model and variable selection.

*Important note about bias*: If an estimate is free from bias, then it is correct *on average*. In other words, if we were to repeat our experiments and statistical analyses many times, the average of our estimates will converge towards the true effect we are studying. But the estimate we obtain from the one regression we do perform can still be wildly inaccurate – much too large or much too small. This is why freedom from bias is only one of our goals. Precision is another, and we evaluate precision by examining standard errors.

2) The *standard errors* are computed correctly and are small enough to generate precise estimates of the parameters of interest. Computation of standard errors under ordinary least squares requires strong assumptions about independence of observations and the distribution of the errors. If these assumptions are violated, you will need to modify how you compute the standard errors.

3) The findings are *robust*. Researchers are often uncertain about the best way to specify their model. *Your findings are robust if they hold up under any plausible specification of the model.*

**The Ten Steps**

The following steps can help prevent the most serious problems that emerge in empirical research and go a long way to assuring that your findings are convincing.

1) Think about a **model**

A well-structured question involves some hypothesized relationships between a dependent variable and causal factors. State the hypothesized relationships as specifically as possible. Are they theoretically plausible?

Your statistical model may be in the form of a simple question, as the following examples illustrate. In each case, there is a research question (q) and a corresponding statistical model (m).

1q) Does a change in the *value of the Euro* affect the *prices charged for imported beers* in the United States?

1m) Price of Imported Beer = f(Value of Euro, other predictors)

2q) Do *larger firms* have *higher health care costs* per employee than do smaller firms?

2m) Health care expenditures per employee = f(firm size, other predictors)

3q) Do *late adopters* of technology tend to use the same *brand* as early adopters?

3m) Choice of brand = f(time of adoption of technology, other predictors)

In each case, the model is expressed in the form $Y=f(\underline{X})$ ,where Y is the dependent variable and $\underline{X}$ represents a *vector* of independent variables (including the main predictor variables of interest and control variables). Each model has a testable hypothesis, namely, that one or more of the X variables has a posited effect on Y. (We do not have to posit the effects of "control" variables included in $\underline{X}$; they are included to improve the precision of the model and help avoid "omitted variable bias".)

2) **Think about causality** and make sure that it is unambiguous. Assuring that you can state the direction of causality is known as the *identification problem*.

Your goal is to determine how a change in some factor under management control, X, will affect some outcome, Y. A regression coefficient $\beta_x$ tells you if the factor X is correlated with the outcome Y, holding constant all other variables in the regression. But it is not sufficient to learn the value of $\beta_x$. You must also determine the direction of causality.

You may recall that there are several reasons why two variables X and Y may be correlated in the data:

1) X causes Y, as implicitly hypothesized by the model. If so, then changing X would definitely cause Y to change.

2) Y causes X (reverse causality).

3) Some third variable Z causes both X and Y to move in lockstep.

In all three cases, the regression coefficient is positive. But only in case (1) would changing X cause Y to change. If (2) or (3) are possible, the model is not identified and you cannot determine if and how changing X will affect Y.

As this discussion suggests, there are two types of identification problems.

· Confusion about whether X causes Y or Y causes X is called *simultaneity bias.*

· If some factor Z causes both X and Y to move in lockstep, then failure to include Z in the regression induces an *omitted variable bias*.

These problems are closely related (the underlying mathematics is nearly identical) and econometricians often use the term *endogeneity bias* to refer to either one of them. I will have much more to say about endogeneity bias later on in the course.

*The Importance of Story Telling*

A good way to think about identification is to "tell stories." For example, suppose that you want to evaluate whether firms should increase CEO compensation as a way to promote performance. You observe that when CEOs are paid more, their firms enjoy higher returns on assets (ROAs). In other words, $\rho$(ROA, CEO pay) > 0.[1] This correlation supports the view that higher compensation causes better performance. But it is not definitive.

To determine whether the direction of causality is identified, you should see if you can another story that might explain the observed correlation. Here are two alternatives:

· Many corporate boards reward CEOs when their firms do well. In this case, causality runs from performance to compensation, not vice versa. (Alternatively, you could note that many firms have incentive-laden CEO compensation contracts and that the use of such incentives causes both high pay and good performance.)

· In some industries, large firms tend to outperform smaller firms. Large firms also tend to pay their CEOs more (perhaps because it is harder work). Thus, firm size causes both high pay and good performance.

It is not necessary to be definitive about these alternatives. Any or all of them are enough to cause doubts about causality; sorting them out requires similar empirical approaches that we will discuss in class.

Good story-tellers have a grasp of the institutions as well as economic theories. They understand the possible relationships among variables, and can clearly articulate possible directions of causality.

[1] The symbol $\rho$ (which should be the Greek letter rho if our computers share character sets) stands for correlation.

3) **Obtain, organize, and clean the data**

It is ideal to think about the underlying model before you obtain your data. Otherwise you may convince yourself to rule out certain theoretical possibilities merely because you lack the data to test them. That could lead to disastrous consequences.

Data come from many sources – government web sites, consulting firms selling proprietary survey results, market research firms selling scanner data, etc. Most statistical software packages permit you to merge data from different sources. It is essential that you can *link* the data through a variable that is common in both data sets. It helps (but is not essential) if the link is numerical, such as the CUSIP code for firms, or census codes for U.S. geographic regions.

Once you have the data, you need to *clean* it. This means doing the best job you can to assure that there are no errors in the data, or that data which made sense for someone else to use is ready for you to use. For example, one of my data sets includes the zip codes of consumers. U.S. zip codes range from 00001 to 99999. I was dismayed to find zip codes as large as 998750135. After giving the matter some thought, I realized that the "bad" zip codes were actually nine-digit zips. I recovered the five digit zip codes by lopping off the last four digits.

This is an example of *dirty data*. *Dirty data is not ready for analysis because the values are incorrect or inappropriate.* You will often judge for yourself whether your data is clean enough to analyze. So get to know your data! Here are some steps you can follow.

1) Print out a distribution; look at the minimum, maximum and modal values. Do these make sense?

2) Examine outliers. (Outliers are extreme values of variables, or, in the case of regressions, observations that have extreme values of residual.) Determine if the outliers

are plausible. For example, a distribution of population by U.S. county should have an enormous range; a distribution of per capita income by county should not.

3) Pick some variables that you know from theory should be correlated with each other. Are they? If not, try to determine what went wrong. Perhaps there one enormous outlier is wreaking havoc with the correlation coefficient.

If the data look wrong, feel wrong, or even smell wrong, then they might be wrong. Clean them!

You may have to make judgement calls when cleaning data. If you have a good reason to believe that a value is incorrect and you think you know the correct value, then correct it! The zip code example is a no-brainer. Here is an example of an obvious typo. You have pricing data for ice cream where the prices range from 2.50 to 5.00, except one firm has prices of 350-400. It seems certain that the "outlier" prices have been entered in pennies rather than dollars. Once again, it is okay to rely on your judgment, change the values, and move on. It is sometimes not so easy to clean dirty data. If the price of ice cream was 375732, you would be hard pressed to figure out what correct price was supposed to be.

What should you do when you are uncertain about an outlier? One thing to bear in mind is that ordinary least squares (OLS) regression minimizes the *squared* residuals; this implies that an outlier dependent variable can have a really big squared residual. It follows that *one bad value of a dependent variable can profoundly influence your analysis*. The same is true to a lesser extent about predictor variables.

As important as it is to weed out incorrect outliers, it is equally important not to weed out correct values that happen to be out of line with the rest of your data. One good but unusual data point may be important to your analysis.

To summarize:

· It is sometimes easy to recognize data entry errors. By all means, go ahead and make the changes when you can.

· Sometimes, you will be unable to figure out whether the data is correct. You may wish to delete observations when the dependent variable is clearly in error. If a predictor variable is in error, you can stop short of deleting the entire observation, using techniques that will be described later in the course.

*Reminder: Cleaning data can be time consuming, but is a vital step in your analysis.*

4) **Match** the data to the theory

Data can be classified as either *primary* or *secondary*. Researchers collect primary data for specific projects. Secondary data is collected by one researcher but used by another. The advantage of primary data is that the researcher can choose what to measure. But even with primary data, and especially with secondary data, *it may take a leap of faith to accept that the variable for which you have data matches the theoretical variable of interest*. For example, you may hypothesize that new product introductions are a function of a firm's technical know-how. How would you define a product introduction? How would you measure "technical know-how?

When assessing the appropriateness of your variables, be sure to tell stories. You might convince yourself that higher levels of R&D spending would be a good measure of technical know-how. But what if firms spend more on R&D when they are trying to catch up with more technically savvy competitors? In that case, higher expenditures could indicate lesser know-how! (Bonus question: Do you see that R&D spending might be *endogenous*?)

Another measure of know-how might be accumulated patents. This is an intuitively plausible measure that does not seem to suffer from the causality problem described above. Thus, accumulated patents might be a fine *proxy* for know-how. It is easy to measure, correlated with know-how (we hope), and appropriate for use in regression analysis.

5) Once you are satisfied that your data are well matched to the theory, it is time to **delve deeply into your data.**

- Look for correlations, trends, etc.

- Find the *action* in the data – do the predictor variables move around a lot? Does the dependent variable show any patterns?

Most students remember the concept of multicollinearity from DS 431. When it is a problem, standard errors can get very large and coefficients on the correlated variables may "flip" signs. Unfortunately, most students have an overblown fear of multicollinearity, which is rarely considered a major problem in real world empirical research. Moreover, it is very easy to detect and correct for multicollinearity, as we will discuss at length in a later lecture.

*I will say this more than once before the quarter ends: Correlation among variables does not imply that you have a problem with multicollinearity!*

6) **Begin your empirical work by briefly examining a simple model, but then quickly settle on a richer *core model***

Start with a stripped-down approach to your analysis. For example, you could assess the claim that size is an important determinant of profitability by dividing the sample of firms into two groups, "big" and "small", and comparing mean profits across them. If this simple comparison fails, then odds are good that a richer comparison will fail as well.

Do not dwell too long on simple analyses, because *you may be omitting key control variables* and this could seriously bias your results. Most of your analysis should center on a *core model* that includes those variables that theory says matters most.

7) After an initial look at the core model in OLS, settle on an **appropriate empirical method**

OLS regression is usually a good starting point for analysis but, depending on the problem at hand, can generate misleading results and imprecise estimates. Remember that OLS assumes that the dependent variable Y is generated by some model along the lines of $Y = B_0 + B_1X_1 + B_2X_2 + \varepsilon$ where $\varepsilon$ is a normally distributed error term. This is not always the best way to describe the relationship between the predictor variables and the dependent variable. There may be other models that do a better job of describing how the world works.

There are many issues to address when selecting an empirical method and later lectures will cover many of them. The field of econometrics is largely devoted to identifying the appropriate empirical methods for different problems.

Here are some key questions that will guide your choice of methods.

- Is the dependent variable discrete (e.g., 0 or 1) rather than continuous?
- Is the dependent variable categorical (E.g., the choice among brands of automobiles)?
- Is the dependent variable truncated at 0?

If you answered "yes" to any of these questions, then a method other than OLS may be indicated.

You may need to *transform* the values of your variables to generate a more appropriate empirical specification. Here are some key questions that will guide your variable specification:

- Are any of the independent variables categorical? If so, OLS may be okay, but you might need to create "dummy" variables.
- Do you need to include interaction terms?
- Do you need to take logs, include polynomial terms, etc?

8) Assess the **error structure** for violations of OLS assumptions

OLS generates unbiased standard errors only when certain assumptions about the error terms are satisfied. Two important assumptions are:

· The errors are homoscedastic – that is, they have identical distributions

· The errors are independent

*Heteroscedasticity*

· Are the *magnitudes* of your errors systematically correlated with any right hand side variables, or any other variables you may think of?

· If so, you have heteroscedasticity. Your coefficients are probably unbiased. But the standard errors may be incorrect. We will discuss several tests and corrections.

*Non-independence*

· Are your data naturally "grouped," with multiple observations for a given firm, state, individual, etc.?

· If you have such groups, then the intragroup errors (e.g., the errors for each division within the firm) are unlikely to be independent. This means that you do not have as many degrees of freedom as suggested by the sheer number of rows of data. You must account for the grouping of the data, or again you will obtain the wrong standard errors. We will discuss some alternatives for doing this.

9) **Determine whether your results are robust**

As you analyze your data, you may decide that there are several models that are equally defensible. *Your results are robust if you obtain similar key results from a number of defensible models.* This implies that you can save yourself a lot of time if you do not try to find a mythical "best" model, but instead verify that your results are robust across several good models. If your results are not robust, then they are not convincing.

Robustness can be a confusing concept, so remember this *critical robustness question*:

· Are your key findings immune to *defensible* changes in the regression model?

Here are some secondary robustness questions:

· Do coefficients vary much as you add potentially important control variables to what seems to be a rich model?

· Are the results of "fancy" methods different from the results from OLS?

· Are the results driven by outliers?

If you answered “yes” to any of the above questions, then you may have a robustness problem. At the very least, you should try to gain a better understanding of relationships in the data that are causing these problems.

If your results are not robust, then perhaps you have done a poor job of specifying the model. More likely, the patterns in the data are not consistent enough to warrant a strong conclusion.

10) **Interpret** your results

No statistical analysis is complete until you interpret the results. Recall the purpose of the analysis — to answer a question of interest. Your interpretation should largely relate back to that question, focusing on two elements:

- What is the *direction* of the observed relationship between the predictor variable of interest and the dependent variable? Can you reject the null hypothesis of no relationship?

- What is the *magnitude* of the relationship? It is not enough to state that when a CEO dies unexpectedly, firms tend to lose value. You must go further and state that on average, these firms lose X percent of their value. This way, you move from theory to practicality. Always give a real world interpretation of your results!

**Here is a Summary of the Ten Steps**:

1) Write down a model

2) Think about causality

3) Obtain, organize, and clean the data

4) Match the data to the theory

5) Get to know your data

6) Examine briefly a simple model and then focus on a "core" model

7) Implement the appropriate empirical method

8) Assess the error structure and make necessary corrections to the model

9) Test for robustness

10) Interpret the results