Lecture 21 – "Doing" Applied Econometrics

Throughout this course I have attempted to give you a good grounding in the theoretical and applied aspects of econometrics at what amounts to an "advanced beginner" level. Though we have touched on some theoretical topics, we have not delved deeply into this arena in an effort to focus on the "doing" of econometrics.

The data that you recently had fun with on your take home exam is an example of some of the problems you can run into when doing econometrics. There is a world of difference between applied and theoretical econometrics. In this course you have been taught a wide variety of econometric techniques but our focus has been on the mechanics of estimation and testing with the exception of the midterm where you were asked to apply what you had learned to a data set with issues that were not identified for you.

It is unfortunate that you cannot follow a simple cook book approach to estimate any/every model you might wish to in any/every situation. What I hoped to do in this course is to give you a good basis in theory, and a good understanding of the methods by which you can address "bad" things that happen in the estimation process. Identifying these issues will now be up to you.

I do not call myself an econometrician. Perhaps I am merely an economist with above average analytical skills. Perhaps the truest statement is that I am an unapologetic empiricist. I find that the data will usually tell us what we need to know if only we are willing to listen, rather than torturing it into confessing crimes it did not commit.

So, I leave you with some advice that is due in large part to Peter Kennedy; The Ten Commandments of Applied Econometrics.

Rule 1. Use Common Sense (and economic theory).
Common sense is not all that common. A common sense approach does not necessarily require complicated econometric techniques. An example would be the selection of a functional form for your model that corresponds with the requirements of economic theory (e.g. homogeneity, returns to scale, technical change). Also, don't infer confuse causation from correlation.

Rule 2. Avoid Type III errors.
A type III error occurs when a researcher produces the right answer to the wrong question. A corollary of this rule is that an approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.

This issue here is that the relevant objective/hypothesis/specification may be completely different from what is initially suggested. Ask questions, especially seemingly foolish questions to ensure that you have a full understanding of the context of the problem. Be sure that you have formulated your approach to the research question appropriately.

Rule 3. Know the Context
This is an extension of rule 2. It is extremely important that the researcher become intimately familiar with the phenomenon being investigated, its history, institutions, constraints, measurement peculiarities, cultural customs, etc. all of which go well beyond a thorough literature review. Again you have to ask questions like how the data were gathered, what formulas were used to determine the values of the variables, what accounting conventions were followed, how was the questionnaire worded, how closely do the measured variables match their theoretical counterparts?

I often work with plant scientists on issues regarding plant growth, nutrient uptake, and plant diseases. I have received data including a column labeled "market yield". What exactly does that mean? It is something you have to pursue rather than assuming that the person who provided the data had done some work for you and was thinking about the problem in the correct economic context. Don't forget that you must know more than your computer.

Rule 4. Inspect the Data
Even if you know the context of the problem, you need to become intimately familiar with your data. This means you need to look at the numbers themselves, compute summary statistics, and above all graph the data so that you can get a feel for the way in which it is behaving. This rule also includes running diagnostics for influential observations and leverage points, VIF's and a condition index.

The advantage of graphing is that it will usually make you aware of things that you might not have previously considered such as outliers, trends, structural shifts and the like. You should also graphically examine the residuals from your OLS regressions.
Rule 5. KISS

Keep it Sensibly Simple – not to be confused with Keep it Simple, Stupid. Sometimes simple is stupid. It is okay to begin with a simple model as long as you understand that the point is to find a better way to mimic the data generating process. The drawback to this approach however, is that it may result in a good deal of pre-test bias. A good rule of thumb is to make your model as simple as possible avoiding the addition of variables that may or may not belong. Occam's razor is appropriate to remember here, the simplest explanation of any phenomenon is often the most accurate. William (of) Occam said it like this "*Numquam ponenda est pluralitas sine necessitate*." To quote Karl Popper "We prefer simpler theories to more complex ones because their empirical content is greater; and because they are better testable."

Rule 6. Make sure your Results Make Sense
Are your signs of your estimated coefficients correct given economic theory or expectations? Are the coefficients of reasonable magnitude? Are the implications of your results consistent with economic theory? Are the important variables statistically significant? Are you comfortable with the story that your data is telling?

Rule 7. Understand the Costs and Benefits of Data Mining.
The most undesirable version of data mining occurs when one tailors the specification to fit the data, which may result in a model specification that is misleading. This is precisely what occurs when you run a stepwise regression and allow the computer to select from among the set of all possible regressors, those that result in the "best" specification. Traditional testing procedures applied in this situation are no longer legitimate, because these data, which were used to select the regressors, cannot be considered impartial if used to test that specification.

Data mining that is undertaken to discover empirical regularities (or irregularities) can inform the researcher and benefit the process. This is not a sin. It is incumbent upon you as an applied econometrician to pay attention to what the data is trying to tell you, but never let it take the place of your common sense or knowledge or economic or physical relationships.

Rule 8. Be Prepared to Compromise.
In every economic analysis there is a gap between the problem at hand and the closest scenario which standard econometric theory is applicable. Very seldom does your problem come close to satisfying the assumptions under which econometric theory delivers the optimal solution. As such, you will be forced to compromise and adopt suboptimal procedures. In this class, as in any class on

applied econometrics, you have been given standard solutions to standard problems. In practice, you will be faced with compromises and must be willing to make *ad hoc* modifications to standard solutions.

Rule 9. Do not Confuse Statistical Significance with Meaningful Magnitude.
Very large sample sizes such as those that can be obtained using cross-sectional data give rise to estimated coefficients with very small standard errors. A consequence of this is that sometimes coefficients of trivial magnitude may be deemed significantly different from zero. This creates a misleading impression with respect to what is important. Always look at the magnitude of the coefficient (or more specifically the elasticities) as well as their significance when analyzing the economic importance of your coefficients.

A potentially more serious problem associated with significance testing is that there is a tendency to conclude that finding significant coefficients "sanctifies" a theory with the resulting tendency for researchers to stop looking for further insights. You should always be on the lookout for additional evidence, both corroborating and disconfirming.

Rule 10. Always Report a Sensitivity Analysis.
Econometricians base their analysis on an imaginary assumed data generating process which is viewed as having produced the data used in the estimation. It is, however, possible that your estimated equation doesn't even closely correspond to the true underlying data generating process. Because of this it is important to check to see if your empirical results are sensitive to the assumptions upon which it was based. Therefore, you should relax some of these assumptions and observe how the results change.

For example, you might change the sampling period (use only part of the data – or as new data become available add it to the model) change your functional form, change proxies in your data for different ones – again, be prepared to compromise.

Keep in mind that published research is almost never a description of the path or approach actually followed by the researcher.

These rules give rise to the Ten Commandments of applied econometrics as given by Peter Kennedy:

1. Thou shalt use common sense and economic theory.
2. Thou shalt ask the right questions.
   Corollary: thou shalt place relevance before mathematical elegance.
3. Thou shalt know the context.
   Corollary: thou shalt not perform ignorant statistical analysis.
4. Thou shalt inspect the data.
   Corollary: thou shalt place data cleanliness ahead of econometric godliness.
   Corollary: every number is guilty unless proven innocent.
5. Thou shalt not worship complexity.
   Corollary: thou shalt not talk Greek without knowing the English translations.
6. Thou shalt look long and hard at thy results.
   Corollary: thou shalt apply the laugh test.
7. Thou shalt beware the costs of data mining.
   Corollary: thou shalt not worship R^2.
   Corollary: thou shalt not hunt statistical significance with a shotgun.
   Corollary: thou shalt not worship the 5 percent significance level.
8. Thou shalt be willing to compromise.
   Corollary: thou shalt not worship textbook prescriptions.
9. Thou shalt not confuse significance with substance.
   Corollary: thou shalt not ignore power.
   Corollary: thou shalt seek additional evidence.
10. Thou shalt confess in the presence of sensitivity.
    Corollary: thou shalt anticipate criticism.

Keeping these ten commandments is not enough to guarantee quality applied work. Much of the skill in applied econometrics is judgmental and subjective. This can only be learned by experience, doing, writing, and publishing.

Other issues in applied econometrics:

Getting the Wrong Sign

A very common occurrence in applied work is to run your *a priori* favorite model specification and get a “wrong” sign. Rather than considering this a disaster, you should look upon it as an opportunity – a friendly message that some more work needs to be done.

The first step should be to check economic theory. It is amazing how after the fact people can conjure up rationalizations for incorrect signs. Be clear about what theory says regarding the behavior of economic agents. If there is good reason *a priori* to expect a different sign, then you have a moral obligation to seek econometric explanations for the wrong sign before suggesting that your theory be changed.

The Top Ten Reasons for Wrong Signs:

1. Omitted Variables. If you leave an important explanatory variable out of your model, this may result in the coefficient estimates for other variables in the model having incorrect signs.

2. High Variances. Collinearity may lead to high variances of estimated coefficients such that the confidence intervals for the parameters straddle zero and a wrong sign is a distinct possibility.

3. Selection Bias. If your data set was not obtained through random sampling (as is often the case in economics) you may need to add a dummy or correction variable to your model to explain phenomenon that are contributing to an incorrect sign.

4. *Ceteris Paribus* Confusion. Make sure you understand the correct interpretation of your parameter estimates. For example, suppose you have regressed house price on square feet, number of bedrooms, number of bathrooms, and a dummy variable to indicate whether or not the house has a family room. You find the coefficient on the dummy variable to be negative. This tells you that, holding everything else constant, if a family room is added, the price of the house falls. If you are holding all other variables constant, one of those is square feet. Thus, adding a family room would

entail a reduction in square footage elsewhere which means all other rooms would be smaller and the net effect of this on price is logically negative.

5. Data definitions/measurement. A common example here is that when you regress crime rate on the per capita number of police you may get a positive coefficient indicating that more police engender more crime. What may actually be happening is that in the face of increased reports of crime, municipalities hire additional police.

6. Outliers. There are times when examining economic phenomenon when you will encounter a true outlier which can obscure the coefficient estimates you would achieve from the rest of the data. Examining your data using DFBETAS and HATDIAG will help detect this. If it is a true outlier then you may delete it from your data set. I caution you on this however, because it is a practice I have seen abused and applied to data points which were not true outliers. I have removed an outlier from a data set only one time in my career.

7. Interaction Terms. Interaction terms may muddy the interpretation of your parameter estimates. Be sure of your calculus when you take the first derivative of the left hand side variable with respect to a right hand side variable. If an interaction term is included for any variable, then it must be included in your assessment of the impact of that variable on the LHS.

8. Specification Error. Incorrectly specifying your functional form or economic model can result in an incorrect sign. Carefully and critically examine your specification.

9. Simultaneity/Lack of Identification. Sometimes the causal flow from the explanatory variables to the dependent variables is not one way. This is particularly true when dealing with the economic phenomenon of supply and demand. Estimating an unidentified equation would produce estimates of an arbitrary combination of supply and demand coefficients, so the sign of a given coefficient can certainly be arbitrary as well.

10. Bad Instrument. Even if you do ensure identification, you may find the slope associated with an instrumental variable to have the wrong sign. This can occur when misidentify the reasons your instrument is correlated with the endogenous variable it replaced. There could be a plausible explanation

for the sign you are estimating that will only become apparent when you take a closer look at your data and the instrument you have chosen.

Other Common Mistakes in Applied Econometrics

1. Interpretation of a significant DW test or heteroscedasticity test as pointing to the need for you to run EGLS. These should initially be interpreted as meaning there is something wrong with your specification.

2. Forgetting interaction or quadratic terms when assessing variable influence. Pay attention to the derivative of your estimated equation with respect to the explanatory variable in question.

3. Using a linear functional form when the dependent variable is a fraction. A linear form could be adequate as long as your dependent variable has no observations close to zero or one. If this is the case you should use a logistic regression.

4. Believing that multicollinearity creates bias, or invalidates inference. No bias is created by multicollinearity. Estimated variances for the coefficients are inflated but the estimates themselves remain unbiased.

5. Using an ordered qualitative variable as a regressor. Using ordered variables rather than separate dummy variables for each category forces the impact of each level of the ordered variable to be the same.

What else do you need to know?

Unfortunately we cannot cover everything in this class. What follows are some of the things we did not cover that you should know well if you intend to be successful as an applied econometrician:

1. Instrumental variables. This technique is incredibly common in applied work.
2. Non-nested testing. Can be someone complex in multiple equation models but are easy to interpret.
3. Bootstrapping. Not always easy to do, but important to understand. Many awkward testing problems can be solved by bootstrapping.
4. MLE. Awkward estimation problems often require that you maximize likelihood rather than minimize the residual sum of squares.
5. ARIMA models. Univariate Box Jenkins models remain some of the most accurate forecasting models around.
6. VAR. Vector autoregressions are a classic method for analyzing time series data. They also do not require the specification of an underlying economic relationship.
7. Panel data. Understanding the difference between fixed and random effects estimation and the circumstances in which one is more appropriate than the other.
8. Nonstationarity. What is it and why do you have to worry about it. Learn how to test for unit roots, what cointegration is, and what the role of error correction models are.