ECOM20001
Econometrics 1

Lecture Note 9
Assessing Studies Based on Multiple Regression

A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 9

Summary of Key Concepts

- Internval and external validity
- Threats to external validity
  - Populations
  - Settings
- Threats to interval validity
  - Omitted variable bias
  - Model misspecification
  - Measurement error
  - Missing data and sample selection
  - Simultaneous causality
  - Sources of inconsistency with standard errors
- Forecasting with regression models

Assessing Empirical Analyses

A mathematician, an accountant and an economist apply for the same job.

The interviewer calls in the mathematician and asks "What do two plus two equal?" The mathematician replies "Four." The interviewer asks "Four, exactly?" The mathematician looks at the interviewer incredulously and says "Yes, four, exactly."

Then the interviewer calls in the accountant and asks the same question "What do two plus two equal?" The accountant says "On average, four - give or take ten percent, but on average, four."

Then the interviewer calls in the economist and poses the same question "What do two plus two equal?" The economist gets up, locks the door, closes the shade, sits down next to the interviewer and says, "What do you want it to equal"?

Assessing Empirical Analyses

- How should we go about assessing a statistical empirical analysis?
- Importantly, how can we develop good empirical analyses, and be able to quicky identify poor ones when we see them
- These are general skills: we are constantly seeing in the media and from other sources (potentially huge/important) conclusions being drawn from (potentially poor/shaky) empirical analyses

Assessing Empirical Analyses

- Two ways to organize your thinking in assessing an empirical analysis are internal and external validity
- Internal validity: when an analysis contains statistical inferences about relationships between variables that are valid for the population being studied
- External validity: when the inferences and conclusions from an analysis can be generalized from the population and setting studied to other populations and settings

Threats to External Validity

1. Differences in populations: when the population being studied is not relevant for understanding populations in other areas
   - would an analysis of minimum wage changes in Australia be relevant for New Zealand? United States? Russia?
   - would an analysis of commuting decisions by Melbournians from 1985 be relevant today?
2. Differences in settings: assuming the population being studied is the same, differences in the settings in which an analysis takes place could undermine the generalizability of the results
   - would the impacts of a change in class sizes at the University of Melbourne generalize to ANU? Similar underlying populations, but different contexts.

- Generally, differences in populations' or settings' characteristics, geography/physical environment, institutions, technologies across space or time can undermine a study's external validity

Threats to Internal Validity

- Internal validity contains two parts:

1. Unbiasedness and consistency: the estimator for the relationship you are interested in musht be unbiased and consistent
   - threats: violation of any of the least squares assumption
2. Hypothesis tests and confidence intervals should be correct: the actual rejection rate of the null in a finite sample should be the same as the desired rejection of the null
   - threats: small samples, heteroskedasticity

- There are 5 key threats to interval validity with multiple regression analysis

Threats to Internal Validity

Omitted Variable Bias

- Omitted variable bias emerges when there is a variable not included in a regression that is correlated with the dependent variable and with one of the independent variables, causing the independence assumption to fail
- If you have data on the variable, then including it in the regression fixes the problem
- Alternatively, if you have many control variables and you can assume mean independence with respect to a variable of interest, then this is another way of side-stepping omitted variable bias

Threats to Internal Validity

Omitted Variable Bias

- In dealing with potential omitted variable bias, you can take a 3-step approach to building up your regression analysis

1. Be clear up front what your coefficient(s) of interest are
2. Try to reason, before running regressions, what are potentially important control variables to hold fixed in estimating the relationship between the dependent variable and coefficient(s) of interest
3. Sequentially add control variables to your regression to see which matter for the sensitivity of the coefficient of interest.
   - this is precisely what you do on Assignments 2 and 3.

- ECOM30001: Econometrics 2 covers other approaches to avoiding omitted variable bias

Threats to Internal Validity

Misspecification of the Regression Function

- Functional form misspecification of the regression function yields biased OLS estimates
- Example: suppose the true population regression function is nonlinear, but the estimated regression function is linear
- Approaches to avoiding misspecification problems:
  - plot/visualize the data using scatter plots to see if nonlinear relationships potentially exist
  - try linear and non-linear specifications and test whether quadratic, cubic, etc., terms in the regression function are statistically significant and/or drastically improve model fit

Threats to Interval Validity

Measurement Error and Errors-in-Variables Bias

- What happens if, for example, in our test scores example we meant to study how class size affects economics tests scores, but accidentally used data on mathematics test scores?
- Economics and mathematics test scores are likely correlated, but they are not the same and the regression coefficients would be biased in this case in a model aimed at understanding economics tests scores
- This is an example of errors-in-variables bias, which is when there is error in the measurement of an independent variable
- Can arise very easily from practical issues such as data recording/entry errors (even at national statistics agencies!), survey respondents not understanding survey questions, or survey respondents not being able to accurately answer questions (say about income)

Threats to Interval Validity

Measurement Error and Errors-in-Variables Bias

- Suppose we are interested in the population regression:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- However, we mismeasure X_i and we have data on $\tilde{X}_i$, where $\tilde{X}_i \neq X_i$
- The actual regression we are running with the data we have is:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i] \\ & \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned}$$

  where $v_i = [\beta_1 (X_i - \tilde{X}_i) + u_i]$. This means the error in the regression we run v_i contains both the population error u_i and the measurement error $(X_i - \tilde{X}_i)$
- If $\tilde{X}_i$ is correlated with the measurement error $(X_i - \tilde{X}_i)$, then there is omitted variable bias in $\hat{\beta}_1$

Threats to Interval Validity

Classic Measurement Error

- Suppose that we have classical measurement error, such that:

$$\tilde{X}_i = X_i + w_i$$

  where w_i is a purely random variable with mean 0 and variance σ_w^2.
- Because it is purely random, we can assume $corr(w_i, X_i) = 0$ and $corr(w_i, u_i) = 0$ such that now there will be no omitted variable bias
- Even in this case, we have that as the sample size gets large, $n \to \infty$:

$$\hat{\beta}_1 \to \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1$$

- This means that even if measurement error is just adding randomness to X_i, $\hat{\beta}_1$ will be inconsistent and biased toward 0, even in large samples, because $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ is less than 1

Threats to Interval Validity

Fixing Measurement Error Problems

- Best way to fix/avoid measurement error problems is to accurately measure X_i
- Another approach to fixing them, which is covered in ECOM30001: Econometrics 2, is to use instrumental variables
- Researchers also sometimes try to directly model account for measurement error in regression models as well, which is a departure from using OLS in estimation

Threats to Interval Validity

Missing Data and Sample Selection

- Often in datasets you will have missing observations
- If observations are missing completely at random, this is not a problem for internal validity; it just reduces sample size and results in less precise OLS estimates
- When variables are missing based on the value of an independent variable X, this is also not a problem for the OLS estimator
  - for instance, in our class size – test score example, we did not see class sizes bigger than 35 students
  - while our regression results would not be informative for classes bigger than 35 students, they were internally valid within the sample that we had

Threats to Interval Validity

Missing Data and Sample Selection

- If data is missing as a function of the dependent variable Y, then there is a risk that X is correlated with u, and omitted variable bias is present. We call this sample selection bias
- Example: suppose we offer an mobile app to customers to help them understand their energy use
  - some customers use the app, others ignore it
  - we track energy usage for customers who used the app and estimate its effect by looking at their behavior before and after they had the app
  - problem: customers who use the app are a non-random selected sample; they are likely to be tech-savvy types
  - given this, we evaluating the effect of the app based on this selected sample would yield a biased estimate of its impact on the population
- Techniques for dealing with sample selection bias are covered in ECOM30002: Econometrics 2

Threats to Interval Validity

Simultaneous Causality

- Throughout the subject, we have implicitly been thinking that X "causes" Y in the subject :

$$X \rightarrow Y$$

- But what if it was the opposite, namely that Y causes X:

$$Y \rightarrow X$$

- Test scores example: we have tried to understand whether:

$$\uparrow \text{ class size } \rightarrow \downarrow \text{ test scores}$$

- What if the government hired teachers and built more classes at schools with low test scores. This would create reverse causality:

$$\downarrow \text{ test scores } \rightarrow \downarrow \text{ class size}$$

Threats to Interval Validity

Simultaneous Causality

- When it is possible for causality to "run backward" from $X \rightarrow Y$, we have the issue of simultaneous causality, which is yet another way in which omitted variable bias can arise in a regression of Y on X
- We can think of simultaneous equations models as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{2}$$

- Example: (1) would be the class size $\rightarrow$ test scores direction of causality, while (2) would be the test scores $\rightarrow$ class size direction of causality (from the government policy)
- Techniques for dealing with simultaneous equations and related bias are covered in ECOM30002: Econometrics 2

Threats to Interval Validity

Sources of Inconsistency in OLS Standard Errors

- Heteroskedasticity: when the variance of u depends on X in a regression, if not accounted for, you will compute incorrect standard errors and hence have incorrect t-statistics and confidence intervals
- Correlation of the error term across observations: if the error term is correlated across time or space, you can also end up with incorrect standard errors, t-statistics and confidence intervals
- Thus far, we have assumed IID random sampling, and hence ignored the possibility of correlated errors

Threats to Interval Validity

Sources of Inconsistency in OLS Standard Errors

- There are two common situations where correlation can arises:

1. Temporal correlation: when u is correlated over time
   - example: suppose you follow the test scores at the same school over time; we would expect the test scores and errors in the test–score vs. class–size regression to be correlated year-by-year
2. Spatial correlation: when u is correlated across space time
   - example: suppose you follow the test scores at two neighbouring schools in the same district; we would expect their test scores and errors in the test–score vs. class–size regression to be correlated across the nearby schools

- Techniques for dealing with correlated errors are covered in ECOM30002: Econometrics 2

Internal and External Validity with Forecasting

- Throughout the subject, we have focused on obtaining unbiased and consistent regression coefficient estimates, in possibly non-linear models, focusing on variable(s) of interest
- Another focus and use of regression models is forecasting, where we: (1) estimate a regression model for Y; and (2) use it to predict a value of Y for any combinations of X's
- In short, for any given set of X's, we can generate a forecast of Y using predicted values from the regression: $\hat{Y}$
- The corresponding forecast error is then the difference between what occurred and what was forecasted/predicted:

$$Y - \hat{Y}$$

Internal and External Validity with Forecasting

- In forecasting problems, interval validity is not a major concern. In fact, we do not care about the interpretation of the regression coefficient estimates at all. All we are concerned with is minimizing the forecast error
- External validity, however, is critical in forecasting problems
- We need to be sure that the underlying population that we drew our sample from in estimating the model can generalize in such a way that we can forecast outcomes in other situations/contexts

Internal and External Validity with Forecasting

- Netflix is a good example of forecasting in practice:
  - my past TV viewing behaviour is likely to be informative/predictive of what future TV shows I will like
  - my kid's TV viewing behaviour is definitely not informative/predictive of what future TV shows I will like
  - So estimating a regression model of TV show choice based on my wife's viewing behaviour would not have external validity for forecasting my TV viewing behaviour

Assessing Empirical Analyses: Fried Chicken Kills You (!?)

USA TODAY NEWS SPORTS LIFE MONEY TECH TRAVEL OPINION 32° CROSSWORDS MORE Subscribe Sign In

Eating fried chicken weekly could lead to early death, study finds

Josh Hafner, USA TODAY Published 3:22 p.m. ET Jan. 24, 2019 | Updated 12:58 p.m. ET Jan. 27, 2019

CONNECT TWEET LINKEDIN COMMENT EMAIL MORE

Ah, fried chicken. *(Photo: Pimonpim Tangosol, Getty Images)*

How would you like your death: regular, or extra early?

A new study links regular fried chicken consumption – one serving a week or more – to a 13 percent increased risk of premature death, at least among older women.

Researchers looked at nearly 107,000 postmenopausal women who tracked their diets in the 1990s for a nationwide study that followed them until 2017. But a fatal link to fried chicken may exist for other groups, too.

Ad closed by Google

Report this ad

Why this ad?

What do Economists Really Do?

Examples of important empirical research that is of the highest quality

- Inequality
  - Raj Chetty (Harvard), Opportunity Insights
    http://www.equality-of-opportunity.org/
- International Development
  - Esther Duflo (MIT), Poverty Action Lab
    https://www.povertyactionlab.org/
- Environment
  - Michael Greenstone (Chicago), Climate Impact Lab
    http://www.impactlab.org/
- Health
  - Amy Finkelstein (MIT), Poverty Action Lab
    https://economics.mit.edu/faculty/afink/
- (Social) Media
  - Matt Gentzkow (Stanford), Gentzkow-Shaprio Lab
    https://gentzkow.people.stanford.edu/

What do Economists Really Do?

- "What do Economists Really Do?" keynote lecture by Prof. Oriana Bandiera
  - 1-minute summary video: https://www.youtube.com/watch?v=pWxmcETjkPc
  - Full lecture: https://www.youtube.com/watch?v=iiYKRD8ochA
- "Reviving the American Dream and Big Data" (Raj Chetty)
  - TEDx talk: https://www.youtube.com/watch?v=u2U9-Wq2ub0