More on Specification and Data Issues

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Introduction

Most of the remaining lessons on OLS address problems with the 4th Gauss-Markov assumption (the error term's mean independence from the regressors):

$$E(u|x_1, \ldots, x_k) = E(u) = 0.$$

Examples of how this assumption can be compromised without omitting a relevant variable:

- functional form misspecifications and
- measurement error.

How functional form specifications (for *x* variables) may be tested and the properties of OLS are under measurement error.

Measurement error illustrates a more general practical problem of missing or imprecise data, and it can compromise the randomness of the sample (MLR.2). It merits attention because of this risk.

The proxy variable solution to the omitted variable problem is discussed.

- The last textbook chapter on OLS with cross-sectional data, so it serves to close that discussion.

Outline

Functional Form Misspecification.

Using Proxy Variables for Unobserved Explanatory Variables.

- Lagged Dependent Variables.

Properties of OLS under Measurement Error.

- In the Dependent Variable.
- In an Explanatory Variable.

Missing Data.

Nonrandom Samples.

Outliers and Influential Observations.

Functional form misspecification

Technically the mean independence of the error term is violated if there is a non-linear relationship, e.g., quadratic or logarithmic, between *x* and *u* that the empiricist fails to estimate.

The problem is comparatively easy to solve, though, if a bit labor intensive.

The *F* test allows the researcher to test sets of exclusion restrictions, and he could test the coefficients on several non-linear, say quadratic, *x* terms to verify whether they are statistically significant.

Functional form misspecification example (from examples 8.3 & 9.1)

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 avgsen^2 + \beta_4 ptime86 + \beta_5 ptime86^2 + \beta_6 qemp86 + \beta_7 inc86 + \beta_8 inc86^2 + \beta_9 black + \beta_{10} hispan + u,$$ where

$narr86 \equiv$ times arrested in 1986,

$pcnv \equiv$ proportion prior arrests $\rightarrow$ conviction $\approx \Pr(\text{conviction}|\text{arrest})$,

$avgsen \equiv$ (average prior sentence|conviction), months,

$ptime86 \equiv$ months in prison during 1986,

$qemp86 \equiv$ quarters employed during 1986,

$inc86 \equiv$ legal income in 1986 ($100s),

$black \equiv$ indicator for black,

$hispan \equiv$ indicator for latino.

Functional form misspecification (continued)

This model, estimated using individual level data on arrestees, enables one to test the joint significance of the three square terms embodied in the hypothesis:

$$H_0: \ \beta_3 = \beta_5 = \beta_8 = 0.$$

This algorithm for inclusion and testing of expanded non-linear terms should lead to a specification with an accurate specification of the regressors.

The eventual specification ought to satisfy the regression specification error test (RESET), though, to confirm this.

RESET

The premise for using RESET is adding non-linear functions of the regression's fitted values ($\hat{y}$) to the model because the fitted values are functions of the regressors, after all.

So one might estimate:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u, \text{where}$$

$\hat{y}$ is the fitted value from the regression of *y* only on x_1 through x_k.

If the null hypothesis that the regression is correctly specified is correct, then the square and cube of the fitted values should be insignificant and H_0: $\delta_1 = \delta_2 = 0$.

- In this example the test could be performed as an *F* statistic for those two exclusion restrictions, with joint significance indicating a model misspecification.

Functional form misspecification (continued)

The RESET does not offer much guidance about the *particular* misspecification in a model, though.

For instance, it doesn't tell you whether the specification should be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \text{or}$$

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + u.$$

Functional form misspecification (concluded)

Using the logic of the RESET, the fitted values of the incorrect specification ought to be insignificant in the correct specification, i.e.,

$$y = \beta_0 + \beta_1 x_1 \ + \beta_2 x_2 + \theta_1 \hat{y}_{log} + u,$$

should not reject the null that $\theta_1 = 0$, where $\hat{y}_{log}$ is fitted from the model estimated in logs.

If this (Davidson-MacKinnon) test passes and the alternative fails, i.e.,

$$y = \beta_0 + \beta_1 \ln(x_1) \ + \beta_2 \ln(x_2) + \theta_1 \hat{y}_{levels} + u, \text{and } \theta_1 \neq 0,$$

it is evidence in favor of the regression in levels.

- But passing each test is not mutually exclusive—nor is failing each test.
- So the outcome may not be decisive.

Using proxy variables for unobserved explanatory variables

What does an empiricist do when the relevant information that is not in the data set? This is a ubiquitous problem in economics because collecting and organizing accurate and precise data is expensive and time-consuming.

- Also many relevant variables are exceedingly difficult to measure and observe in the first place.

Nonetheless omission of relevant variables biases OLS estimates, so a solution to unobserved variables is crucial.

Sometimes one can identify a proxy variable to substitute for a relevant unobserved variable.

What is a proxy variable? The glossary in the text defines it as:

- "An observed variable that is related but not identical to an unobserved explanatory variable in multiple regression analysis."

So what are the properties of a *good* proxy variable?

Proxy variables

A proxy variable is related to the unobserved variable that it "proxies".

- The relationship can be modeled linearly like a regression with a stochastic part and a non-stochastic part.
- If x_3 is a proxy for the unobserved variable, x_3^*, the relationship might look like:

$$(1)\ x_3^* = \delta_0 + \delta_3 x_3 + v_3.$$

The "error," v_3, represents the non-identical relationship between the proxy and unobserved.

δ_3 captures the relationship between them; if $\delta_3 = 0, x_3$ is not a good proxy.

Substituting (1) into a regression of *y* on $x_1, x_2, \text{and } x_3^*$ shows how the proxy may improve estimates that would otherwise suffer from omitted variable bias.

$$(2)\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u.$$

Proxy variables (continued)

$$(1) \text{ and } (2) \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\delta_0 + \delta_3 x_3 + v_3) + u.$$

Rearranging some terms yields a new estimable model with different intercept and error:

$$(3)\ y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + (\beta_3 v_3 + u).$$

For consistent estimates of β_1 and β_2, which is usually the goal, MLR.4 must hold in (3).

The new error term, which Wooldridge calls "e", must satisfy the requirements:

$$E(e|x_1, x_2, x_3) = 0, \text{ which is satisfied if,}$$

$$E(u|x_1, x_2, x_3) = E(v_3|x_1, x_2, x_3) = 0.$$

Proxy variables (continued)

This is like saying that:

- a) the proxy would not belong in the actual regression if we could observe x_3^* (we wouldn't need the proxy if we had the "real thing") and
- b) the proxy relationship (1) correctly omits x_1 and x_2.

In the textbook example, x_3 is an intelligence quotient ("IQ"), x_1 and x_2 are education and labor force experience, and IQ proxies for unobserved "ability"—a vague and difficult to measure concept that affects both wages and education.

- So one's IQ score does not depend on education or work experience.
- Were the relationship between ability and IQ to depend on education or experience, the 2nd equality on the previous slide would not hold, and one could not estimate β_1 and β_2 consistently (or without bias).

Proxy variables (continued)

Also to economize on notation, the intercept and coefficient on x_3 are collapsed into single parameters—which are all that can actually be estimated.

$$(3) \Leftrightarrow y = \alpha_0 + \beta_1 x_1 \ + \beta_2 x_2 + \alpha_3 x_3 + e;$$

$$\alpha_0 \equiv (\beta_0 + \beta_3 \delta_0), \alpha_3 \equiv \beta_3 \delta_3, \text{and } e \equiv \beta_3 v_3 + u.$$

Performing the regression above will not yield unbiased or consistent estimates of the coefficient on x_3 ("IQ") or the intercept, but generally they aren't the variables of interest.

If IQ is a "good" proxy (as defined above), you will get unbiased and consistent estimates of β_1 and β_2.

Proxy variables (concluded)

Even if the proxy variable chosen is not perfect, the bias (inconsistency) in OLS may be better with than without the proxy.

- If IQ score increases with education and experience, it is straightforward to show the bias:

$$(1')\ x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3.$$

It makes the regression, (3),

$$(3')\ y = \alpha_0 + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \alpha_3 x_3 + (\beta_3 v_3 + u).$$

Since β_3, δ_1 and δ_2 would all probably be positive, the returns to education and experience would both still be biased upward unless $\delta_1 = \delta_2 = 0$.

- The promise of this specification lies in controlling for ability "somewhat" with an imperfect proxy.
- And making the estimates of $\delta_1$and δ_2 relatively small, reducing the error compared to the complete omission of ability.

Using a lagged dependent variable as a proxy

One clever use of proxy variables relies on the information observed by looking at the dependent variable *y* in a preceding time period.

When the outcome is correlated over time, e.g., "persistent" or "inertial," y_{t-1} can be a proxy for omitted factors that are also persistent, yet unobserved.

Adding the lagged dependent variable as a proxy enables a researcher to be agnostic about *what* the unobserved variable is exactly, as long as it persistently explains *y* across time periods, i.e., in periods t and $t-1$.

Using a lagged dependent variable as a proxy (continued)

Including y_{t-1} as a proxy controls for "whatever made *y* higher or lower last period".

- Whatever it is contributes to explaining variation in *y* in the present as well, and y_{t-1} "controls" for it.
- Conditioning on a lagged dependent variable enables you to interpret the effect of interest as follows.

In the crime rate example in the text, using$crime_{t-1}$ as a proxy for unobserved factors means that the effect of law enforcement spending is "holding previous year's crime rate constant"—which could certainly be a source of omitted variable bias (correlated with current crime rate as well as current law enforcement spending).

Properties of OLS under measurement error

Another empirical shortcoming frequently encountered is imprecisely measuredvariables.

We consider this issue successively for dependent and explanatory variables.

Measurement error in the dependent variable

Consider first a *y* variable that is measured noisily, i.e., with a stochastic error ("e_0") around the measured values.

$$e_0 \equiv y - y^*; y^* \text{is the right value, and } y \text{ is observed.}$$

The "right" value, y^*, is what the economic agents actually act on, e.g., how many hours they work per week, their gross annual income, how many car trips they make per week.

It is observed with measurement error in the data, nonetheless y^* is the basis for the population model:

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

Measurement error in *y* (continued)

Using the observed value instead is equivalent to substituting in

$$y^* = y - e_0,$$

which adds to the error term:

$$y - e_0 = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$$

$$\Leftrightarrow y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0.$$

This model is estimable, and it still consistent as long as e_0 satisfies some conditions.

- Namely it must be mean independent of the regressors, i.e., it must satisfy the same requirements that *u* does, since it is now in the error term.

Measurement error in *y* (concluded)

If the two parts of the new error ($u + e_0$) term are uncorrelated (a common assumption), the error now has larger variance:

$$\sigma_u^2 + \sigma_e^2 > \sigma_u^2,$$

so the variance of the estimators is larger than in the absence of measurement error.

If e_0 is uncorrelated with the regressors but has non-zero mean, it only biases the intercept.

The assumption that the measurement error is independent of the regressors is not always valid, though, and in those instances OLS is biased.

Measurement error in an explanatory variable

Were the measurement error problem to afflict an explanatory variable instead, the model would look as follows:

$$y = \beta_0 + \beta_1 x_1^* + u; x_1^* = x_1 - e_1, \text{where}$$

e_1 is the measurement error and $E(e_1) = 0$.

Performing the regression (which is assumed to satisfy Assumptions MLR.1 through MLR.4) using the noisy measure entails estimating:

$$y = \beta_0 + \beta_1(x_1 - e_1) + u = \beta_0 + \beta_1 x_1 + u - \beta_1 e_1.$$

Measurement error in an *x* variable (continued)

As with the dependent variable case, regardless of whether "the noise is uncorrelated with the signal," the variance is larger with measurement error as well as the OLS standard errors:

$$\sigma_u^2 + \sigma_{e1}^2 > \sigma_u^2;\ u \text{ and } e_1 \text{are uncorrelated.}$$

If the noise is uncorrelated with the signal ($Cov(x_1^*, e_1) = 0$), the noise (measurement error) has to be correlated with the noisy measure.

If

$$x_1 = x_1^* + e_1, \text{and } Cov(x_1^*, e_1) = 0,$$

$$Cov(x_1, e_1) = E(e_1 x_1^* + e_1^2) = 0 + \sigma_{e1}^2.$$

Measurement error in an *x* variable (continued)

This is called the classical errors-in-variables (CEV) scenario.

- And it produces negative covariance between the noisily-measured variable and the error term (as well as bias and inconsistency).

$$Cov(x_1, -\beta_1 e_1) = -\beta_1 \sigma_{e1}^2, \text{but}$$

since the inconsistency contains the coefficient, β_1 can be factored out and the inconsistency expressed as a multiplicative constant (in the limit).

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1 - \frac{\beta_1 \sigma_{e1}^2}{\sigma_x^2} = \beta_1 - \frac{\beta_1 \sigma_{e1}^2}{\sigma_{e1}^2 + \sigma_{x*}^2} = \beta_1 \left(\frac{\sigma_{x*}^2}{\sigma_{e1}^2 + \sigma_{x*}^2} \right) < \beta_1.$$

Measurement error in an *x* variable (concluded)

So the consequence of CEV is a downward (toward zero) bias in the estimate of β_1, which is known as attenuation bias.

- Measurement error in one variable biases the estimates of the other regressors' coefficients as well (except under very strong assumptions).

Finally the CEV scenario will often not hold.

- Suffice it to say that measurement error that is correlated with the true value also presents problems for OLS estimates, but showing their properties is more difficult than is appropriate for this class.

Whether measurement errors obey the CEV or not, there are (non-OLS) methods to solve the problem. These must be postponed until a later chapter, though.

Missing data

The reality of empirical work in economics is that most data sources contain gaps (missing observations) for some variables of interest.

- At first this doesn't sound like a big deal, but since the OLS minimand* is a mathematical function of all the variables for all observations, a missing value negates the whole observation.

Stata records missing values in its Data Editor with a ".".

- It automatically excludes any observations with a missing value of a variable in the regression you ask it to estimate.

If the nature of the missing data is random, this merely reduces the sample size, along with the problems for inference that come with a smaller (but still representative and random) sample.

*Sum of squares of errors.

Nonrandom samples

Perhaps surprisingly sampling that depends on the regressors (either because of missing data or convenience sampling) is exogenous and creates no problems for Assumption MLR.2.

- Sampling probability that depends on an element's value of an explanatory variable is not ideal, but the estimates are still unbiased and consistent because the regression model, once specified, holds for any subset of the population of interest.

The same is not true if sample selection is related to elements' values of the dependent variable: a case of endogenous sample selection.

This will lead to bias in the OLS estimates because:

$$E(y|\boldsymbol{x}) \neq E(y|\boldsymbol{x}, y^*), \text{where } y^* \text{ determines sample selection probability.}$$

Nonrandom samples (continued)

A common example, similar to the one in the textbook, is estimating labor supply elasticity.

- For employed individuals ($y > 0$), the wages are observed, but for non-employed individuals ($y = 0$) the wage they *would* earn is not observed (certainly not by an empiricist).
- Labor supply has an intensive and an extensive margin. And excluding the missing observations amounts to conditioning on people who have either comparatively high wage offers, low reservation wages or both.
- The basic problem is that labor supply elasticity among those observed working is unlikely to represent the elasticity for the whole population.

Once again there are methods that deal with selection issues like this, but they are beyond the purview of OLS and not discussed here.

Outliers and influential observations

OLS is sensitive to extreme values of one or more variables (outliers) because it is based on variances and covariances, and a single (few) "large" observed deviation(s) from a variable's mean can dominate these calculations.

When excluding one or several such observations changes the OLS estimates in practically significant ways, a researcher should be concerned about those observations.

This is particularly dangerous in a small sample.

Outliers and influential observations (continued)

This topic is contained under the chapter on “Data Issues” because very often outliers result from mistakes in data recording or decoding.

Examples:

- 1) data entry is susceptible to typographical errors, such as hitting a key twice or entering a “7” instead of a “1”,
- 2) consider the question from the NSCW* below about years of labor force participation.

Years of experience is recorded as the actual number if observed, but a code (998 or 999, respectively) is used if the respondent “doesn’t know” or refuses to answer.

- If labor market experience is a regressor and includes several (technically missing) observations, they will immediately turn into large outliers: 998 or 999 years of experience!

*National Study of the Changing Workforce.

Question from the NSCW

Q660/QEB39 And how many years in total have you done any work for pay since you were 18 years old -- including part-time and full-time jobs?

(INTERVIEWER NOTE: PROBE: Your best estimate will be fine.)
(INTERVIEWER NOTE: ENTER NUMBER OF YEARS, ROUND FRACTIONS OF 1/2 YEAR (6 MONTHS) OR MORE UP TO THE NEXT YEAR, ROUND FRACTIONS LESS THAN 1/2 DOWN: LESS THAN 1 YEAR ALWAYS = 0) (INTERVIEWER: ENTER 998 FOR DK & 999 FOR REF.)

|_|_|_|
(248, 250)

[RANGE: 0 - 99, 998, 999]

CLEAN YOUR DATA!!!

The life-saving lesson from this is that the first thing you must do with data is "clean it up":

- identify outliers by looking at the summary statistics,
- figure out why you have outliers by using the codebook*, and change them to missing or correct them as the situation warrants.

Do not risk the disgrace of estimating and presenting results based on erroneous ("unclean") data!!! This is the equivalent of a surgeon not washing his hands before an operation.**

*Many surveys have a separate (sometimes lengthy) downloadable document explaining the coding of each variable at what survey question it is based on. This is called the codebook.

**For more read: Hamermesh, Daniel S. 2000. "The Craft of Labormetrics." *Industrial and Labor Relations Review*, Volume 53, Number 3: 363-380.

Outliers and influential observations (continued)

You may also encounter examples of outliers that have nothing to do with datacoding.

- Maybe they are correctly observed but pulled from a distribution that is prone to extreme values.
- Maybe the observations do not really belong to the population of interest.

What to do about cases like that? Depends on project-specific factors:

- how much the estimates change when the outliers are included (excluded),
- what the convention in the related literature is, et al.

Depending on these factors, the outliers may be included in or excluded from estimation.

Outliers and influential observations (concluded)

In STATA the leverage of each observation can be computed after a regression by typing:

predict [newvar], hat.

And an instructive graphical diagnosis of the role of outliers comes from plotting the leverage ("how far each observation is from the means") against the squares of the residuals.

This is accomplished in STATA by typing **lvr2plot** after running a regression.

- The idea is to identify observations that have a lot of leverage *and* large residual, as well as those that have a lot of leverage and a small residual.
- The former may simply not be appropriate for including in the population of interest (unusual and model predicts poorly), and the latter may exert undue influence on the estimates and excluding them would be useful to see if the results hold in their absence.

Conclusion

This is only an overview of the issues involved in turning a raw data set into a clean, useable product.

For the most part the exercises in this class use very clean data with few or no major issues.

We could spend a whole semester studying data issues and what can be done about them, but that is not the main purpose of this course.

In your daily life as an empiricist, however, these issues will occupy more of your time than running regressions and presenting the results.

This is the underappreciated “behind-the-scenes” work that is a prerequisite for a high quality research project.

Lags and leads

The terminology, “lagged,” refers to the same variable observed in a pastperiod.

A lagged variable is usually denoted with a subscript, e.g., $t-1$, to distinguish it from the current period observation (which is subscripted with a t).

The analogous term for a future period is called a “lead”.

The chapters on time series regressions will elaborate on this in more detail.

Back.

Selection probability

Wooldridge Chapter 17 has a nice model of sample selection, i.e., the observation is in the data set if all its values are observed for all relevant variables:

$$s_i = \begin{cases} 1 \text{ if all observed} \\ 0 \text{ otherwise.} \end{cases}$$

The whole model gets interacted with this binary variable when performing the regression calculations (variance, covariance), dropping out the observations for which $s_i = 0$.

$$y_i s_i = \beta_0 s_i + \beta_1 s_i x_i + s_i u_i$$

Selection probability (continued)

The analog of Assumption SLR.4, here, is that the expectation of the error term, conditional on *x*, is zero:

$$(A1)\ E(u_i s_i | x_i) = 0,$$

where you can assume *u* is a well-behaved error that satisfies SLR.4 ($E(u|x) = 0$).

So satisfying (A1) comes down to whether

$$E(u|x, s = 1) = E(u|x, s = 0) = E(u|x).$$

If the first equality does not hold, you have endogenous sample selection and a bias in the estimates of the βs.

Back.

Leverage-residuals plot