Causal Inference with Observational Data

7. Tricks of the Trade I

Marc F. Bellemare

May 2018

Introduction

This deck of slides and the next are a loose collection of bits of applied econometric knowledge I have acquired over the years, and which do not fit neatly into one of the previous chapters.

My hope in teaching you those tricks is that you will be able to use them in order to show your readers (in particular, your reviewers) that you know what you are doing, thereby maximizing the probability that your manuscripts will land in good journals.

Measurement Error in a Binary Treatment Variable

It is not uncommon for a binary treatment variable to be measured with error—this is what we call misclassification.

In Bellemare et al. (2015), for example, we are looking at the relationship between whether a woman has undergone female genital cutting (FGC) and whether she supports the practice of FGC, and there is some evidence from the field of public health that physical examinations and self-reporting yield different rates of FGC prevalence (our treatment variable) and that institutional features such as bans on FGC lead to misreporting of FGC.

Measurement Error in a Binary Treatment Variable

In that paper, we are very careful to always talk of the relationship between one *reporting* having undergone FGC and one's *reported* support for FGC, but it is possible to do better if you have a good idea of how much misclassification there is in either direction, i.e., true zeroes reported as ones (call this proportion p), and true ones reported as zeroes (call this proportion r).

Measurement Error in a Binary Treatment Variable

Bollinger (1996) came up with a method to calculate bounds on γ in the regression

$$y_i = \alpha + \beta x_i + \gamma D + \epsilon_i \tag{1}$$

when D is dichotomous and suffers from misclassification.

In short, suppose you get a coefficient estimate $\widehat{\gamma} > 0$ from a regression of y on misclassified D. Bollinger's method allows you to put bounds a and c on $\widehat{\gamma}$ such that $a < \widehat{\gamma}$ and $c > \widehat{\gamma}$. Better yet, if you actually know (or have a good idea of) the rates of misclassification, you can get even better (i.e., tighter) bounds—something like d and f, where $a < d < \widehat{\gamma} < f < c$.

Goodness of Fit in Binary Choice Models

In econometrics, goodness-of-fit measures tell us what percentage of the variation in a dependent variable is explained by the explanatory variables.

If you've ever taken a statistics class, you are almost surely familiar with the R^2 measure. In a regression of, say the logarithm of wage on age, gender, and education level, the R^2 is simply the fraction of the total variation in wage that is explained by variation in age, gender, and education level.

Goodness of Fit in Binary Choice Models

Given the foregoing, you'd think R^2 is a great measure, since it tells you how much of the variation in y all of the variables x explain.

In truth, R^2 is actually not all that interesting, because you can throw in any variable on the right-hand side—for example, the color of one's eyes in the log wage regression above—and R^2 can only increase, because there is bound to be a (spurious) correlation between the color of one's underwear and one's wage.

Even the adjusted R^2, which corrects for how many variables there are in x, isn't that great, since that correction is somewhat arbitrary.

Goodness of Fit in Binary Choice Models

With binary outcomes, people often like to use the percentage of ones and zeroes correctly predicted, and report that as a measure of goodness of fit. Kennedy, in his classic econometric treatise, argued that this was not a very good measure:

> *It is tempting to use the percentage of correct predictions as a measure of goodness of fit. This temptation should be resisted: a naïve predictor, for example that every* $y = 1$*, could do well on this criterion. A better measure along these lines is the sum of the fraction of zeros correctly predicted plus the fraction of ones correctly predicted, a number which should exceed unity if the prediction method is of value. See McIntosh and Dorfman (1992).*

Goodness of Fit in Binary Choice Models

This could use a bit of explanation: Suppose we have $y = (0, 1, 1, 1, 1, 1, 1, 1)$, and we have a vector of predicted values of $\widehat{y}$ be $(0, 1, 1, 0, 1, 1, 1, 0)$.

The usual percentage-of-correct-predictions measure would be 0.75, since 75% of observations are correctly predicted. But one can do even better by guessing "all ones."

Indeed, if you were to guess all ones, you'd get 87.5% goodness of fit!

Goodness of Fit in Binary Choice Models

What McIntosh and Dorfman (1992) suggest instead is to add up (i) the fraction of correctly predicted zeroes (in the example above, 100%) and (ii) the fraction of correctly predicted ones (in the example, 50%).

In that example, then, the total McIntosh-Dorfman goodness-of-fit measure would be 1.5 which, by McIntosh and Dorfman criterion standards, would be deemed a good fit, since it exceeds 1.

Goodness of Fit in Binary Choice Models

In a referee report I received a few years ago, a reviewer was faulting me for a low pseudo R^2 measures on a probit, and suggested that I report the percentage of correct predictions.

Notwithstanding the fact that pseudo R^2 measures are pretty bad (see Estrella, 1998), I responded with the Kennedy quote above, and in the published version, I actually report three measures: the pseudo R^2 (0.081), the percentage of correct predictions (0.63), and the Dorfman-McIntosh measure (1.29).

The Use and Misuse of R-Square

Table 5. *Probit estimation results for the first stage of the treatment regressions*

Variable	Marginal effect	(Std. Err.)
Dependent variable: = 1 if participates in contract farming; = 0 otherwise		
Household size	0.025	(0.021)
Dependency ratio	−0.132	(0.214)
Single	0.068	(0.201)
Female	−0.449*	(0.236)
Migrant	0.066	(0.138)
Age	−0.021***	(0.007)
Education	−0.005	(0.014)
Experience	0.013*	(0.007)
Member of peasant organization	0.546***	(0.110)
Fady days	−0.003*	(0.002)
Working capital	0.005	(0.004)
Assets	0.002	(0.002)
Landholdings	0.001**	(0.000)
"Yes" to $12.5 investment	0.382***	(0.148)
"Yes" to $25 investment	0.406***	(0.140)
"Yes" to $37.5 investment	0.454***	(0.137)
"Yes" to $50 investment	0.539***	(0.148)
"Yes" to $62.5 investment	0.326*	(0.192)
"Yes" to $75 investment	0.727***	(0.181)
Intercept	0.260	(0.285)
Number of observations	1178	
District fixed effects	Yes	
F-statistic (instruments)	24.55	
p-Value (joint significance, all coefficients)	0.000	
Goodness of fit measure (McIntosh & Dorfman, 1992)	1.29	
Percentage correct predictions	0.630	
Pseudo *R*-square	0.081	

Note: These estimation results correspond to Eqn. (3) in the body of the paper. Estimation results are probability-weighted.

* Significance at the 1% levels.

** Significance at the 5% levels.

*** Significance at the 1% levels.

The Use and Misuse of R-Square

A few years ago at a conference I attended, a presenter talked about paper in which he had run a randomized controlled trial to determine the effect of a treatment variable D on an outcome y, randomizing D and collecting information on a number of control variables x in addition to collecting information on y.

When presenting his results, the presenter did what we commonly do in economics, which is to show a table presenting several specifications of the regression of interest, from the most parsimonious (i.e., a simple regression of y on just D) to the least parsimonious (i.e., a regression of y on D and all the available controls x).

The Use and Misuse of R-Square

The problem, however, was that the R^2 measure—the regression's coefficient of determination—for the simple regression of y on just D (i.e., the most parsimonious specification) was about 0.01, meaning that the treatment variable D explained about 1 percent of the outcome of interest.

The Use and Misuse of R-Square

This is interesting, given that if policy makers make a big deal about the relationship between D and y, one of the findings of the paper should be that policy makers should really spend their time on other things.

Indeed, if D explains only 1 percent of the variation in y, focusing on D in order to stimulate y is unlikely to be cost effective.

In other words, there are other factors out there that explain 99 percent of the variation in y, and it is likely that among those factors, at least one or two will play a significant role—or at least, a role that is much more important than D.

p-Values Are Thresholds, Not Approximations

I was once working with a grad student. The two of us were running some rough-cut regressions, taking a first stab at some data we had just received.

As is often the case, we realized we had to cluster our standard errors at the relevant level. So we did that, and the coefficient of interest, which had hitherto been significant, now had a p-value of 0.102 because of the clustering.

p-Values Are Thresholds, Not Approximations

The grad student said: "Well it's significant, but only barely." I asked the grad student to explain her reasoning, because I was curious about what she saw that I wasn't seeing.

She then said "Well, the *p*-value rounds down to 0.10, doesn't it?"

It was then that I had to tell the student that *p*-values are thresholds, not approximations, and that if a *p*-value is greater than 0.10, then the estimate is not significant at any of the conventional levels. Likewise, if a *p*-value is 0.051, the estimate is only significant at the 10 percent level, no matter how you want it to be significant at the 5 percent level.

p-Values Are Thresholds, Not Approximations

Relatedly, there is such a thing as *p*-hacking, or data dredging, the phenomenon whereby "the use of data mining to uncover patterns in data that can be presented as statistically significant, without first devising a specific hypothesis as to the underlying causality."

A recent article (Brodeur et al., 2016) shows that although we should expect the distribution of *p*-values reported in economics journals to be smooth around the critical thresholds of 0.10, 0.05, and 0.01, it turns out that they are anything but smooth—there are serious discontinuities around those thresholds, which indicates that there is a significant amount of *p*-hacking going on in our discipline.

p-Values Are Thresholds, Not Approximations

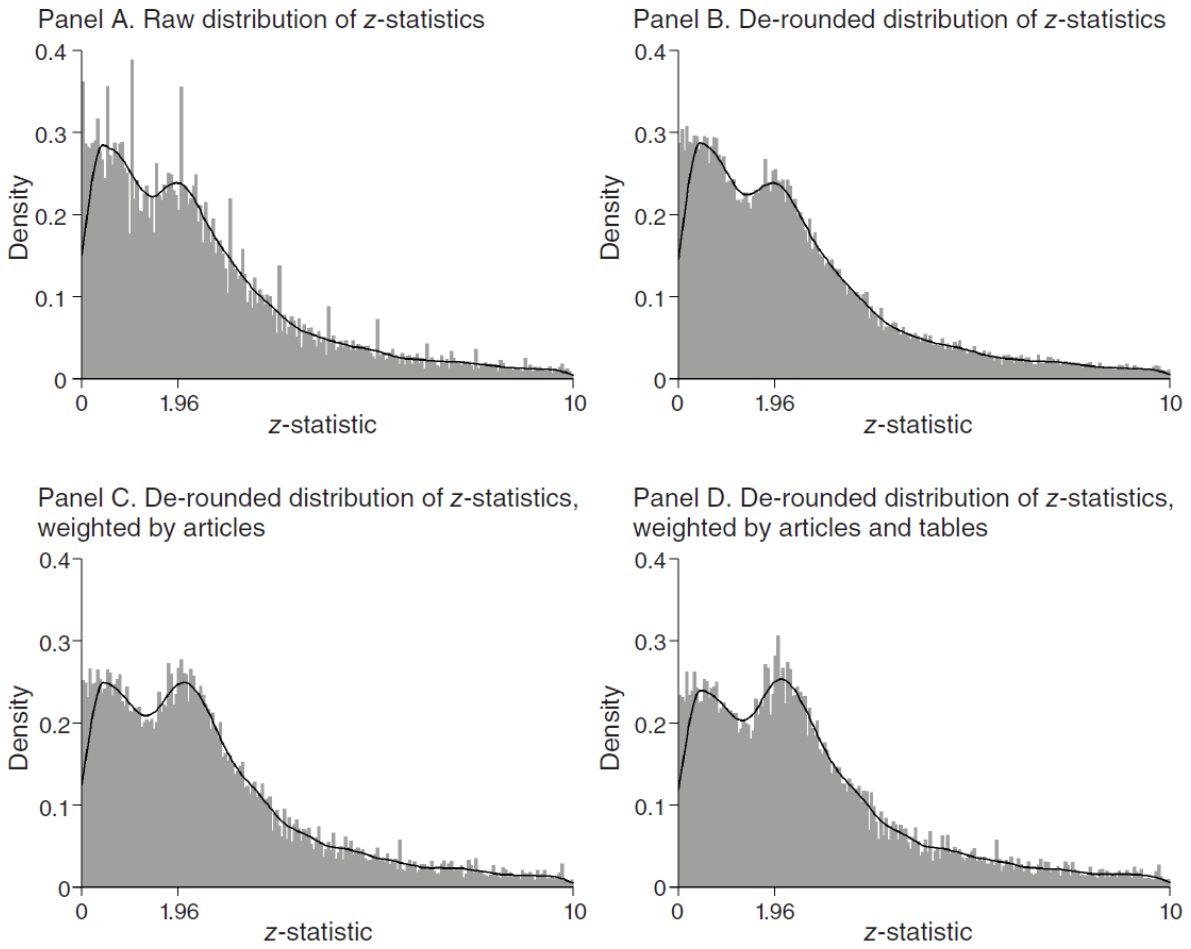

FIGURE 1. DISTRIBUTIONS OF *z*-STATISTICS

Notes: See the text for the de-rounding method. The distribution presented in the subfigure C uses the inverse of the number of tests presented in the same article to weight observations. The distribution presented in subfigure D uses the inverse of the number of tests presented in the same table (or result) multiplied by the inverse of the number of tables in the article to weight observations. Lines correspond to kernel density estimates.

Source: American Economic Review, Journal of Political Economics, and Quarterly Journal of Economics (2005–2011)

"Do Both"

One of the questions I often have to answer is of the form: "Should I do A or B?" Specifically, questions like

- Should I estimate a linear probability model, a probit, or a logit?
- Should I use sampling weights or not?
- Should I cluster my standard errors or not?
- Should I take the logarithm of my dependent variable, or just use its level?

"Do Both"

- Should I estimate my spline regression with three, four, five, or more knots?
- Should I estimate this in level or in first differences?
- Should I express my variables in per capita terms, or just include them as is and control for population size?

"Do Both"

Almost always, the question is asked as though there is a single answer.

But this is where economics becomes more art than science, more rhetoric than dialectic, and where students have to learn that there is more than one way to skin a cat.

"Do Both"

In applied work, especially in applied work relying on observational data (a second-year paper rarely allows one the time to collect one's own experimental data), the key to convincing your readers that x causes y is to show that your core result holds over and over, no matter how you slice the data, and to show that if there are some cases where x does not seem to cause y, you have a good story for why that is the case.

As such, sky is the limit, and you should always have an appendix that is not for publication, but just for the reviewers, in which you present all those extra results.

What to Do with Endogenous Controls?

Suppose you have observational data, and you are interested in estimating the causal effect of your variable interest D on your outcome of interest y, and you also have access to a vector of control variables x. For the sake of argument, let's assume there is only one control variable in the equation

$$y_i = \alpha + \beta x_i + \gamma D + \epsilon_i \tag{2}$$

The parameter of interest is γ. If you have observational data, then you know that in most cases $E(D'\epsilon) \neq 0$, i.e., D is endogenous to y, and γ does not capture the causal effect of D on y.

What to Do with Endogenous Controls?

But what about x? It often happens that x is also obviously endogenous to y—say, because x is a decision variable which is determined by each individual respondent's expectation of y, which would constitute a case of reverse causality.

What to Do with Endogenous Controls?

In terms of the peer-review process one thing I would not encourage you to do is to try to find an instrumental variable for x.

Why is that? To put it simply (if a bit cynically): Because D is your variable of interest, and it is difficult enough to deal with the fact that D is endogenous—that is, how well you do so will determine how well your paper is received by reviewers and editors—that attempting to deal with the endogeneity of your control variable exponentially expands the number of reasons why your reviewers might recommend that your paper be rejected.

What to Do with Endogenous Controls?

I still sometimes see papers where the authors are looking at the effect of some variable of interest D on some outcome of interest y, but where they spend a considerable amount of time trying to deal with x.

But that is really besides the point, because it is D that is the variable of interest, not x.

What to Do with Endogenous Controls?

So how do we deal with endogenous controls? First, let's think about what an endogenous controls means: An endogenous control x means that $E(X'\epsilon)$ is different from zero, which obviously means that the estimated β in equation 1 will be biased.

What to Do with Endogenous Controls?

An endogenous control x also means that the OLS estimator for γ—the parameter of interest—will be biased, since x appears in the formula for the OLS estimator of γ.

Moreover, Frölich (2008) discusses how both OLS and 2SLS will be inconsistent in the presence of endogenous controls. That is, they do not converge to the true value of the parameter of interest.

Excluding the endogenous control x means that x is now in the error term ϵ, and so if x is correlated with D, then your estimate of γ is also biased.

What to Do with Endogenous Controls?

This suggests the following: If D and x are uncorrelated, then it is better to leave x out of your regression altogether, because in that case, it does not bias your estimate of γ, no matter how much variation in y is explained by x.

What to Do with Endogenous Controls?

If D and x are correlated, then you have a problem either way.

Omitting x means that you have an omitted variable bias. Including it means that your estimates are inconsistent.

What should you do? The middle-of-the-road approach is the usual "do both," that is to present results both with and without the endogenous control, and see what changes. But even that is not terribly satisfactory, since there is bias in both cases, and "get a better research design" is even less helpful.

What to Do with Endogenous Controls?

Ideally, you would find a good (i.e., valid and relevant) IV for x, but those are difficult to find, and if the IVs used for endogenous variables of interest D in the papers I have seen trying to tackle the of endogenous controls x were usually not the best, the IVs used for those endogenous controls were even worse.

What to Do with Missing Data?

Suppose you observe D for everyone in your sample, but you have missing data for x. What should you do? Here are a few options:

- Ignore the problem. With missing data, there is an implicit assumption that is made when you ignore the problem, viz. that data are missing at random. If you are going to ignore the problem, you should think carefully about whether data are likely to be missing at random.

What to Do with Missing Data?

- Run a balancing test. If you want to have an idea of how missing data may bias your sample, you can also run balancing tests. That is, use a t-test to compare the mean of y for those observations with missing x versus those observations with x present, and do the same for D. If you fail to reject the null hypotheses that (i) the mean of y is equal for those with x and those with missing x, and (ii) the mean of D is equal for those with x and those with missing x, you can be a bit more confident that your missing values for x appear to leave the sample intact. If you find, say, that there are systematic differences in some variable between those with x and those with missing x, that tells you how those missing values might bias your sample.

What to Do with Missing Data?

- Run the sub-regression $y_i = \alpha + \gamma D + \epsilon_i$ with and without those observations for which x is missing. Is γ roughly the same across samples? If so, then that is an additional reason not to worry about missing values for x, given that γ is the parameter of interest. Of course, if you have missing values for D, that is a different problem.

What to Do with Missing Data?

- Use "missing dummies" to keep those observations. You can create a dummy variable–let's call it z–equal to 1 if x is missing and equal to zero otherwise. Then, create a variable x' equal to x if x is nonmissing and equal to zero otherwise, and estimate $Y = \alpha + \beta x' + \zeta z + \gamma D + \epsilon$. This has the advantage of retaining all observations. This is something a reviewer once asked me to do, and though it feels like a bit of a kludge, it is fine when presented alongside the results of a regression where you treat the missing values of x as missing at random.

What to Do with Missing Data?

- "Do both." By now, you know this is pretty much my mantra when it comes to applied econometrics, which is more like rhetoric than dialectic, and in which you need to show that your finding holds over and over in different specifications, building your case for it like a lawyer would build his client's case in court. So don't be afraid to do all of 1 to 4 above.
- Another thing you can do is to impute those missing values. That is, regress x on D and get the predicted values of x, i.e., $\widehat{x}$, and replace missing values of x with the $\widehat{x}$ values. This also feels like a bit of a kludge, but when used with other methods, and not as your only solution, it should be all right.

What to Do with Missing Data?

- Finally, should you be lucky enough to have an instrumental variable that (i) is relevant, i.e., it is correlated with missing values, and (ii) is valid, i.e., it only affects y through x, you can try to estimate a 2SLS or selection correction model, but this seems like a lot of work, and it is rare that we have a good IV for D, not to mention for x.
- Get better data.

What to Do with Missing Data?

The foregoing presupposes that you have a sizable proportion of your sample with missing x.

If you only have five cases where x is missing out of 500 observations, I don't think anyone will seriously mind if you treat those missing values as missing at random.

But if, say, more than 5% of your sample is missing, you might want to run through the list above–and even that is an arbitrary rule of thumb. The best thing to do, as always, is to be forthcoming about the problem, explore how it might compromise (i.e., bias) your results, and try to show robustness as best as you can.

Pesky Proxies

It often happens in the course of doing empirical work that we wish study the relationship between some variable of interest D and some outcome y, but that we don't have access to a good measure of D.

Rather, what we have instead is a proxy for D, which Wiki defines as "a variable that is not in itself directly relevant, but that serves in place of an unobservable or immeasurable variable.

In order for a variable to be a good proxy, it must have a close correlation, not necessarily linear or positive, with the variable of interest."

Pesky Proxies

For example, we may observe a dummy variable for whether one has started a business as a proxy for entrepreneurial ability.

Or we may observe one's IQ as a proxy for intellectual ability. Or we may observe the frequency of elections as a proxy for democracy.

The possibilities here are endless.

Pesky Proxies

For the sake of argument, then let's denote our proxy variable—what we actually observe in lieu of D—as D^*, so that

$$D^* = f(D) + u, \tag{3}$$

where $f(\cdot)$ is a mapping from D to D^* and u is some kind of error term to make the relationship between D and D^* stochastic, for if that relationship were deterministic and D^* were equal to $f(D)$, then observing the proxy D^* would be as good as observing the variable of interest D.

Pesky Proxies

Our ideal goal is to estimate the coefficient γ accurately in the usual regression but the best we can do is to estimate

$$Y = \alpha + \beta x + \gamma f(D) + \epsilon^* \tag{4}$$

We now have to contend with u being in the error term $\epsilon^* = \epsilon + u$, and so if u is correlated with any of the variables on the right-hand side, then we are dealing with an endogeneity problem.

Pesky Proxies

In the best-case scenario, u is uncorrelated with the variables on the right-hand side, but that isn't always the case, and it isn't even clear that this is frequently the case.

And then there care cases where the variable that you use as a proxy really does not have a monotonic relationship with y, and in which case any statistical test related to γ is unidentified because you don't know what to test for.

In Bellemare and Brown (2010), we showed that using income or wealth as a proxy for risk aversion in a test of risk sharing leads to a test that is unidentified, which invalidated the widespread use of income or wealth as a proxy for risk aversion in the applied contract theory literature.

Outliers

Outliers cause estimation problems because they bias point estimates.

They cause inference problems because they cause standard errors to be too large, thereby making it more likely that one will fail to reject a false null, i.e., a type II error.

For example, if you collect data on a random sample of the population, the bulk of the people in your data might be between 18 and 80 years old, but you might also have someone in there who is 110 years old, that person is an outlier.

Outliers

Important distinction: outliers vs. leverage points.

An outlier is an observation whose residual is significantly larger than that of other observations (i.e., an outlier is measured along the y-axis).

A leverage point is an observation that has an exceedingly low or large value of an explanatory variable (i.e., a leverage point is measured along the x-axis).

Outliers

The issue with outliers and leverage points is that they can drive your results. Usually, the best way to detect influential observations is exploratory data analysis—plot the data and see whether there are such observations.

If there are, Kennedy advises taking a look at each such observation, and try to determine whether it has a story to tell (e.g., a household may report a yield of zero because lightning fell on its plot and burned the entire crop), or whether it looks like an error (e.g., a typo in data entry, or a respondent trolling the enumerator). When an observation is influential because it looks like an error, it is reasonable to throw it out.

Outliers

If you keep those influential observations (say, because they have a story to tell), Kennedy suggests five different "robust" estimators in his book, including M-estimators, which assign weights to each observation that are not increasing in their error (OLS weights each observation in an increasing manner as it moves away from the average because it squares each error).

Outliers

What I have done in my own work has been one of a few things:

- Estimate a median regression version of my regression of interest, which estimates the median instead of the mean regression slope, the median being less sensitive to outliers than the mean. It's what I have done, for example, in this article on whether mobile phones are associated with higher prices for farmers.

Outliers

- Estimate a number of other robust specifications, e.g., *M*-estimators, *MM*-estimators, *S*-estimators, and *MS*-estimators. Vincenzo Verardi has done a bunch of work on outliers, and he has written a Stata add-on command to estimate those.
- Adopt a rule of thumb for deletion of outliers—say, drop all observations that are more than 2, 2.5, or 3 standard deviations from the mean of each explanatory variable—and re-estimate everything.

Outliers

Ultimately, what you should aim for is to show what happens across a number of estimators, i.e., OLS with outliers arbitrarily removed, robust $M/S/MM/MS$ estimators, median regression, OLS with rule-of-thumb deleted observations, etc.

If your core results are essentially the same in sign and significance across all specifications, then you should be good to go.

Data Cleaning

Many textbooks now come with a number of data sets that readers can use to apply various techniques and replicate the examples in the book (for example, Wooldridge's textbook), which is great.

The problem with those data sets is that they are "perfect." That is, no data are missing, no values are the product of an obvious typo, all the data are in one neat file, and so on.

Data Cleaning

Very often, however, the data you will want to use for a research project is not clean. It will come in several files covering different questionnaire modules across different years. Monetary values will have been recorded in nominal terms. Some people will have refused to answer some questions; others will have trolled the enumerators with crazy answers. Whoever entered the data will have made typos.

Data Cleaning

The list of issues is almost endless, and each data set has its unique set of data-cleaning issues, which is why it is very difficult to actually teach students how to clean data. But if there is one thing that you need to remember on the data-cleaning front, it's this:

Document everything.

Data Cleaning

Cleaning data will typically involve running a program file wherein

- You merge different data files together. This can range from easy if you only have to match observations with themselves (i.e., individuals' answers to demographic questions with the same individuals' answers to financial questions) to very tricky if you have to ascribe several sub-observations (e.g., a household's individual plots) to one "master" variable (e.g., the household itself), and you might want to check that step several times over to make sure everything is okay, going so far as inspecting a few observations to see if they line up with the actual values recorded in the survey questionnaire.

Data Cleaning

- You tab each variable to see whether there are obvious irregularities: missing values, outliers, censoring, truncation, etc. For cases where you have several sub-observations per unit (say, several country-year observations), you might want to check that the time-invariant values are indeed time-invariant, checking the mean of those variables by country. Here, you might also want to plot your dependent variable against each right-hand side variable, just to get a visual sense of what is going on as well as to detect outliers and leverage points.

Data Cleaning

- You drop some observations because of missing values, outliers, typos, etc.
- You transform some variables by taking a log, applying an inverse hyperbolic sine transformation, expressing them in real terms, converting two-week recall into seasonal data, dividing by 1,000 to have estimated coefficients more in line with your other estimated coefficients, and so on.

Data Cleaning

- You generate new variables from those you currently have, whether this means adding variables together (e.g., to calculate household size), creating dummies from continuous variables (e.g., to break up income into income brackets), creating ratios of two variables (e.g., to use firms' price-earnings ratios as a regressor), etc.
- You perform other operations that will lead to a nice, clean data set you can just run a parsimonious estimation program file on.

Data Cleaning

So what I suggest—and what I try to do myself—is to write a program file that begins by loading raw data files (i.e., Excel or ASCII files) in memory, merges and appends them with one another, and which documents every data-cleaning decision via embedded comments (in Stata, those comments are lines that begin with an asterisk) so as to allow others to see what assumptions have been made and when.

This is like writing a chemistry lab report which another chemist could use to replicate your work.

Data Cleaning

Another important rule is to never, ever save over (i.e., replace) a data file.

If you replace a data file from which you have dropped something or in which you have transform the data in some irreversible way (say, because you failed to follow the "Document everything" rule and did not document what you did to the data), then that file is gone forever.

Multicollinearity

Suppose you have the following regression model

$$y_i = \alpha + \sum_{j=1}^{K} \beta_j x_{ji} + \epsilon_i. \tag{5}$$

You have N observations which you use to estimate the regression. If $N < K$, you will not be able to estimate the vector of parameters β because you have fewer equations than you have unknowns in your system—recall from your middle-school algebra classes that you need at least as many equations as you have unknowns in order to solve for those unknowns.

Multicollinearity

So in econometrics, $N < K$ means that you cannot "solve" for β (i.e., it is under-determined), $N = K$ means that your equation has a unique solution for β (i.e., it is exactly determined), and $N > K$ means that your equation has several solutions for β (i.e., it is over-determined).

Multicollinearity is the problem that arises when N is too small relative to K, or what Arthur Goldberger called "micronumerosity," referring to too small a number of observations relative to the number of parameters. The most extreme version of multicollinearity is $N < K$, in which case you cannot estimate anything.

Multicollinearity

A less extreme version of multicollinearity is when there is an exact linear relationship between two variables.

Suppose x_1 and x_2 above are respectively dummy variables for whether one is male and whether one is female.

Barring the unlikely case where the data include one or more intersex individuals, trying to estimate the equation will lead to one of the two variables being dropped, simply because $x_1 + x_2 = 1$, i.e., there is an exact linear relationship between the two. If you were to try to "force" that estimation, your statistical package would not be able to invert the matrix $X'X$ necessary to estimate β by least squares, and the only way to include both variables would be to estimate the equation without a constant.

Multicollinearity

The more common version of the multicollinearity problem is when the correlation between two or more variables is "too high," meaning that there is an approximate linear relationship between those variables.

A good example would be between the amount of food one purchases which one consumes, the amount of food one purchases which one wastes, and the total amount of food one purchases. Food consumed and food wasted need not sum up to one's total food purchases–sometimes one gives food to someone else–but the correlation is high.

Multicollinearity

When that happens, the OLS estimator is still unbiased, and as Kennedy (2008) notes, the Gauss-Markov theorem still holds, and OLS is BLUE.

Rather, the problem is that the standard errors blow up, and β is imprecisely estimated, and so hypothesis tests will tend to fail to reject the null hypothesis that the components of β are not statistically different from zero.

Multicollinearity

Unless you have perfect collinearity, in which case Stata will drop a regressor, detecting multicollinearity is tricky, given that having imprecise estimates is not uncommon with observational data.

One thing I see often in the manuscripts I review or am in charge of as an editor is a correlation matrix, which shows the correlation between the variables in a regression. But this is only useful insofar as you have multicollinearity issues between two variables; if the multicollinearity issue stems from an approximate linear relationship between three or more variables, the correlation matrix is near useless.

Multicollinearity

What to do when you suspect you are dealing with a multicollinearity problem? Kennedy offers a few ideas; I am listing those that strike me as the most practical:

- Do nothing. This is especially useful if your coefficient estimates turn out to be statistically significant–if you do get significance even with imprecisely estimated coefficients, you're in relatively good shape.
- Get more data. See the discussion above for why that might be a good idea. This can be a costly option, however, and by "costly," I mean "impossible."

Multicollinearity

- Drop one of the collinear variables. That would have been my default prior to writing this post, but this only is a workable solution if that variable adds nothing to the regression to begin with, i.e., if its estimated coefficient is zero. But then, how can you tell whether that is the case if that coefficient is imprecisely estimated? Moreover, doing this introduces bias, so you need to think carefully about whether you're willing to deal with bias in order to mitigate imprecision.
- Use principal components or factor analysis. This boils down to creating an index with the multicollinear variables or estimating some linear combination of those same variables which is then used as a single regressor. The latter is especially useful when you have several variables that aim to measure the same thing, and you want to include them all.

Multicollinearity

I must confess that I hardly ever worry about collinearity in my own work. That's because if the problem gets too extreme, Stata will drop one of the collinear variables, and if the problem is not extreme, it is hard to determine whether a statistically insignificant coefficient estimate is imprecisely estimate because of multicollinearity or because of there being no statistically significant relationship.

Correlation Isn't Necessarily Transitive

When y is correlated with D, and D is correlated with z, y isn't necessarily correlated with z. That is, correlation is not always transitive.

Correlation Isn't Necessarily Transitive

From my choice of labels for those variables, you have probably guessed why this is matters for applied econometrics: It is perfectly possible that in a regression of y on D where you are interested in the causal relationship flowing from D to y, you have an otherwise valid instrument z (i.e., a variable that is plausibly exogenous to outcome y and which is also relevant in explaining D).

Obviously, z being relevant means that it is correlated with D. Assuming that D is also correlated with y, the fact that correlation isn't necessarily transitive means that the IV is not necessarily correlated with the outcome.

Correlation Isn't Necessarily Transitive

What this means in practice is that when doing IV, you should always show a reduced-form regression of y on z, whose purpose is to reassure your readers that your IV actually affects your outcome variable.

This is a point that is made by Angrist and Pischke in *Mostly Harmless Econometrics*.

Correlation Isn't Necessarily Transitive

You might be tempted to think that finding that the coefficient on z is not statistically significant in a reduced-form regression of y on z is a good thing, because it proves that the IV is uncorrelated with the outcome of interest.

Correlation Isn't Necessarily Transitive

Drawing such a conclusion would be misguided, however, first because non-rejection of the null of no statistical significance in this case is not definitive proof that the two are uncorrelated—null results can be about evidence of absence, but they can also be about absence of evidence—and second because what we want here is not for z to be uncorrelated with y.

Though we often people say that a good IV is uncorrelated with y, what we actually want is for the IV to be correlated with y, but only through D.

Correlation Isn't Necessarily Transitive

So what should you do in cases where your reduced-from regression of y on z shows that the two do not appear to be correlated?

Much like there are three ways to get to play at Carnegie Hall—practice, practice, and practice—you should probably strive to explain, explain, and explain some more why the IV is still valid in such cases.

One way out of this might be to explain that failing to reject the null in this case is simply absence of evidence, but this is probably only convincing in cases where your sample is small.

Type III Errors

Kennedy (2008) writes:

> *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.*

Type III Errors

How common are type III errors? And what does a type III error look like in practice?

There is no hard evidence on the extent of type III errors beyond the anecdotal, but how often do we run into applied researchers who are bent on applying a specific technique to a given problem?

Type III Errors

Relatedly, I remember sitting in a seminar wherein a colleague asked "How did you come to work on this topic?" to the presenter, who had just shown us findings behind which the causal story appeared tenuous to my colleague.

When the presenter said "Well, I noticed that I had this source of exogenous variation in my data, so I decided to look for an outcome to apply it to...," he lost about half of the audience.

I guess the lesson is to make sure you have a good story for why your empirical setup is interesting, and why it matters for policy, business, or other.

Type III Errors

How do you protect yourself from making type III errors? As with many other things in applied econometrics, the answer is not contained in a theorem or lemma.

Rather, the answer is to talk to colleagues about what you are doing, and to see whether what you are doing passes the laugh test.

Type III Errors

Though you should often expect a certain amount of skepticism when you first explain what you are doing (after all, most of the obvious research questions have already been answered), you should be able to dissipate said skepticism pretty quickly with facts.

So in a typical regression of y on D, that means explaining how much the effect of D on y costs to society, how many people it affects, how much it shortens life expectancy, etc.

Statistical vs. Economic Significance

Every so often, you run into a paper in which the authors have a good story, a good identification strategy, and robust, statistically significant findings, but in which there is little to no discussion of the findings' economic significance.

Statistical vs. Economic Significance

What is economic significance? For the purpose of this discussion, let's define statistical significance in its usual sense—Is the null hypothesis that the coefficient of interest is statistically different from zero rejected at the 90, 95, or 99 percent levels of confidence?

Similarly, let's define economic significance as how much something matters in the real world—Put simply: Is the treatment effect big or small in real-world terms?

Statistical vs. Economic Significance

So I guess I don't have much more of a point than "Make sure you discuss the economic significance of your findings on top of their statistical significance."

Basic econometrics courses are of no help here, as they tend to be too generic to get into economic significance beyond broad recommendations.

Statistical vs. Economic Significance

Applied courses tend to be a lot better; suppose you have a good identification strategy to estimate the effect of a policy in which some consumers receive a lump-sum transfer on those consumer's marginal propensity to consume (MPC), which you find is significant at the 99 percent level of confidence.

What if that estimate tells you that the effect of that lump-sum transfer is to change MPC by 0.02 percent? Chances are the policy isn't really worth it. But if your estimate says that change is equal to 20 percent, the story changes.

Statistical vs. Economic Significance

Keeping statistical and economic significance in mind, there are four possible cases:

- A finding is statistically significant and economically significant. This is the ideal case, and the one that makes your job easiest when it comes to convincing readers that you have a publishable finding.

Statistical vs. Economic Significance

- A finding is statistically significant but economically insignificant. For me, this is second-best. You may be tempted to gloss over economic significance in such cases because you worry about the consequences of being honest about reporting an economically insignificant finding, but I think the consequences of trying to hide this are much worse than the consequences of being up-front about it. Besides, there is something to learn from such cases: Some things just don't work, or they don't work as well as previously thought.

Statistical vs. Economic Significance

- A finding is statistically insignificant and economically significant. This is very likely to happen when you have too small a sample size and you don't have much statistical power. For such cases, I recommend taking a look at an old, underappreciated article by Don Andrews titled "Power in Econometric Applications" (Andrews, 1989).

Statistical vs. Economic Significance

- A finding is statistically insignificant and economically insignificant. This is the most difficult case to work with. In order to publish such null findings, you have to work very, very hard to show that you are demonstrating evidence of absence of an effect rather than dealing with absence of evidence (i.e., low statistical power). I have managed to publish one such finding once, but it took (i) my contradicting widespread conventional wisdom, (ii) several robustness checks, and (iii) a sympathetic editor for this to happen.