PLSC 503: "Multivariate Analysis for Political Research"

Variable Selection and Specification Bias
March 21, 2017

Introduction

Model specification encompasses a number of things. While it tends to be thought of in terms of including irrelevant or excluding relevant independent variables, it also deals with a lot of other issues, such as functional forms, interaction effects, and others – all of which we'll talk about after today.

Model specification is often thought of in terms of what variables belong on the right-hand side of a regression equation. In some respects, it's more valuable to think of it in terms of coefficient restrictions; that is, *what restrictions on the impact of your* $\mathbf{X}$*s on* Y *do the model you estimate imply?* The important meta-lessons to take away from that way of thinking about specification in this way is that model specification is hard, that theory is (concomitantly) important, and that in the absence of strong theories it is often valuable to be a bit more flexible in one's model specification. But, we'll talk more on that later; before we get to model specification, we should take a brief aside to discuss the issue of random covariates.

Random Covariates

Stock treatments of OLS regression begin with the assumption that the covariates $\mathbf{X}$ are fixed in repeated sampling – that is, that, were we to repeat the data-gathering process again, the values of $\mathbf{X}$ would remain the same (and only the values of $\mathbf{Y}$ would change). This makes sense in a few instances; for example, in experiments, we might randomize respondents across values of some $\mathbf{X}$ variable and then take a second, third, etc. measure on their responses.

However, the assumption is harder to justify when (as is typically the case in political science) we have nonexperimental, observational data. This is particularly the case when one person's response variable (be it international trade, ethnic fractionalization, party identification, or whatever) is another person's covariate.

So, we should ask the question of what happens if one or more elements of $\mathbf{X}$ are random. As it happens, the answer is, not much; for our OLS estimator to have nice properties with random regressors, we need only to make a couple additional, rather mild assumptions about the model. Fox has a nice discussion of all this; but, specifically, with a stochastic $\mathbf{X}$, the regular best linear unbiased properties of OLS will hold when:

1. $\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$ – that is, the covariates and the errors are independent, and
2. The distribution of $\mathbf{X}$ does not depend on either $\boldsymbol{\beta}$ or σ^2.

If these conditions are met, then stochastic regressors pose no problem. As a practical matter, this is not too difficult; the first of these is generally required of our models anyway, while the second requires only that the random component of the $\mathbf{X}$s not vary systematically with either their influences on $\mathbf{Y}$ or the degree of variability in the errors $\mathbf{u}$.

Model Specification: Basics

Consider a very simple two-variable OLS model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{1}$$

where all the (other) usual OLS assumptions are met. Implicit in this specification is that $\beta_1 \neq 0$ and $\beta_2 \neq 0$; otherwise there would be no need to have both X_1 and X_2 in the model.

Consider first what happens if we exclude X_2 from the regression model we actually estimate:

$$Y_i = \gamma_0 + \gamma_1 X_{1i} + e_i \tag{2}$$

Note that this is precisely equivalent to imposing the restriction $\beta_2 = 0$ on (1) (along with all the other higher-order restrictions implied; more on that in the next couple weeks). Note as well that, in (2),

$$e_i = \beta_2 X_{2i} + u_i$$

where $\text{E}(u_i) = 0$, $\text{Cov}(X_{2i}, u_i) = 0$, and all the other usual assumptions about X_2 and u hold. At the outset, it is clear that, since $\beta_2 \neq 0$,

$$\begin{aligned} \text{E}(e) &= \text{E}(\beta_2 X_2 + u) \\ &= X_2 \text{E}(\beta_2) + \text{E}(u) \\ &\neq 0 \end{aligned}$$

In other words, the expectation of the "error" term in (2) is not zero; that, in turn, means that the basic condition for unbiasedness in our estimates of γ_0 and γ_1 is not met. **As a result, $\hat{\gamma}_0$ and $\hat{\gamma}_1$ will be biased.** The nature of that bias depends on the relationship between X_1 and X_2. Formally, we can write the expectation of γ_1 as:

$$\begin{aligned} \text{E}(\gamma_1) &= \beta_1 + \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2}\beta_2 \\ &= \beta_1 + b_{X_2 X_1}\beta_2 \end{aligned} \tag{3}$$

where $b_{X_2X_1}$ is the "slope" coefficient one obtains from regressing X_2 on X_1. This means that the degree of bias that results from omitting X_2 depends on the extent of the relationship between X_1 and X_2:

1. If X_1 and X_2 are *uncorrelated* – that is, $\text{Cov}(X_1, X_2) = 0$ – then

   - $\hat{\gamma}_1$ will be an *unbiased* estimate of β_1.
   - While this seems unlikely to be a very useful result, it can be in certain circumstances, particularly when randomization is used to insure that a variable of interest is uncorrelated with other ("control") variables. Also,
   - It is possible to show that $\hat{\gamma}_0$ will be a *biased* estimate of the true intercept β_0.

2. If X_1 and X_2 are *correlated* – that is, $\text{Cov}(X_1, X_2) \neq 0$ – then

   - $\hat{\gamma}_1$ will be a *biased* estimate of β_1, and $\hat{\gamma}_0$ will be a *biased* estimate of β_0 as well.
   - Mathematically, this is because $b_{X_2X_1} \neq 0$.
   - Moreover, for the simple model in (1), (3) implies that
     - if X_1 and X_2 are *positively* correlated, $\hat{\gamma}_1$ will *overestimate* β_1, while
     - if X_1 and X_2 are *negatively* correlated, $\hat{\gamma}_1$ will *underestimate* β_1.
   - Intuitively, this is because some of the variability of Y is common to both of the Xs; with X_2 excluded, all of that joint (co)variance is "attributed" to X_1.

In other words, if – as is likely to be the case – you omit one (or more) variable(s) that are correlated with one or more "included" variables, the result is to bias your estimates of the effects of those included variables on Y.

Omitted Variables and Inference

Not surprisingly, in addition to introducing bias, model misspecification also louses up our ability to make correct inferences about the parameters of interest as well. This is because we're basing our estimate of the variance of the errors $\hat{\sigma}^2$ on the estimates of the model in (2); call that $\hat{\sigma}_e^2$. Recall that, for a bivariate model in general,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N - k}$$

and

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

Also, recall – from the lecture on collinearity – that we can write the variance of $\hat{\beta}_1$ in the model with two right-hand-side covariates as:

$$\widehat{\mathrm{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2(1 - R^2_{X_1X_2})}$$

where $R^2_{X_1X_2}$ is the R^2 from regressing X_2 on X_1 (that is, the squared Pearson correlation between X_1 and X_2). This means that the variance of $\hat{\beta}_1$ will always be at least as large in the model that includes X_2 as in the model that omits it, so long as $\mathrm{Cov}(X_1, X_2) \neq 0$.

But, of course, that's only part of the problem. Because, in general, $\hat{e}_i \neq \hat{u}_i$, the estimate of σ^2 based on the misspecified model ($\hat{\sigma}^2_e$) will be biased and inconsistent. This in turn means that the standard error estimates based on $\hat{\sigma}^2_e$ will be biased. More generally, it is possible to show that:

$$\mathrm{E}(\sigma^2_e) = \sigma^2_u + f\left(\beta_2, X_1\right) \tag{4}$$

(see, e.g., Greene 1997, 402-4). That is to say, the estimate of the disturbance variance σ^2_u is biased. As a general rule, its impossible to say which way it will be biased, since that direction depends on both the extent to which the two Xs covary and the extent to which X_2 (the omitted variable) and Y covary (that is, how much the estimated error variance σ^2 is reduced by adding X_2 to the equation). The result is that inferences based on the standard estimate of the variance–covariance matrix of the βs will generally be incorrect.

Multivariate Matters

In the multivariate case, consider a basic model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{5}$$

which meets all our usual OLS criteria. Now suppose we estimate an alternative model:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{e} \tag{6}$$

where $\mathbf{Z} \subset \mathbf{X}$. This gives an estimator

$$\boldsymbol{\Gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$

Substituting, we see that:

$$\begin{aligned} \boldsymbol{\Gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \end{aligned} \tag{7}$$

Recalling the assumption that $\mathrm{E}(\mathbf{u}) = 0$, we see that

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Gamma}) &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}\boldsymbol{\beta} \end{aligned} \tag{8}$$

where $\mathbf{P} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. This is, in effect, a vector of the estimated coefficients from regressing each of the variables $\mathbf{X}$ on the variables in $\mathbf{Z}$ and arranging their coefficients in a vector. $\mathbf{P}$ is thus a measure of the extent of bias in the estimated coefficients in the misspecified regression; the direction and extent of the bias will, again, depend (in a complicated way) on the extent to which the variables in $\mathbf{X}$ omitted from $\mathbf{Z}$ are correlated with those in $\mathbf{Z}$.

Overspecification

So far, we've focused on the omission of variables that are "supposed to be" among the covariates in the model. But what if instead one includes one (or more) irrelevant variable(s) in your estimation? Formally, we can think about this as in (5) and (6), above, but where $\mathbf{X} \subset \mathbf{Z}$, rather than the other way around. That is, we have a "correct" model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

which we estimate using

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{e}$$

where $\dim(\mathbf{Z}) \geq \dim(\mathbf{X})$. Define $\mathbf{W}$ as the "irrelevant" variables in $\mathbf{Z}$, so that we are estimating a model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\Theta} + \mathbf{u} \tag{9}$$

where, by assumption, $\boldsymbol{\Theta} = 0$.

Estimating (9) can be thought of as incorrectly failing to impose the constraint that $\boldsymbol{\Theta} = 0$. Put a little differently – and in contrast to the omitted-variable case discussed above – we can think of estimating (9) not as *using incorrect information* about the influences on $\mathbf{Y}$, but instead as *failing to use (some) correct information* about that relationship.[1]

As a result of this distinction, the ill effects of including "irrelevant" variables are mild. In particular, incorrectly estimating (9) in place of (5) will yield

- Estimates $\hat{\boldsymbol{\beta}}$ that are unbiased and consistent estimates of $\boldsymbol{\beta}$, and
- An estimate of the error variance σ^2 that is also unbiased. However,
- The estimates produced will generally be *inefficient* (that is, they will have larger than optimal variances).

This makes intuitive sense:

[1] In this sense, it is akin to not including a data point that "fits" the model, while omitted variable bias comes from including data that do not.

- Inclusion of one or more irrelevant variables among the covariates doesn't interfere with the impact of the relevant ones on $\mathbf{Y}$; as a result, there is no bias introduced in our estimates of $\boldsymbol{\beta}$. Similarly,
- Their inclusion has no effect on the estimated residuals, and so also doesn't affect the estimate of σ^2.
- What it *does* do is eat up one degree of freedom for each "extraneous" covariate, thus unnecessarily reducing the efficiency of the model.

Practical Implications, and Pre-Test Bias

In general, then,

- Omitting important variables leads to estimates which are biased, incorrect standard errors, and generally worthless results.
- If those omitted variable(s) happen to be uncorrelated with the included ones, then the bias is limited to the intercept, but the bias and inconsistency in the standard errors remains, rendering inference problematic. Conversely,
- Including irrelevant variables does not yield bias or inconsistency in any estimated parameters, but leads to inefficiency.

All of this would seem to suggest that including extraneous variables is better than omitting relevant ones; and, that is true, to a point. However, overspecification can pose problems, particularly when inefficiency is high (due to lots of included covariates), and even more so when those variables are highly collinear with each other. The latter, as we know, leads to larger (albeit not incorrect) standard error estimates, which in turn makes finding "statistically significant" results more difficult.

The result, as noted by (e.g.) Greene, is that model specification and collinearity often present applied researchers with a choice: either

- Include many, many variables to avoid bias, and deal with high levels of collinearity (and correspondingly "nonsignificant" results), or
- Omit variables to "beef up" one's statistical significance, but at the risk of omitted variable bias.

In these circumstances, analysts often choose a "third way," first estimating a series of models with combinations of variables and including or excluding covariates on the basis of t-tests and significance (α) levels. Greene and others refer to this as a **pre-test estimator**. The Judge et al. (1985, pp. 72-82) volume has a nice discussion of these estimators; here, suffice it to say that, as a model-building strategy, pre-test estimators leave a lot to be desired. In particular,

- Pre-test estimators are *biased*. In particular, **the properties of pre-test estimators depend on the significance levels α, which, in applied work, is usually chosen in an ad hoc manner**.
- In mean-squared error terms, the pre-test estimator is inferior to "unrestricted" OLS unless the model that results from the pre-test estimator is precisely properly specified.
- As Greene notes, "The pre-test estimator is the least precise of the three (restricted, unrestricted, and pre-test estimators) when the researcher is most likely to use it" (2003, 150).

The point of this mini-diatribe is thus to point out that, contrary to what many people think, **model specification ought not be guided by t-tests and other inferentially-based approaches**.

Omitted Variable Bias: A (Simulated) Example

As a simple illustration of over- and underspecification, I generated some "fake" data ($N = 100$) on four variables. The model is

$$Y_i = 0 + 1.0X_{1i} - 2.0X_{2i} + u_i \tag{10}$$

where $X_1 \sim N(0, 1)$, $u_i \sim N(0, 2)$, and X_1 and X_2 are correlated at -0.5. I also generated a fourth variable, Z, that is $\sim N(0, 10)$ and unrelated to any of the other three variables:

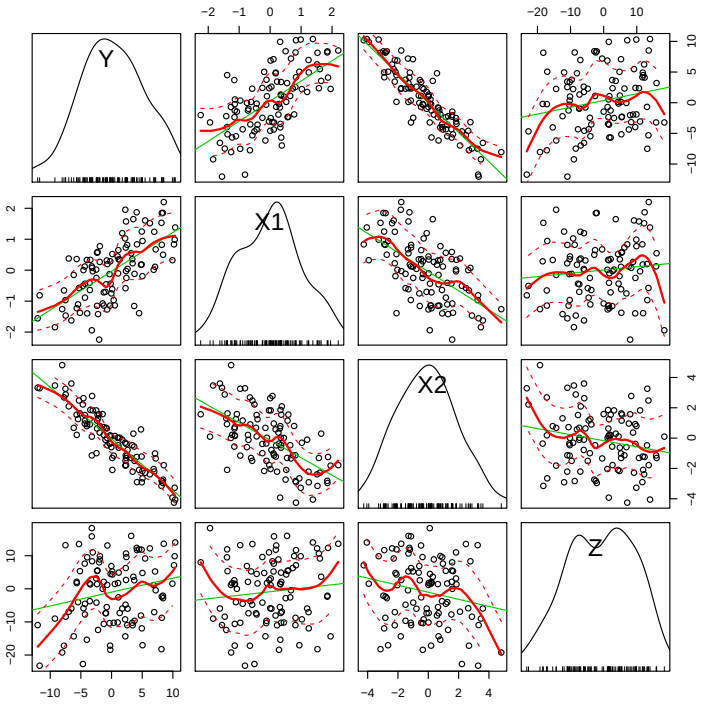
```
> N <- 100
> X1<-rnorm(N)              # <- X1
> X2<-(-X1)+1.5*(rnorm(N)) # <- correlated w/X1
> Y<-X1-(2*X2)+(2*(rnorm(N))) # <- Y
> Z<-10*rnorm(N)            # <- irrelevant
> data <- data.frame(Y=Y,X1=X1,X2=X2,Z=Z)
```

A scatterplot matrix of the variables so generated is in Figure 1. The "correctly" specified model is then:

```
> correct<-lm(Y~X1+X2)
> summary(correct)

Residuals:
   Min     1Q Median     3Q    Max
-5.721 -1.209  0.093  1.198  5.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03311    0.21249  -0.156  0.87651
```

```
X1           0.81690    0.26718   3.057  0.00288 **
X2          -2.13652    0.13844 -15.433  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.116 on 97 degrees of freedom
Multiple R-squared:  0.8295,Adjusted R-squared:  0.826
F-statistic:   236 on 2 and 97 DF,  p-value: < 2.2e-16
```

Figure 1: Scatterplot Matrix of X_1, X_2, and Y (Simulated Data, $N = 100$)



If we estimate the "correct" model but also include Z (the "irrelevant" variable), we get the following:

```
> overspec<-lm(Y~X1+X2+Z)
> summary(overspec)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9809 -1.0442 -0.0265  1.2609  6.0201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01570    0.21420  -0.073  0.94173
X1           0.82148    0.26785   3.067  0.00281 **
X2          -2.11735    0.14105 -15.011  < 2e-16 ***
Z            0.01662    0.02202   0.755  0.45220
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 2.12 on 96 degrees of freedom
Multiple R-squared:  0.8306,Adjusted R-squared:  0.8253
F-statistic: 156.8 on 3 and 96 DF,  p-value: < 2.2e-16
```

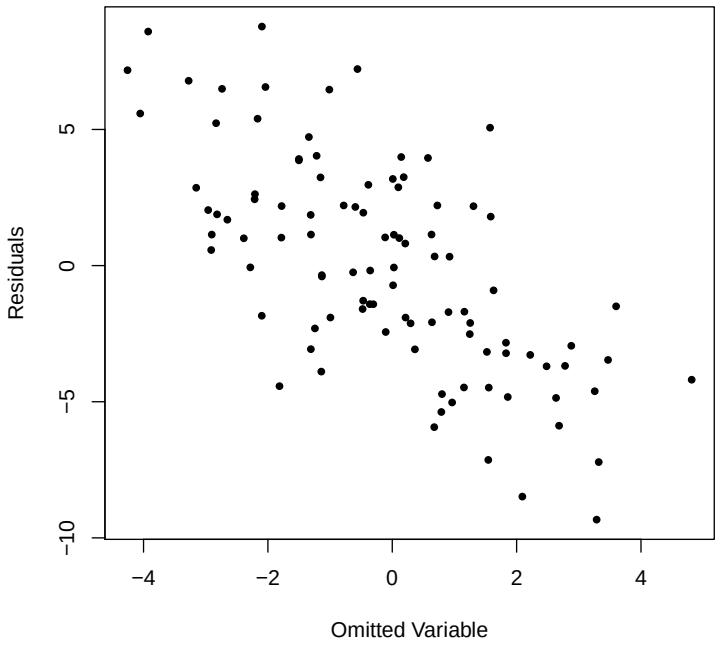
Here, the effects for X_1 and X_2 are more or less perfectly estimated, and at very little loss of efficiency. On the other hand, if we estimate a model that is underspecified (by omitting X_2 from the estimation), we get:

```
> incorrect<-lm(Y~X1)
> summary(incorrect)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3297 -2.9762 -0.0672  2.4828  8.7787

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2704     0.3913   0.691    0.491
X1            3.2783     0.3964   8.270 6.71e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.913 on 98 degrees of freedom
Multiple R-squared:  0.411,Adjusted R-squared:  0.405
F-statistic: 68.39 on 1 and 98 DF,  p-value: 6.714e-13
```

Note that:

- Our estimate for the effect of X_1 ($\hat{\gamma}_1$, in the terminology discussed above) is badly overestimated. This is because both $\text{Corr}(X_1, X_2) < 0$ and $\text{Corr}(X_2, Y) < 0$. As a result, the model incorrectly "assigns" some of the variation in Y "due to" X_2 to X_1.
- The estimate for the intercept term is also biased, albeit not as much.
- We can also say with confidence that the estimated standard error for $\hat{\beta}_1$ is wrong, though precisely *how* it is wrong we need not go into.
- A scatterplot of the residuals from the underspecified model against X_2 (presented in Figure 2) clearly illustrates that the residuals $\hat{u}_i$ are negatively correlated with X_2 – a strong sign that X_2 should be included in the model:

Figure 2: Scatterplot Matrix of Residuals from Misspecified Model, Against X_2 (The Omitted Variable)



Detecting and Dealing With Misspecification

In the matter of model specification, one fact rises above all others:

Nothing Beats a Good Theory. Period.

Note a few things about this otherwise simple-sounding statement:

- "Good" is often (but not always) a quite different adjective from the following:
  - Famous
  - Complicated
  - Formal
- In fact, this is precisely the *wrong* class in which to learn what constitutes a good theory; that's why most of what we teach you here has little or nothing to do with statistics per se, and why this class is (arguably) the least important one in your graduate education.

- The pessimistic view is that all theory is imperfect, and therefore that you'll *always* misspecify your model(s), and so your inferences will *always* be wrong. If you believe that, then your alternative is to do anthropology, area studies, or something akin to them.
- The more optimistic perspective is that theories (and therefore models) are all imperfect, but that we can still learn from them (or, as my own Ph.D. thesis advisor once said, "If it's worth doing, it's worth doing badly").

Beyond that, there are a few other rules of thumb / pointers to keep in mind:

- *Unnecessary variables will often have small, insignificant coefficients.* Note that the "small" is as important as the "insignificant," since if you have a large N even tiny coefficients can be statistically different from zero.
- *Look at your residuals.* They can tell you a lot about model specification, provided that you have some sense of the question you're investigating (and, if you don't, you have much bigger problems).
- *Resist the urge to overspecify.* Including "irrelevant" variables is one thing; blindly including variables that are intervening, or causal in some other way, is just as bad as (arguably worse than) leaving them out.

Model Specification Tests

Statisticians and (mostly) econometricians have devised some tests for omitted variable bias. All are based on the idea that, if you omit a relevant variable, that variable will be correlated with the residuals from the misspecified model. So, for example:

- One can conduct a Durbin-Watson d test on cross-sectional data after sorting the data on some omitted variable; if it is significant, it suggests that there is something systematic in the pattern of the residuals across observations.
- Ramsey's RESET test is another widely-seen test. The procedure is to
    1. Estimate the regression and obtain the residuals,
    2. Generated predicted values of Y (that is, $\hat{Y}$),
    3. Reestimate the regression, including all the original covariates in $\mathbf{X}$ as well as some form of the $\hat{Y}$s,
    4. Use an F-test to compare the two models: if the F-test says that inclusion of the $\hat{Y}$s improves the fit of the model, then the model is likely misspecified.

Thankfully, these tests have never really taken off in political science the way they have in economics; in fact, it is likely that you'd be roundly chastised by panel discussants and journal reviewers if you used such tests to determine your model's specification.

Proxy Variables

Another (more common) situation is where you know that a relevant variable ought to be included, but one isn't available. For example, suppose we were interested in explaining individual-level campaign contributions, using survey data. We'd like to include a variable for each respondent's income/wealth, since that is very likely related not only to how much they contribute, but also to whom. However, data on income in mail and/or telephone surveys is notoriously poor; it is either unavailable (missing), or mis-measured, or coded in (say) ordinal categories.

In cases like this, one alternative is to include in the model a variable that tracks closely with the thing you want to measure; this is known as a "proxy" variable. With respect to income, for example, one might include a variable for *years of formal education* instead of income; this has the advantages of

- Being closely related to income,
- more commonly measured, and
- more accurately measured.

More generally, proxies are imperfect measures of the thing they're substituting for. Because of that fact, proxy variables introduce the possibility of measurement error. However, as we've seen before, that's often a more tractable problem with which to deal than is specification bias.

Conclusion

Finally, note that this entire discussion has hinged on which covariates to include in a regression model, and which to exclude. Equally important – and, thus far, undiscussed – are such weighty matters as:

- *How* to include the variables – that is, their *functional form*.
- *Mediated effects* – are the influences of each element of $\mathbf{X}$ constant for all other values of all other elements of $\mathbf{X}$? If not, then the $\mathbf{X}$ variables *interact*.
- *Causality*. Do some variables in $\mathbf{X}$ "cause" other elements of $\mathbf{X}$? If so, what should be done?

All of this heady stuff will form the basis for discussions in future classes.