Wooldridge, Introductory Econometrics, 4th ed.

Chapter 9: More on specification and data problems

Functional form misspecification

We may have a model that is correctly specified, in terms of including the appropriate explanatory variables, yet commit functional form misspecification–in which the model does not properly account for the relationship between dependent and observed explanatory variables. We have considered this sort of problem when discussing polynomial models; omitting a squared term, for instance, and constraining $\partial y/\partial x$ to be constant (rather than linear in x) would be

a functional form misspecification. We may also encounter difficulties of this sort with respect to interactions among the regressors. If omitted, the effects of those regressors will be estimated as constant, rather than varying as they would in the case of interacted variables. In the context of models with more than one categorical variable, assuming that their effects can be treated as independent (thus omitting interaction terms) would yield the same difficulty.

We may, of course, use the tools already developed to deal with these problems, in the sense that if we first estimate a general model that allows for powers, interaction terms, etc. and then "test down" with joint F tests, we can be confident that the more specific model we develop will not have imposed inappropriate restrictions along the way. But how can

we consider the possibility that there are missing elements even in the context of our general model?

One quite useful approach to a general test for functional form misspecification is Ramsey's **RESET** (regression specification error test). The idea behind RESET is quite simple; if we have properly specified the model, no nonlinear functions of the independent variables should be significant when added to our estimated equation. Since the fitted, or predicted values ($\widehat{y}$) of the estimated model are linear in the independent variables, we may consider powers of the predicted values as additional regressors. Clearly the $\widehat{y}$ values themselves cannot be added to the regression, since they are by construction linear combinations of the x variables. But their squares, cubes,... are not. The RESET formulation reestimates the original equation, augmented by powers of $\widehat{y}$ (usually squares, cubes, and fourth powers are sufficient) and conducts an F-test for the joint null

hypothesis that those variables have no significant explanatory power. This test is easy to implement, but many computer programs have it already programmed; for instance, in Stata one may just specify `estat ovtest` (omitted variable test) after any regression, and the Ramsey RESET will be produced. However, as Wooldridge cautions, RESET should not be considered a general test for omission of relevant variables; it is a test for misspecification of the relationship between y and the x values in the model, and nothing more.

Tests against nonnested alternatives

The standard joint testing framework is not helpful in the context of "competing models," or nonnested alternatives. These alternatives can also arise in the context of functional form: for instance,

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad (1)\\
y &= \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u
\end{aligned}
$$

are nonnested models. The mechanical alternative, in which we construct an artificial model that contains each model as a special case, is often not very attractive (and sometime will not even be feasible). An alternative approach is that of Davidson and MacKinnon. Using the same logic applied in developing Ramsey's RESET, we can estimate each of the models in (1), generate their predicted values, and include them in the other equation. Under the null hypothesis that the first form of the model is correctly specified, a linear combination of the logs of the x variables should have no power to improve it, and that coefficient should be insignificant. Likewise, one can reestimate the second model, including the predicted values from the first model. This

testing strategy–often termed the Davidson-MacKinnon "**J test**"–may indicate that one of the models is robust against the other.

There are no guarantees, though, in that applying the J test to these two equations may generate zero, one, or two rejections. If neither hypothesis is rejected, then the data are not helpful in ranking the models. If both are rejected, we are given an indication that neither model is adequate, and that a continued specification search should be conducted. If one rejection is received, then the J test is definitive in indicating that one of the models dominates (or subsumes) the other, and not vice versa. However, this does not imply that the preferred model is well specified; again, this test is against a very specific alternative, and does not deliver a "clean bill of health" for the preferred model should one emerge.

Proxy variables

So far, we have discussed issues of misspecification resulting from improper handling of the x variables. In many economic models, we are forced to employ "proxy variables": approximate measures of an unobservable phenomenon. For instance, admissions officers use SAT scores and high school GPAs as proxies for applicants' ability and intelligence. No one argues that standardized tests or grade point averages are actually measuring aptitude, or intelligence; but there are reasons to believe that the observable variable is well correlated with the unobservable, or **latent**, variable. To what extent will a model estimated using such proxies for the variables in the underlying relationship be successful, in terms of delivering consistent estimates of its parameters? First, of course, it must be established that there is a correlation between the observable variable and the latent variable. If we consider the

latent variable as having a linear relation to a measurable proxy variable, the error in that relation must not be correlated with other regressors. When we estimate the relationship including the proxy variable, it should be apparent that the measurement error from the latent variable equation ends up in the error term, as an additional source of uncertainty. This is an incentive to avoid proxy variables where one can, since they will inexorably inflate the error variance in the estimated regression. But usually they are employed out of necessity, in models for which we have no ability to measure the latent variable. If there are several potential proxy measures, they might each be tested, to attempt to ascertain whether bias is being introduced to the relationship.

In some cross-sectional relationships, we have the opportunity to use a lagged value of the dependent variable as a proxy variable. For instance, if we are trying to explain cities' crime

rates, we might consider that there are likely to be similarities—irregardless of the effectiveness of anti-crime strategies—between current crime rates and last year's values. Thus, a prior value of the dependent variable, understandably independent of this year's value, may be a useful proxy for a number of factors that cannot otherwise be quantified. This approach might often be used to deal with factors such as "business climate," in which some states or municipalities are viewed as more welcoming to business; there may be many aspects to this perception, some of them more readily quantifiable (such as tax rates), some of them not so (such as local officials' willingness to negotiate infrastructure improvements, or assist in funding for a new facility). But in the absence of radical changes in localities' stance in this regard, the prior year's (or decade's) business investment in the locality may be a good proxy for those factors, perceived much more clearly by the business decisionmakers than by the econometrician.

Measurement error

We often must deal with the issue of measurement error: that the variable that theory tells us belongs in the relationship cannot be precisely measured in the available data. For instance, the exact marginal tax rate that an individual faces will depend on many factors, only some of which we might be able to observe: even if we knew the individual's income, number of dependents, and homeowner status, we could only approximate the effect of a change in tax law on his or her tax liability. We are faced, therefore, with using an approximate measure, including some error of measurement, whenever we might attempt to formulate and implement such a model. This is conceptually similar to the proxy variable problem we have already discussed, but in this case it is not a latent variable problem. There is an observable magnitude, but we do not necessarily observe it. For instance, reported income is

an imperfect measure of actual income, while IQ score is only a proxy for ability. Why is measurement error of concern? Because the behavior we're trying to model–be it of individuals, firms, or nations–presumably is driven by the actual measures, not our mismeasured approximations of those factors. To the extent that we fail to capture the actual measure, we may misinterpret the behavioral response.

If measurement error is observed in the dependent variable–for instance, if the true relationship explains y^*, but we only observe $y = y^* + \epsilon$, where ϵ is a meanzero error process, then ϵ becomes a component of the regression error term: yet another reason why the relationship does not fit perfectly. We assume that ϵ is not systematic, in particular, that it is not correlated with the independent variables X. As long as that is the case, then this form of measurement error does no real harm; it

merely weakens the model, without introducing bias in either point or interval estimates. If the magnitude of the measurement error in y is correlated with one or more of the x variables, then we will have a problem of bias.

Measurement error in an explanatory variable, on the other hand, is a far more serious problem. Say that the true model is

$$y = \beta_0 + \beta_1 x_1^* + u \tag{2}$$

but that x_1^* is not observed; instead, we observe $x_1 = x_1^* + \epsilon_1$. We can assume that $E(\epsilon_1) = 0$ with generality. But what should be assumed about the relationship between ϵ_1 and x_1^*? First, let us assume that ϵ_1 is uncorrelated with the observed measure x_1 (that is, larger values of x_1 do not give rise to systematically larger (or smaller) errors of measurement). This can be written as $Cov(\ \epsilon_1,\ x_1) = 0$. But if this is the case, it must be

true that $Cov(\ \epsilon_1,\ x_1^*) \neq 0$: that is, the error of measurement must be correlated with the actual explanatory variable x_1^*, so that we can write the estimated equation (in which x_1^* is replaced with the observable x_1) as

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 \epsilon_1) \qquad (3)$$

Since both u and ϵ_1 have zero mean and are uncorrelated (by assumption) with x_1, the presence of measurement error merely inflates the error term: that is, $Var\,(u - \beta_1 \epsilon_1) = \sigma_u^2 + \beta_1^2 \sigma_{\epsilon_1}^2$, given that we have assumed that u and ϵ_1 are uncorrelated with each other. Thus, measurement error in x_1^* does not negatively affect the regression of y on x_1; it merely inflates the error variance, like measurement error in the dependent variable.

However, this is not the case that we usually consider under the heading of **errors-in-variables**. It is perhaps more reasonable to

assume that the measurement error is uncorrelated with the true explanatory variable: $Cov($ $\epsilon_1,\ x_1^*) = 0$. If this is so, then $Cov(\ \epsilon_1,\ x_1) = Cov(\epsilon_1, \left(x_1^* + \epsilon_1\right)) \neq 0$ by construction, and the regression (3) will have a correlation between its explanatory variable x_1 and the composite error term. The covariance of $(x_1, u - \beta_1\epsilon_1) = -\beta_1 Cov(\epsilon_1,\ x_1) = -\beta_1\sigma_{\epsilon_1}^2 \neq 0$, causing the OLS regression of y on x_1 to be biased and inconsistent. In this simple case of a single explanatory variable measured with error, we can determine the nature of the bias:

$$\begin{aligned} \texttt{plim}(b_1) &= \beta_1 + \frac{Cov\left(x_1, u - \beta_1\epsilon_1\right)}{Var(x_1)} \\ &= \beta_1 \left(\frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_{\epsilon_1}^2}\right) \end{aligned} \tag{4}$$

demonstrating that the OLS point estimate will be **attenuated**–biased toward zero–since

the bracketed expression must be a fraction. Clearly, in the absence of measurement error, $\sigma^2_{\epsilon_1} \rightarrow 0$, and the OLS coefficient becomes unbiased and consistent. As $\sigma^2_{\epsilon_1}$ increases relative to the variance in the (correctly measured) explanatory variable, the OLS coefficient becomes more and more unreliable, shrinking toward zero.

What can we conclude in a multiple regression equation, in which perhaps one of the explanatory variables is subject to measurement error? If the measurement error is uncorrelated to the true (correctly measured) explanatory variable, then the result we have here applies: the OLS coefficients will be biased and inconsistent for all of the explanatory variables (not merely the variable measured with error), but we can no longer predict the direction of bias in general terms. Realistically, more than one explanatory variable may be subject to measurement

error (e.g. both reported income and wealth may be erroneous).

We might be discouraged by these findings, but fortunately there are solutions to these problems. The models in question, in which we suspect the presence of serious errors of measurement, may be estimated by techniques other than OLS regression. We will discuss those **instrumental variable** techniques, which may also be used to deal with problems of simultaneity, or two-way causality, in Chapter 15.