constructing 2SLS estimates of a wage equation that treats AFQT scores (an ability test used by the armed forces) as an endogenous control variable to be instrumented. The instruments for AFQT are early schooling (completed before military service), race, and family background variables. They estimated a system that can be described like this:

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_{0i}'\pi_{10} + \pi_{11}'\mathrm{Z}_i + \xi_{1i} \\ \mathrm{Y}_i &= \alpha_0'\mathrm{X}_{0i} + \alpha_0'\mathrm{X}_{1i} + \rho\hat{s}_i + [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)]. \end{aligned}$$

This looks a lot like manual 2SLS.

A closer look, however, reveals an important difference between the equations above and the usual 2SLS procedure: the covariates in the first and second stages are not the same. For example, Griliches and Mason included age in the second stage but not in the first, a fact noted by Cardell and Hopkins (1977) in a comment on their paper. This is a mistake. Griliches' and Mason's second stage estimates are not the same as 2SLS. What's worse, they are inconsistent where 2SLS might have been fine. To see why, note that the first-stage residual, $\mathrm{S}_i - \hat{s}_i$, is uncorrelated with X_{0i} *by construction* since OLS residuals are always uncorrelated with included regressors. But because X_{1i} is not included in the first-stage it is likely to be correlated with the first-stage residuals (e.g., age is probably correlated with the AFQT residual from the Griliches and Mason (1972) first stage). The inconsistency from this correlation spills over to all coefficients in the second stage. The moral of the story: put the same exogenous covariates in your first and second stage. If a covariate is good enough for the second stage, it's good enough for the first.

Forbidden Regressions

Forbidden regressions were forbidden by MIT Professor Jerry Hausman in 1975, and while they occasionally resurface in an under-supervised thesis, they are still technically off-limits. A forbidden regression crops up when researchers apply 2SLS reasoning directly to nonlinear models. A common scenario is a dummy endogenous variable. Suppose, for example, the causal model of interest is

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\mathrm{D}_i + \eta_i, \tag{4.6.1}$$

where D_i is a dummy variable for veteran status. The usual 2SLS first stage is

$$\mathrm{D}_i = \pi_{10}'\mathrm{X}_i + \pi_{11}'\mathrm{Z}_i + \xi_{1i}, \tag{4.6.2}$$

a linear regression of D_i on covariates and regressors.

Because D_i is a dummy variable, the CEF associated with this first stage, $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i]$, is probably nonlinear. So the usual OLS first-stage is an approximation to the underlying nonlinear CEF. We might, therefore, use a nonlinear first stage in an attempt to come closer to the CEF. Suppose that we use Probit to model $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i]$. The Probit first stage is $\Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, where π_{p0} and π_{p1} are Probit coefficients, and the fitted values are $\hat{\mathrm{D}}_{pi} = \Phi[\mathrm{X}_i'\hat{\pi}_{p0} + \hat{\pi}_{p1}'\mathrm{Z}_i]$. The forbidden regression in this case is the second stage equation created by substituting $\hat{\mathrm{D}}_{pi}$ for D_i:

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{\mathrm{D}}_{pi} + [\eta_i + \rho(\mathrm{D}_i - \hat{\mathrm{D}}_{pi})]. \tag{4.6.3}$$

The problem with (4.6.3) is that only OLS estimation of (4.6.2) is guaranteed to produce first-stage residuals that are uncorrelated with fitted values and covariates. If $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i] = \Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, then residuals from the nonlinear model will be asymptotically uncorrelated with X_i and $\hat{\mathrm{D}}_{pi}$, but who is to say that the first stage CEF is really Probit? With garden-variety 2SLS, in contrast, we do not need to worry about whether the first-stage CEF is really linear.[35]

A simple alternative to the forbidden second step, (4.6.3), avoids problems due to an incorrect nonlinear first stage. Instead of plugging in nonlinear fitted values, we can use the nonlinear fitted values *as instruments*. In other words, use $\hat{\mathrm{D}}_{pi}$ as an instrument for (4.6.1) in a conventional 2SLS procedure (as always, the exogenous covariates, X_i, should also be in the instrument list). Use of fitted values as instruments is the same as plugging in fitted values when the first-stage is estimated by OLS, but not in general. Nonlinear-fits-as-instruments has the further advantage that, if the nonlinear model gives a better approximation of the first-stage CEF than the linear model, the resulting 2SLS estimates will be more efficient than those using a linear first stage (Newey, 1990).

But here, too, there is a drawback. The nonlinear-fits-as-instruments procedure implicitly uses nonlinearities in the first stage as a source of identifying information. To see this, suppose the causal model of interest includes the instruments, Z_i :

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \gamma'\mathrm{Z}_i + \rho\mathrm{D}_i + \eta_i. \tag{4.6.4}$$

Now, with the first stage given by (4.6.2), the model is unidentified and conventional 2SLS estimates of (4.6.4) don't exist. But 2SLS estimates using X_i, Z_i, $\hat{\mathrm{D}}_{pi}$ do exist, because $\hat{\mathrm{D}}_{pi}$ is a nonlinear function of X_i and Z_i that is excluded from the second stage. Should you use this nonlinearity as a source of identifying information? We usually prefer to avoid this sort of back-door identification since its not clear what the underlying experiment really is.

As a rule, naively plugging in first-stage fitted values in nonlinear models is a bad idea. This includes models with a nonlinear second stage as well as those where the CEF for the first stage is nonlinear. Suppose,

[35] The insight that consistency of 2SLS estimates in a traditional SEM does not depend on correct specification of the first-stage CEF goes back to Kelejian (1971). Use of a nonlinear plug-in first-stage may not do too much damage in practice - a probit first-stage can be pretty close to linear - but why take a chance when you don't have to?

for example, that you believe the causal relation between schooling and earnings is approximately quadratic (as in Card's [1995] structural model). In other words, the model of interest is

$$\text{Y}_i = \alpha'\text{X}_i + \rho_1\text{s}_i + \rho_2\text{s}_i^2 + \eta_i. \tag{4.6.5}$$

Given two instruments, it's easy enough to estimate (4.6.5) treating both s_i and s_i^2 as endogenous. In this case, there are two first-stage equations, one for s_i and one for s_i^2. You need at least two instruments for this to work, of course. It's natural to use Z_i and its square (unless Z_i is a dummy, in which case you'll need a better idea).

You might be tempted, however, to work with a single first stage, say equation (4.6.2), and estimate the following second stage manually:

$$\text{Y}_i = \alpha'\text{X}_i + \rho_1\hat{s}_i + \rho_2\hat{s}_i^2 + [\eta_i + \rho_1(\text{s}_i - \hat{s}_i) + \rho_2(\text{s}_i^2 - \hat{s}_i^2)].$$

This is a mistake since $\hat{s}_i$ can be correlated with $\text{s}_i^2 - \hat{s}_i^2$ while $\hat{s}_i^2$ can be correlated with both $\text{s}_i - \hat{s}_i$ and $\text{s}_i^2 - \hat{s}_i^2$. On the other hand, as long as X_i and Z_i are uncorrelated with η_i in (4.6.5), and you have enough instruments in Z_i, 2SLS estimation of (4.6.5) is straightforward.

4.6.2 Peer Effects

A vast literature in social science is concerned with peer effects. Loosely speaking, this means the causal effect of group characteristics on individual outcomes. Sometimes regression is used in an attempt to uncover these effects. In practice, the use of regression models to estimate peer effects is fraught with peril. Although this is not really an IV issue *per se*, the language and algebra of 2SLS helps us understand why peer effects are hard to identify.

Broadly speaking, there are two types of peer effects. The first concerns the effect of group characteristics such as the average schooling in a state or city on individually-measured outcome variable. This peer effect links the average of one variable to individual outcomes as described by another variable. For example, Acemoglu and Angrist (2000) ask whether a given individual's earnings are affected by the average schooling in his or her state of residence. The theory of human capital externalities suggests that living in a state with a more educated workforce may make everyone in the state more productive, not just those who are more educated. This kind of spillover is said to be a *social return* to schooling: human capital that benefits everyone, whether or not they are more educated.

A causal model which allows for such externalities can be written

$$Y_{ijt} = \delta_j + \lambda_t + \gamma\overline{S}_{jt} + \rho\text{s}_i + u_{jt} + \eta_{ijt}, \tag{4.6.6}$$