Causal inference for binary regression

Austin Nichols

July 14, 2011

Selection and Endogeneity

The "selection problem" can refer to many distinct problems, but we examine the case where the observational units select their own treatment based on characteristics they "observe" and we do not (unobservables, in the parlance of econometrics), the situation in almost all nonexperimental data, leading to endogeneity of treatment and biased estimates of the impact of treatment. We frequently have a similar problem in experiments with imperfect compliance, where an experiment essentially generates observational data.

In a linear model, we can use panel methods to difference away unobservables that do not vary along some dimension (e.g. person-level characteristics that do not change over time), or instrumental variables (IV) and regression discontinuity (RD) methods to deal with other unobservables (see Nichols 2007,2008 for an overview).

A regression with a binary outcome y presents special difficulties. Panel methods typically require absurdly strong assumptions; the cross-sectional instrumental variables solution may not be obvious, particularly when the endogenous regressor of interest is also binary.

Two Examples

Supplemental Nutrition Assistance Program (SNAP)

Suppose we are interested in the impact of food assistance on the incidence of very low food security. If we compare those receiving SNAP (food stamps) to nonrecipients, the recipients look worse off. If we match, reweight, or control for observables, the recipients still look worse off. If we adopt a panel method, those about to receive SNAP **sometimes** look worse off than those who just started receiving it (Wilde and Nord 2005, Nord and Golla 2009), but this could be due to the Ashenfelter (1978) dip: applicants tend to be those who recently experienced a "transitory" dip in well-being and it may be that even if those starting SNAP receipt had been denied benefits, they would be have been better off in later months. Some kind of IV strategy seems in order (Ratcliffe and McKernan 2010).

Two Examples, cont.

Moving to Opportunity (MTO)

Suppose we are analyzing an experiment that gives public housing residents the chance to move to a low-poverty neighborhood, and we want to see if that affects their likelihood of employment three years later: the hypothesis is that they are adversely affected by lack of job networks, and a new neighborhood may solve that (Katz et al. 2000,2001; Kling et al. 2004; Kling et al. 2007). However, only 40 percent of the cases offered the chance take it up, and we want to know the impact of moving on later employment, not the impact of an offer (the "intention to treat" analysis, or ITT, which simply compares the mean outcomes of treatment and control groups). Those who are offered the chance and take it are different in unobserved ways from those who are offered the chance and don't take it; an IV strategy is called for.

Continuous X

Case with binary outcome and all endogenous regressors continuous: can simply use official command **ivprobit** (but see e.g. Altonji, Ichimura, and Otsu 2008 and others on relaxing the normality, linearity, and additivity assumptions).

Description

ivprobit *fits probit models where one or more of the regressors are endogenously determined. By default,* **ivprobit** *uses maximum likelihood estimation. Alternatively, Newey's minimum chi-squared estimator can be invoked with the twostep option. Both estimators assume that the endogenous regressors are continuous and are not appropriate for use with discrete endogenous regressors. See* **[R] ivtobit** *for tobit estimation with endogenous regressors and* **[R] probit** *for probit estimation when the model contains no endogenous regressors.*

Note: No mention of what to do with a binary outcome and a binary endogenous variable!

Binary regressor: general case

Suppose we have a set of individuals who receive treatment ($R = 1$) or not ($R = 0$) and another variable (or set of variables) A that affects the probability of treatment, and covariates X. Let $Z = (X, A)$. With a binary outcome Y, we can write a "threshold model" for some unspecified functions μ_Y and μ_R with unobservable error terms υ and ϵ:

$$Y_i = \mathbf{1}[(\mu_Y(R_i, X_i) > \upsilon_i]$$
$$R_i = \mathbf{1}[(\mu_R(Z_i) > \epsilon_i]$$

This model does not encompass every model of interest (the "non-additive errors" cases) but is already very general; Shaikh and Vytlacil (2011) and others cited there consider various bounds on average treatment effects under this general model. If we assume linearity of functions μ_Y and μ_R and homoskedastic bivariate normal errors for $(\upsilon, \epsilon) \sim N(0, \Sigma)$, we have the bivariate probit of Heckman (1978). With linearity but weaker assumptions on error distributions, various semiparametric estimators are possible.

Generality

Assuming the "threshold model" or additively separable error, per Heckman and Vytlacil (2005), also called "weak separability" of the observed regressors and the unobserved error term, is shown by Shaikh and Vytlacil (2011) to be equivalent to assuming that the expectations of potential outcomes are weakly increasing in the error term (Chesher 2005), or assuming the monotonicity restriction of Imbens and Angrist (1994).

Binary regressor: simple case

If we do maintain linearity and normality, we can write

$$Y_i = \mathbf{1}[(R_i d + X_i b) > \upsilon_i]$$
$$R_i = \mathbf{1}[(Z_i g) > \epsilon_i]$$
$$(\upsilon, \epsilon) \sim N(0, \Sigma)$$

where we normally assume there are some variables in Z not in X; call these A for variables that influence assignment to treatment but have no direct effect on the outcome $Pr(Y = 1)$, the bivariate probit analog of excluded instruments. Then we can estimate in Stata with e.g.:

```
biprobit (Y=X R) (R=X A)
```

Linear models

One approach is merely to estimate a linear probability model using IV (official **ivregress** or **ivreg2** from SSC), which is advocated by Angrist and Pischke (2009:198-204) and supported by much real-world experience comparing partial effects from more plausibly correct models to the partial effects from a linear probability model (see e.g. Wooldridge 2008, Katz et al. 2000 p.28 fn.34). IV has the advantage of easily interpreted coefficients measuring effects in the probability metric, but for those who are used to effect sizes measured in terms of log odds, it may be a less appealing option. In cases where response to treatment varies across individuals, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) point out that using linear IV gives an estimate of the average effect of treatment on the treated (ATT or TOT) for "compliers" (those induced to get treatment by assignment to the treatment group, or who have R=1 because A=1); see also Abadie (2003).

Linear and nonlinear models

However, while the linear IV model is a consistent estimator of an average effect of treatment, it is biased, and its small sample performance may be inferior to a correctly specified maximum likelihood model. The maximum-likelihood bivariate probit or **biprobit** approach (Heckman 1978) is simplest, and we will focus on it in simulations to follow, but there are also **gmm** and semiparametric solutions allowing heteroskedastic and nonnormal errors.

Common views on biprobit v. ivregress

Angrist and Pischke (2009:201) typify one form of received wisdom on **biprobit** and **ivregress**:

> *"Bivariate probit probably qualifies as harmless in the sense that it's not very complicated and easy to get right using packaged software routines."*

But constrast Freedman and Sekhon (2010). Angrist and Pischke (2009:202) again:

> *"Bivariate probit and other models of this sort can be used to estimate unconditional average causal effects and/or effects on the treated. In contrast, 2SLS does not promise you average causal effects, only* **local** *average causal effects."*

Experiments

The best case scenario for any instrumental variables approach is an experimental design with incomplete takeup of the treatment by the group assigned treatment, and no treatment in the control group.

Since assignment status is randomly assigned, the assignment dummy A is guaranteed to be a valid instrument, and its interaction with any exogenous variables will also be a valid instrument. However, it may still be a weak instrument, if takeup is low, but more importantly: the power of any instrumental variables strategy may be very low.

Power is a huge problem for IV strategies generally; too often researchers make a significant coefficient insignificant by instrumenting and then conclude the true effect is zero (even when the original confidence interval is entirely contained in the new IV confidence interval). For an experimental design, we typically have the opportunity to examine power before we collect the data–and to conduct simulations to determine which design is likely to have the greatest power!

biprobit

The **biprobit** approach, thanks to its stronger parametric assumptions, also allows the calculation of various probabilities using the bivariate normal distribution, for various marginal effects. However, note that one of its assumptions is a constant treatment effect d, not d_i, so that average treatment effects for any subpopulation are assumed to be the same as for any other subpopulation or the population (dgp). Still, one can calculate marginal effect of treatment for a subpopulation of "compliers" as an estimate of LATE. Note that the sample estimates of ATE or LATE are estimators for two estimands each: the sample ATE/LATE and the population ATE/LATE.

Whether we characterize our problem as estimating a sample or population ATE (or LATE if true mean treatment effects vary by subsample) seemingly does not affect our choice of estimator, but the mean squared error of an estimator is defined relative to one of these true effects; the rankings could change depending on our estimand.

Calculating ATE

How do we calculate the marginal effect of treatment after biprobit? Three "obvious" approaches: use **margins**, use **predict** to get probabilities, or use **binormal()** with predicted linear indices. The last is more correct, but all should give essentially the same answer.

```
biprobit (y=x R) (R=x A)
margins, dydx(R) predict(pmarg1) force
loc ATEm=el(r(b),1,1)
predict double xb2, xb2
preserve
ren R TR
g R=0
predict double p0, pmarg1
predict double xb0, xb1
replace R=1
predict double p1, pmarg1
predict double xb1, xb1
g double dp=p1-p0
su dp, mean
loc ATE1=r(mean)
su dp if TR==1, mean
loc TOT1=r(mean)
loc r=e(rho)
gen double pdx=(binormal(xb1,xb2,‘r’)-binormal(xb0,xb2,‘r’))/normal(xb2) if TR==1
su pdx, mean
loc TOT2=r(mean)
qui replace pdx=normal(xb1)-normal(xb0)
su pdx, mean
loc ATE2=r(mean)
* ATE2 same as ATE1 above
```

Simulation

Simulation setup: one (or more) excluded binary instrument(s), covariate(s), various correlation structures, sample sizes, random coefficients, heteroskedasticity.

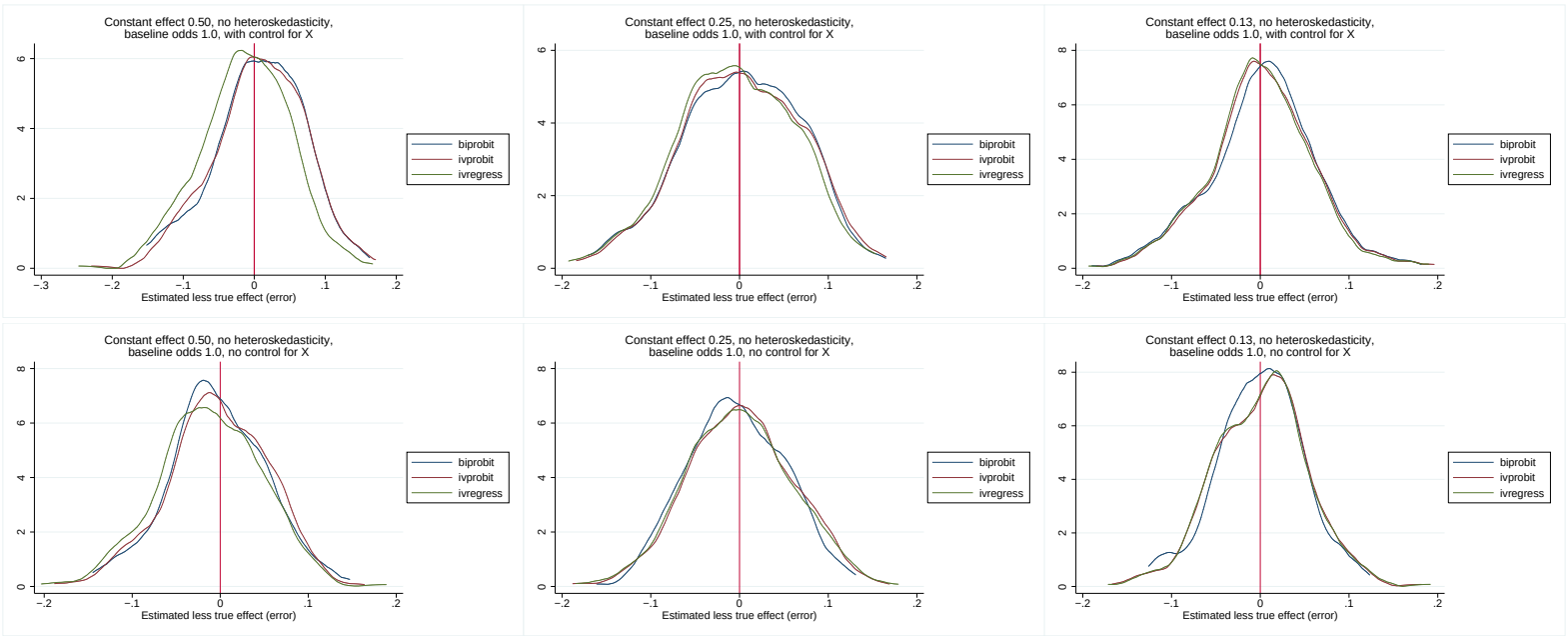
```
mat c=(1,.5,.5 \ .5,1,0 \ .5,0,1)
drawnorm x e z, n(1000) corr(c) clear seed(2)
qui su e
replace e=e/r(sd)
g u=rnormal()
g A=uniform()<.5
g R=A*(x+u>0)
g y=(R/2+e)>0
g y1=R/2+e>0
g y0=e>0
g dy=y1-y0
```

ATE=normal(.5)-.5=0.191

```
ta R dy if A==1, row nokey
           |          dy
         R |         0          1 |     Total
-----------+----------------------+----------
         0 |       241          0 |       241
           |    100.00       0.00 |    100.00
-----------+----------------------+----------
         1 |       208         49 |       257
           |     80.93      19.07 |    100.00
-----------+----------------------+----------
     Total |       449         49 |       498
           |     90.16       9.84 |    100.0
ta R y, row nokey
           |           y
         R |         0          1 |     Total
-----------+----------------------+----------
         0 |       419        324 |       743
           |     56.39      43.61 |    100.00
-----------+----------------------+----------
         1 |        49        208 |       257
           |     19.07      80.93 |    100.00
-----------+----------------------+----------
     Total |       468        532 |     1,000
           |     46.80      53.20 |    100.00
```

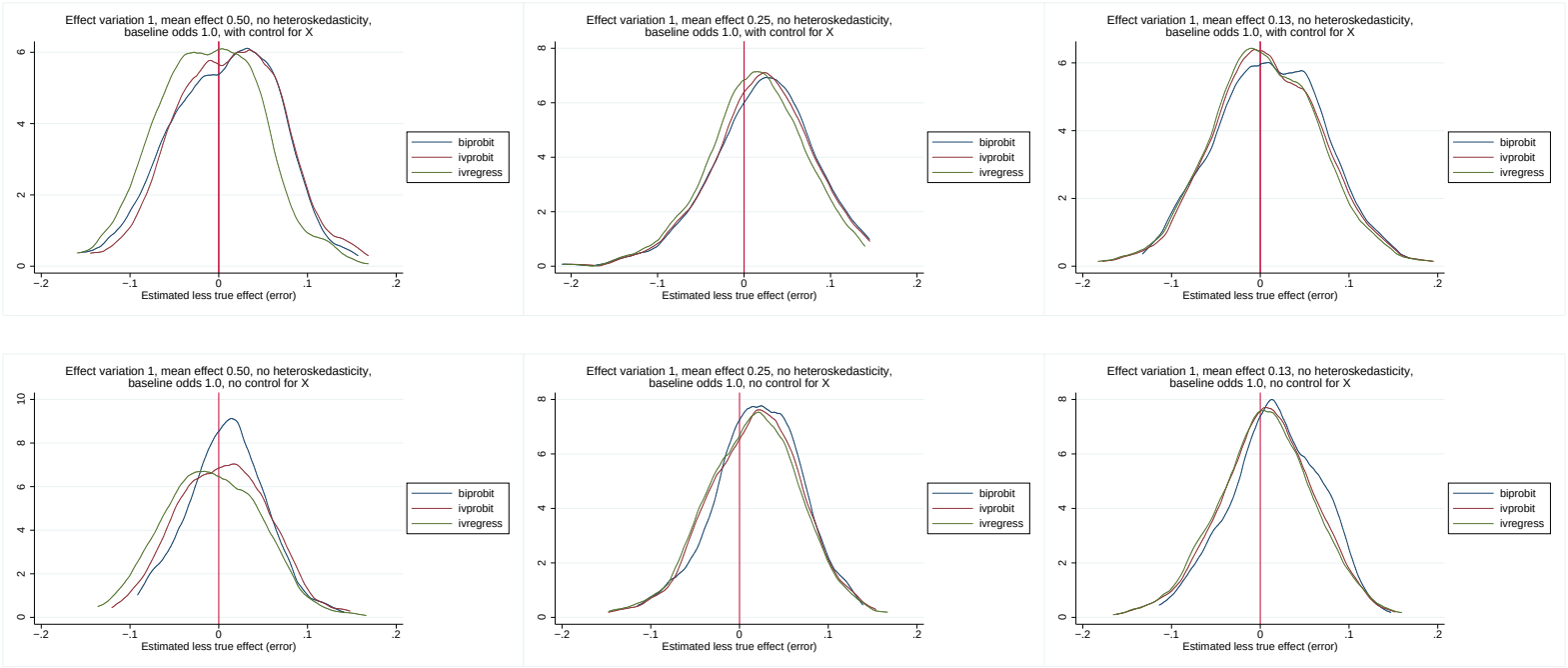
MSE

Estimating population ATE, compare MSE of ivprobit for binary treatment, linear IV, and biprobit, with and without controls for a covariate X that affects treatment takeup probability and the outcome:



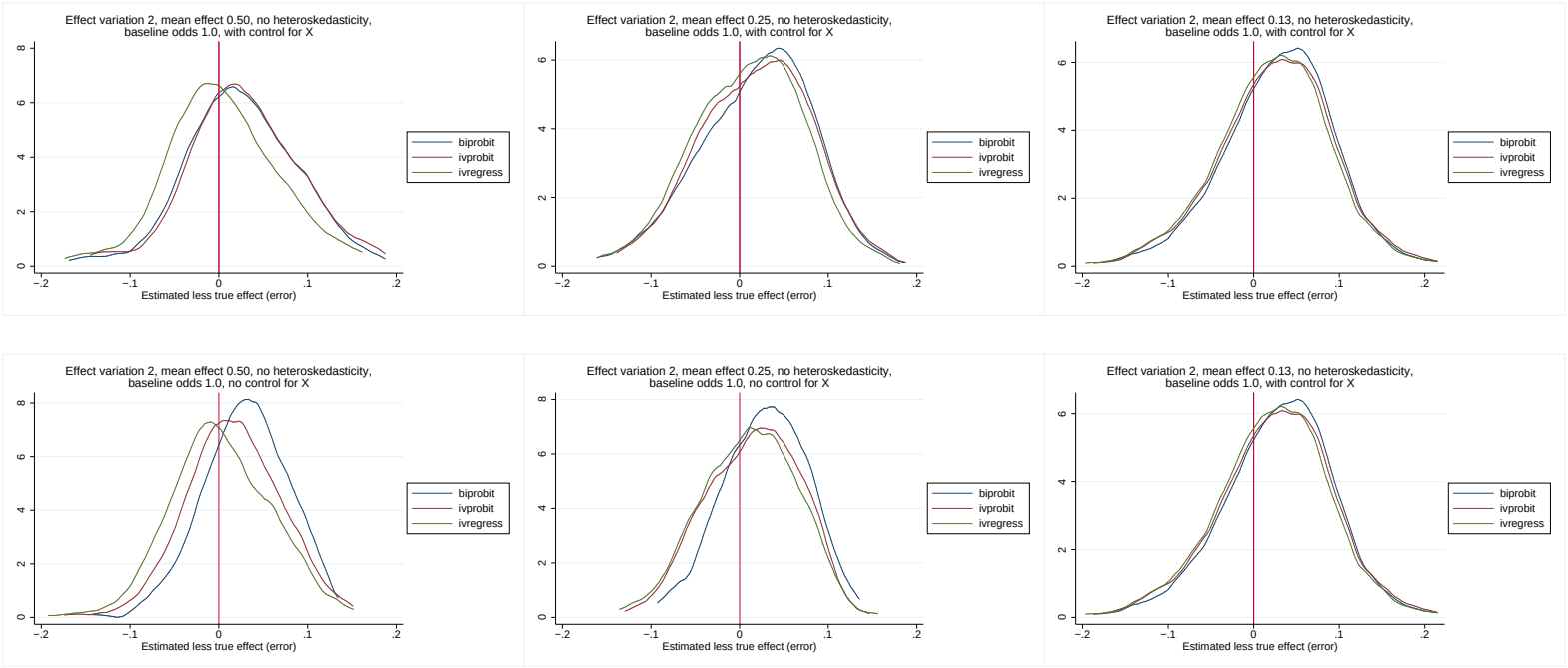
MSE

With random coefficients (SD=.5):



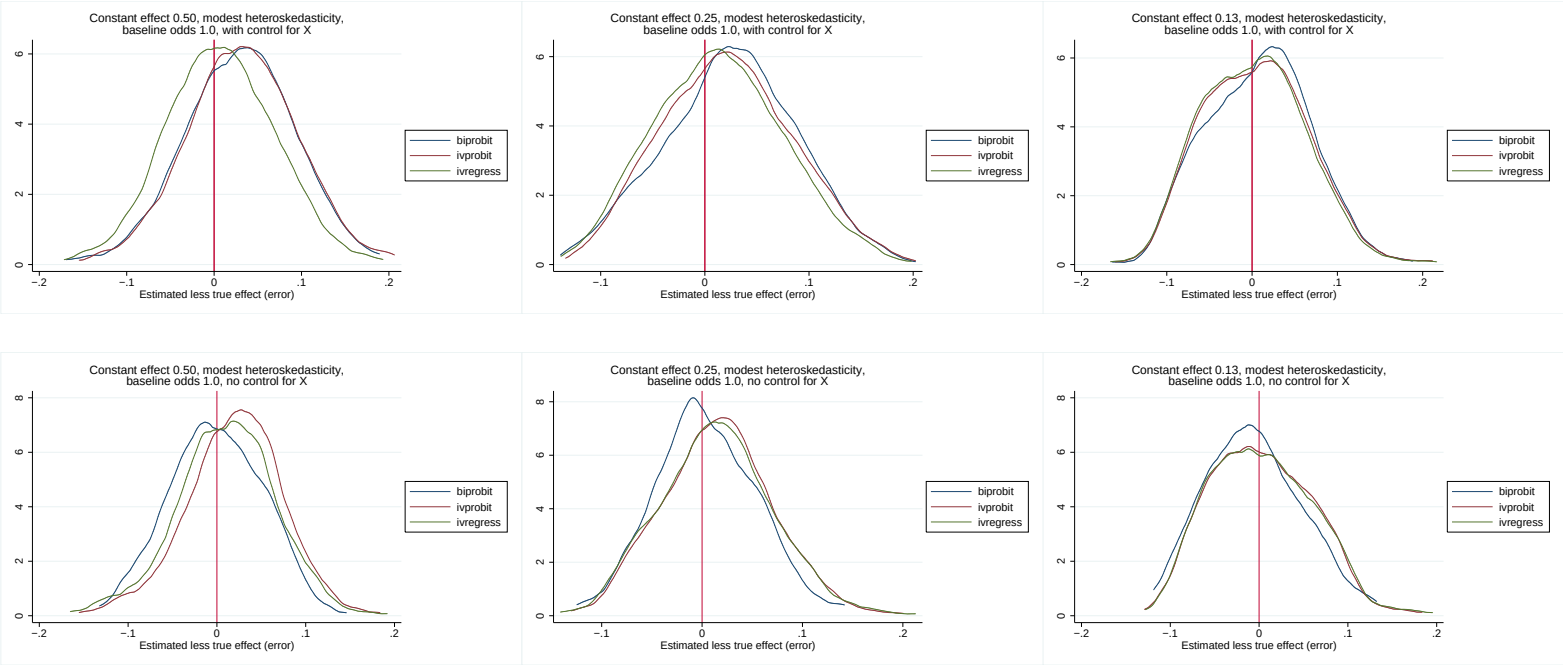
MSE

With random coefficients (SD=1):



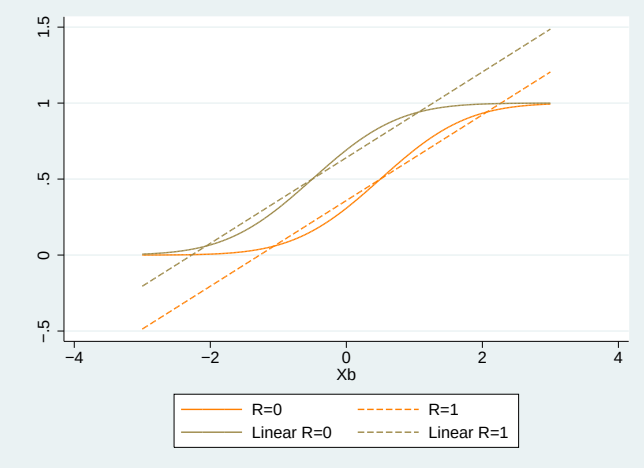
MSE

With heteroskedasticity:



Size of tests

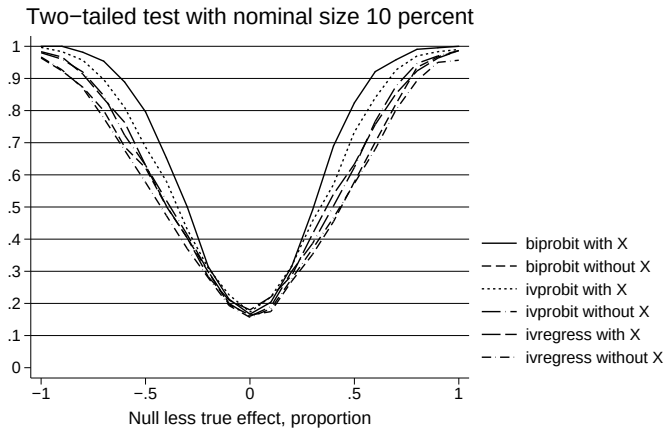
Similar MSE in many cases; patterns are similar for ATE and ATT/TOT, sample and population estimands. Some indications of finite-sample bias away from zero in the bivariate probit and toward zero for linear IV. Same pattern reported in Angrist and Pischke (2009:203).



Size of tests

Bias and MSE are low for various estimators, and the power curve looks similar for each. But it bottoms out well above the nominal size, in the range 17 to 20 percent for a test with nominal size 10 percent and in the range 7 to 12 percent for a test with nominal size 5 percent.

I.e. standard errors are underestimated; bootstrap standard errors are also too small (in many of these settings we should expect no improvement from bootstrap—imagine resampling with no continuous covariates and stratifying by A and R). We will reject a true null hypothesis at much higher rates than our nominal alpha using any of these estimators. One easy solution: adopt a lower size of test, say 3 percent instead of 5.



Alternatives

Why would you **not** want to use **biprobit**, aside from feeling uncomfortable about the strong distributional and functional form assumptions? One reason is that it is a pain to estimate; it frequently takes 10 or 20 times as long as other similar models and Freedman and Sekhon (2010) disparage the ability of Stata and R to find the maximum of the likelihood.

From my own experience estimating millions of **biprobit** regressions, I can offer:

- **Do** use the **difficult** option, which can result in a (circa) threefold speed improvement.
- **Don't** use the **from** option, which can negate the above speed improvement.
- Man, would **biprobit** benefit from some kind of specialized maximizer—it is **slow**!

Alternatives

Another reason **not** want to use **biprobit** is that you suspect endogeneity not only in a single binary regressor; or you want to interact that regressor with exogenous covariates, creating additional endogenous covariates.

There is a natural generalization of **biprobit** with more than one endogenous variable: **cmp** (Roodman 2009) can handle a variety of models using a maximum likelihood approach. As with **biprobit**, one must make strong functional form and distributional assumptions with this approach.

There is also a **gmm** approach, if one defines the proper population moments (Wilde 2008). Or a Bayesian approach (McCarthy and Tchernis 2010). Either can handle multiple endogenous covariates of various types, with additional assumptions.

One can also use a semiparametric model; we will examine these in the single binary regressor case, but they are more easily extended to multiple types of regressors; see esp. Abrevaya, Hausman, and Khan (2009).

Heteroskedasticity

The effect of even modest heteroskedasticity on **biprobit** could be disastrous, but my simulations indicate that **biprobit** is remarkably robust to modest heteroskedasticity, and **ivprobit** slightly less so. Interestingly, **biprobit** and **ivprobit** are both also remarkably robust to variability in the treatment effect (random coefficients) in my simulations.

That is, under heteroskedasticity and random coefficients, in the parameter space I searched, the results are all qualitatively similar, though MSE is higher when required assumptions are violated.

Nonnormal errors

Chiburis, Das, and Lokshin (2011) run simulations similar to mine, and find that when there are no covariates, **biprobit** outperforms IV for sample sizes below 5000, and with a continuous covariate, **biprobit** outperforms IV in all of their simulations. They note that **biprobit** performs especially well when the treatment probability is close to 0 or 1, where linear methods are more likely to produce infeasible estimates. They further note that results of Bhattacharya, Goldman, and McCaffrey (2006), who find **biprobit** robust to non-normality of error terms, do not hold up for all parameter values, but offer "no clear guidance on the parameter values under which the expected bias will be worse." They also recommend a score test due to Murphy (2007) as a specification test (see also Chiburis 2010b; Lucchetti and Pigini 2011), which rejects the model when there is excess kurtosis or skewness in the error distributions. **The impact of pretesting is an important avenue for future research.**

Semiparametric estimators

If we are willing to assume linearity in the basic threshold model, so we have two linear index functions Xb and Zg, but we are not willing to assume a homoskedastic bivariate normal error vector, There are a number of semiparametric estimators, e.g. Abadie (2003), Chiburis (2010a), Shaikh and Vytlacil (2011), Abrevaya, Hausman, and Khan (2009), some offering point identification and some bounds on treatment effects. There is also a semiparametric double-index model proposed by Klein, Shen, and Vella (2010), who follow Klein and Vella (2005) using a similar semiparametric strategy for a **treatreg** type estimator, in turn based on a trick from Klein and Spady (1993):

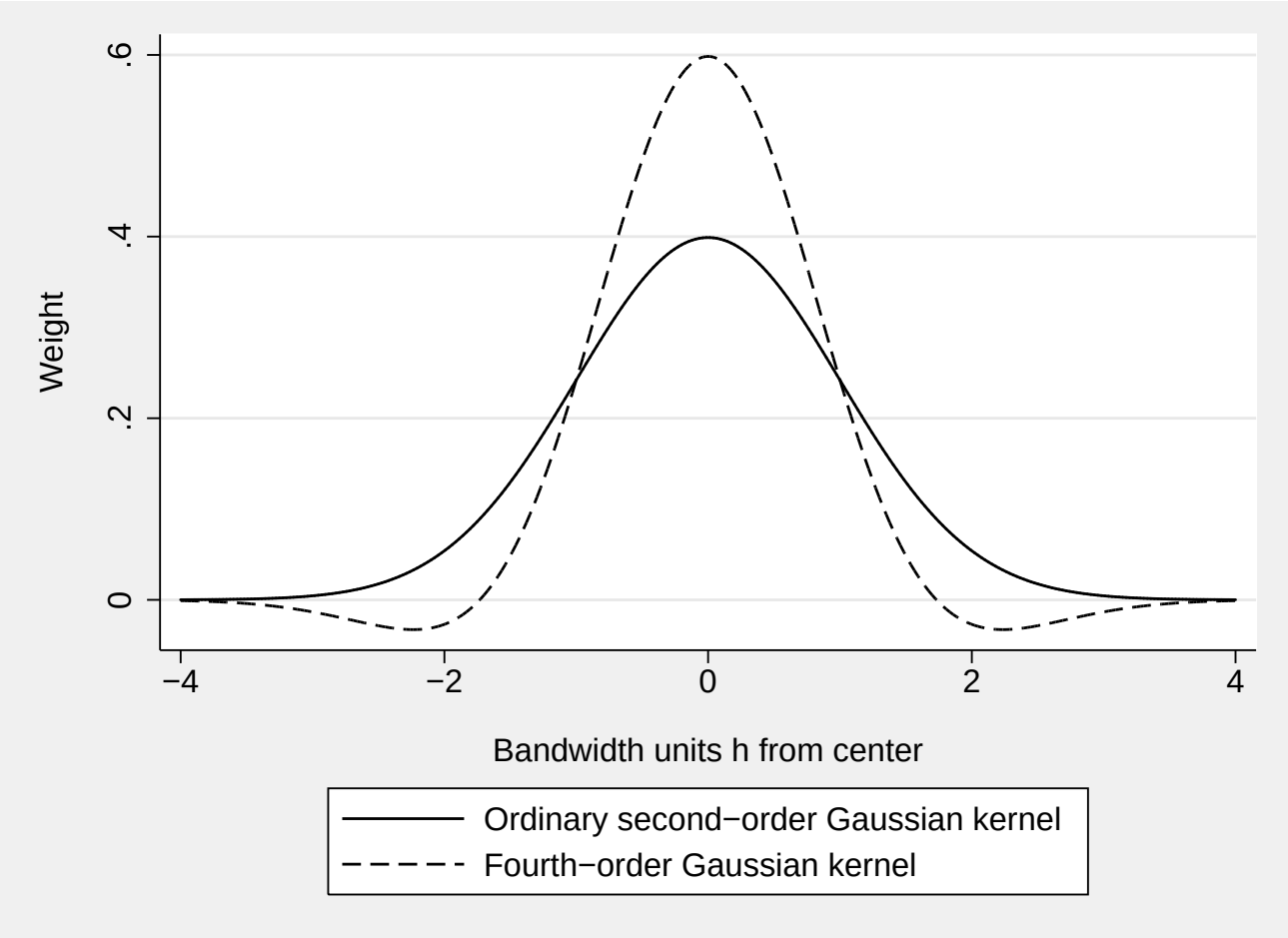
$$E[y|X] = Pr(y = 1|Xb) = Pr(y = 1)\frac{f_1(Xb)|_{y=1}}{f(Xb)}$$

so that the ratio of two nonparametric estimates of the density of the linear index Xb gives an estimate of the probability. The Klein and Spady (1993) estimator attains the Semiparametric Efficiency Bound. See also equation 3(6) in Efron (2003) and Fix and Hodges (1951), or **[MV] discrim knn**.

Higher-order kernels

Usually this literature relies on "bias-reducing kernels" or "higher-order kernels" which have some desirable theoretical properties but can exhibit atrocious small-sample properties, e.g. because they produce negative estimates for a density or a probability. The work by Klein et al. instead uses "local smoothing" with a regular kernel (i.e. a density function symmetric around zero) and trimming (the trimming essentially removes cases where the denominator of that ratio of nonparametric estimates of the density of the linear index Xb may be close to zero).

Higher-order kernels illustrated



Semiparametric estimators MSE

The semiparametric estimators, because they do not assume homoskedastic bivariate normal errors, perform better when those assumptions are violated, and perform about as well when the assumptions are true. However, there are not huge differences between any of the estimators, in my simulations.

Identification without instruments

A similar semiparametric estimator of a double index model can produce a version of IV with a binary endogenous regressor (the **treatreg** environment) where exclusion restrictions are not required (Klein and Vella 2005,2010), but we instead assume that the functional form of heteroskedasticity is in a family of linear index functions. Here we assume Xb is the linear index that mean R depends on ($E(R|X) = F(Xb)$), but that the error variance is $exp(Zg)$, so the two linear indices are again Xb and Zg. An ordinary 2SLS model can include residuals from the first stage only if they are functions of variables excluded from the second stage, but the Klein and Vella (2005,2010) estimator relies on heteroskedasticity for identification. This may sound a bit like a **heckman** selection or **treatreg** model where we rely on functional form for identification, and no additional excluded variables that determine selection. But can offer substantially improved performance if the assumption on the functional form of heteroskedasticity is correct.

Weak IV

There is now a voluminous literature on the dangers of weak instruments, mainly inflated size (overrejection of the null) and bias, due to e.g. Bound, Jaeger, and Baker (1993, 1995), Staiger and Stock (1997), Stock, Wright, and Yogo (2002), and Stock and Yogo (2005). But we have little evidence related to nonlinear models. Since the tests proposed by Stock and Yogo (2005) characterize correlations in the first stage, it is plausible (though unproven) that they work well for any model with a linear first stage and continuous excluded instruments. What about a nonlinear first stage, such as the probit of our current example? Binary excluded instruments?

In our prototypical "best" case scenario, we cannot run a first stage probit, because A=0 implies R=0. That is, no one gets treated who was not assigned to treatment, so a probit cannot be used for the general case to assess the strength of instruments. Various alternatives are possible, but the linear model is a useful starting point.

Half assigned to treatment, half take it up

```
. ta A R
           |           R
         A |         0          1 |     Total
-----------+----------------------+----------
         0 |       500          0 |       500
         1 |       250        250 |       500
-----------+----------------------+----------
     Total |       750        250 |     1,000

. ivreg2 y (R=A)
```

(output omitted)

```
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              499.000
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.000
                                                (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         R
Excluded instruments: A
------------------------------------------------------------------------------
```

Weak IV for biprobit

Note that in the linear model, if the first-stage coefficient for A is one half and the constant is zero, the Wald F statistic is N/2-1, linear in sample size. Linearity of the test statistic in sample size makes it appealing for *ex ante* power analysis.

Note, second, that the critical values for that first-stage Wald F statistic determine the expected actual size of a nominal 5 percent test. The critical value is not 10, as many people still believe, based on work from 15 years ago (Staiger and Stock 1997). Instead, to have an expected size of not more than 10 percent with a nominal size of 5 percent, we need a first-stage Wald F statistic of at least 16 (Stock and Yogo 2005), but this seems like too low a standard; perhaps we really should be aiming for 6 or 7 percent.

Third, those critical values were derived via simulation for a single continuous endogenous variable, and a single continuous excluded instrument, and therefore are wholly inappropriate for our present case. We already know that a first-stage Wald F statistic on the order of 500 gives an expected size roughly twice the nominal size in many of the binary cases.

Weak IV for biprobit

Note that limiting bias of IV to some percentage of ordinary regression is not the binding constraint on instrument strength here; rather the incorrect size is the main issue. If the first-stage Wald F statistic were an adequate measure of weak instruments, then we could run simulations in order to say: if we want to ensure size is no more than q percent with a nominal size of 10 percent, with v excluded binary instruments, then we should observe linear first stage Wald test statistics for excluded instruments on the order of $f(q, v)$, where $f(\cdot)$ is determined by simulation.

Unfortunately, in simulations I have run, the rejection rate is not smoothly declining toward the nominal size in first stage Wald test statistics, and there are no reliable critical values. A new measure of weak instruments for a binary first stage with binary instruments seems to be needed.

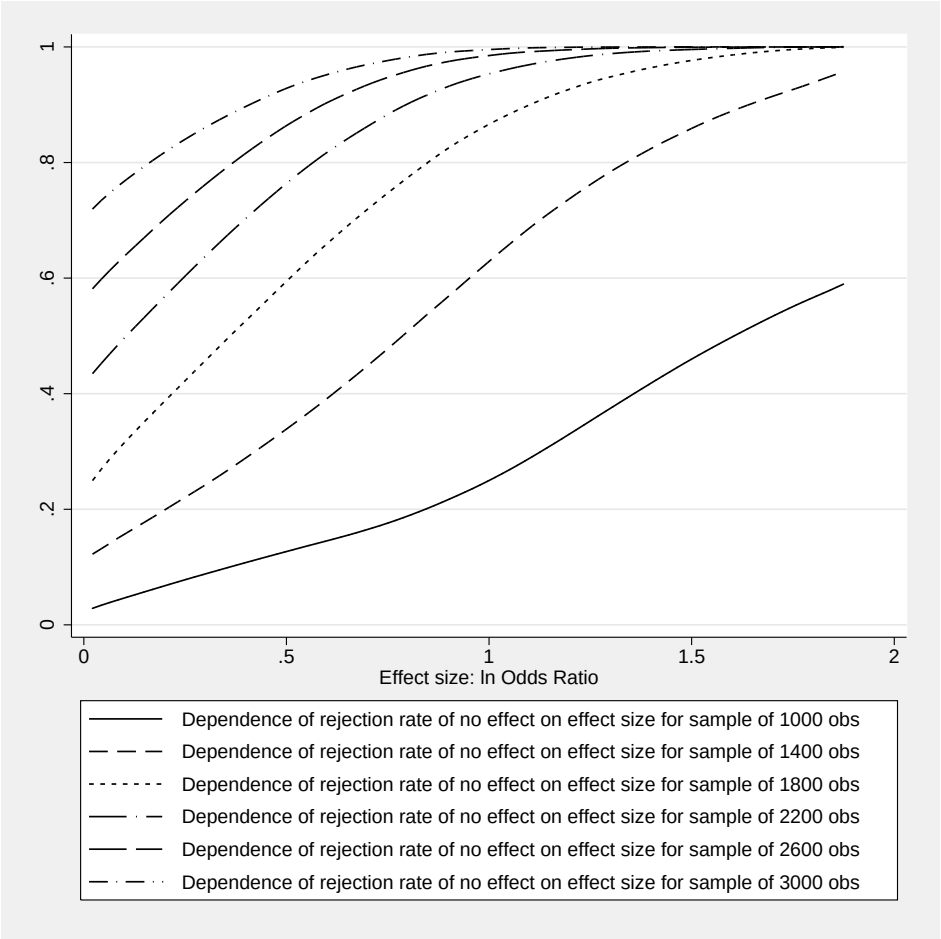
Power analysis for experimental designs

The usual approach to power analysis for social experiments, e.g. in Orr (1999:115-120), referenced in e.g. Kling et al. (2004, p.14 fn.32), is to compute the IV estimate as the ratio of an Intention-To-Treat (ITT) parameter estimate (from a regression of the outcome on the assignment status dummy) divided by the proportion treated (the parameter from a regression of a treatment status dummy on the assignment status dummy), assumed nonstochastic. The ratio comes from the Wald estimator for IV. This is inappropriate in the binary setting: if we are planning to use **biprobit** for analysis, it should used to analyze power. Researchers will commonly claim that the TOT estimate is twice the ITT estimate where takeup was one half, which implies they will use a linear model to analyze binary outcomes and binary treatments. Those extending the ratio approach of IV to power analysis also typically assume that the takeup rate is a fixed proportion, when it is clearly stochastic. See for an example Orr (1999:115-120), the MTO literature (Kling et al. 2004, p.14 fn.32), and Quigley and Raphael (2008) on the power of the MTO experiment.

Power analysis for biprobit

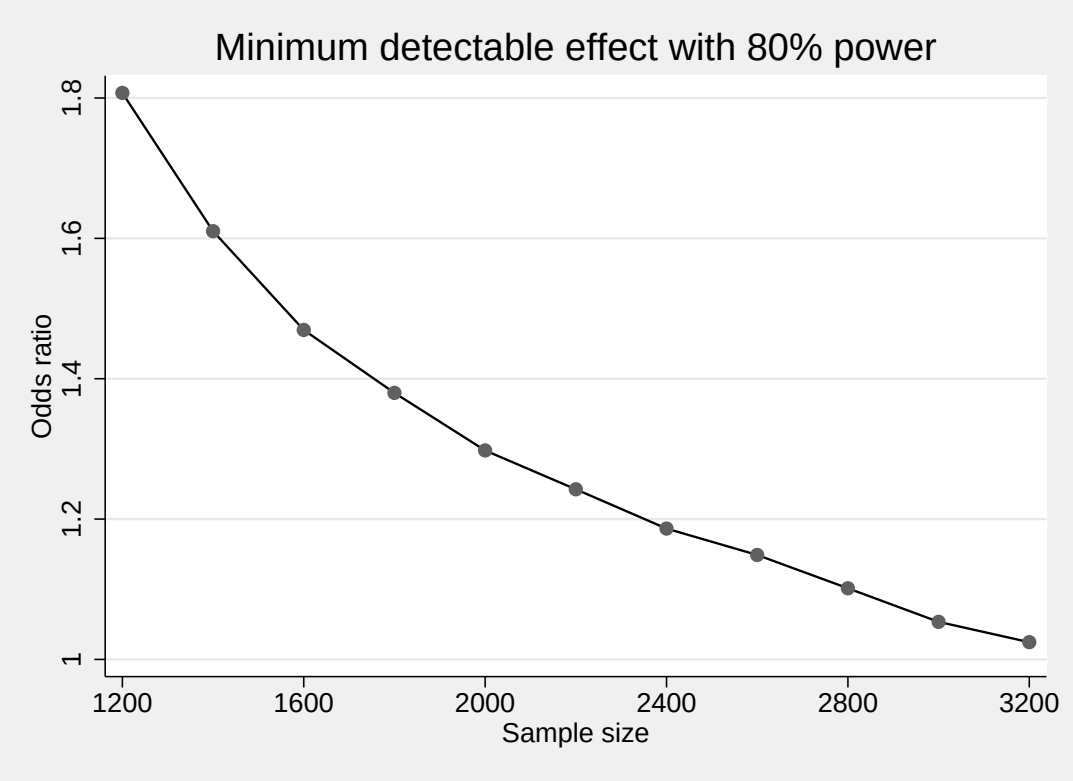
It should be clear that calculations of power, or minimum detectable effects, or required sample sizes, for an experiment or quasi-experiment with a binary outcome and a binary treatment R instrumented by A, must take into account the analysis design. It is straightforward to specify assumed effect sizes and sample sizes, estimation and test and alpha (size of test), then calculate power in a **simulation**. For example, suppose we want to achieve power of 80 percent, using **biprobit**, and we anticipate we will assign treatment to half of our sample and only half of those assigned to treatment take it up. We can trace out the empirical rejection rates.

Power analysis for biprobit, rejection rates



Power analysis for biprobit, minimum detectable effects

Then interpolate to construct minimum detectable effects at various sample sizes (assumes use of analytic SE, but bootstrap is similar):



Conclusions

The first bit of advice in regards to binary regression with a binary endogenous variable is usually one of:

- Use linear IV, and you'll get robust consistent estimates of the ATT.
- Use bivariate probit, and you'll get efficient estimates of the ATE.

Most econometricians would probably prefer a more plausibly correct model that requires fewer assumptions than either of the above.

My simulations indicate that many alternative models give remarkably precise estimates, with low MSE for both sample and population treatment effects. However, I find that the standard errors tend to be dramatically underestimated, even assuming a well-behaved homoskedastic normal error term, if instruments are not exceptionally strong. This leads to overrejection of the null in each model, and we should approach inference in the binary case very cautiously.

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." Journal of Econometrics 113: 231-63.

Abrevaya, Jason; Jerry A. Hausman; and Shakeeb Khan. 2009. "Testing for causal effects in a generalized regression model with endogenous regressors." Working paper.

Altonji, Joseph G.; Hidehiko Ichimura; and Taisuke Otsu. 2008. "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables." NBER Working Paper No. 14161.

Angrist, Joshua D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." Journal of Business and Economic Statistics, 19(1): 2-16.

Angrist, Joshua D. and Alan B. Krueger. 2000. "Empirical Strategies in Labor Economics." in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Angrist, Joshua D.; Guido W. Imbens; and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-472.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton, NJ: Princeton University Press.

Ashenfelter, Orley. 1978. "Estimating the effect of training programs on earnings." *Review of Economics and Statistics* 60:47-57.

Athey, Susan and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74(2): 431-497.

Bhattacharya, Jay; Dana P. Goldman; and Daniel F. McCaffrey. 2006. "Estimating Probit Models with Self-selected Treatments." Statistics in Medicine, 25(3): 389-413.

Bhattacharya, Jay; Azeem Shaikh; and Edward Vytlacil. 2005. "Treatment effect bounds: an application to Swan-Ganz catheterization." NBER working paper 11263.

Bhattacharya, Jay; Azeem Shaikh; and Edward Vytlacil. 2008. "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization." *American Economic Review* 98(2): 35156.

Baum, Christopher F.; Mark E. Schaffer; and Steven Stillman. 2007. "Enhanced routines for instrumental variables/GMM estimation and testing." Stata Journal 7(4): 465-506.

Bound, John; David A. Jaeger; and Regina Baker. 1993. "The Cure Can Be Worse than the Disease: A Cautionary Tale Regarding Instrumental Variables." NBER Technical Working Paper No. 137.

Bound, John; David A. Jaeger; and Regina Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of the American Statistical Association*, 90(430), 443-450.

Chesher, Andrew. 2003. "Identification in nonseparable models. "Econometrica, 71: 1405-1441.

Chesher, Andrew. 2005. "Nonparametric identification under discrete variation." Econometrica, 73: 1525-1550.

Chao, John C. and Norman R. Swanson. 2005. "Consistent Estimation with a Large Number of Weak Instruments." *Econometrica*, 73(5), 1673-1692. Working paper version available online.

Chiburis, Richard. 2010a. "Semiparametric Bounds on Treatment Effects." *Journal of Econometrics,* 159(2):267-275.

Chiburis, Richard. 2010b. "Score Tests of Normality in Bivariate Probit Models: Comment." Working paper.

Chiburis, Richard; Jishnu Das; and Michael Lokshin. 2011. "A Practical Comparison of the Bivariate Probit and Linear IV Estimators." World Bank Policy Research Working Paper 5601.

Efron, Bradley. 2003. "Robbins, Empirical Bayes and Microarrays." *The Annals of Statistics* 31,(2): 366-378.

Freedman, David A. and Jasjeet S. Sekhon. 2010. "Endogeneity in Probit Response Models." Political Analysis, 18(2): 138-150.

Fix, E., and J. L. Hodges. 1951. "Discriminatory analysis: Nonparametric discrimination, consistency properties." In Technical Report No. 4, Project No. 21-49-004. Randolph Field, Texas: Brooks Air Force Base, USAF School of Aviation Medicine. Reprinted 1989 in International Statistical Review 57(3): 238-247.

Heckman, James J. 1976. "The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models." Annals of Economic and Social Measurement, 5: 475-492.

Heckman, James J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica,* 46(6): 931-959.

Heckman, James J. and Edward J. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." Proceedings of the National Academy of Sciences of the United States of America 96:4730-34.

Heckman, James J. and Edward J. Vytlacil. 2000. "The Relationship between Treatment Parameters within a Latent Variable Framework." Economics Letters 66:33-39.

Heckman, James J. and Edward J. Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica,* 73: 669-738.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62(2): 467-75

Imbens, Guido W. and Whitney K. Newey. 2002. "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." NBER Technical Working Paper No. 285.

Katz, Lawrence F.; Jeffrey R. Kling; and Jeffrey B. Liebman. 2000. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." NBER Working Paper No. 7973.

Katz, Lawrence F.; Jeffrey R. Kling; and Jeffrey B. Liebman. 2001. "Moving To Opportunity In Boston: Early Results Of A Randomized Mobility Experiment." Quarterly Journal of Economics, 116(2):607-654.

Klein, Roger W. and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." Econometrica, 61(2):387-421.

Klein, Roger W.and Francis Vella. 2005. "Estimating a class of triangular simultaneous equations models without exclusion restrictions." IFS cemmap Working Paper CWP08/05.

Klein, Roger W.and Francis Vella. 2010. "Estimating a class of triangular simultaneous equations models without exclusion restrictions." *Journal of Econometrics,* 154(2): 154-164.

Klein, Roger W.; Chan Shen; and Francis Vella. 2010. "Triangular Semiparametric Models Featuring Two Dependent Endogenous Binary Outcomes." Unpublished working paper.

Kling, Jeffrey R.; Jeffrey B. Liebman; Lawrence F. Katz; and Lisa Sanbonmatsu. 2004. "Moving To Opportunity And Tranquility: Neighborhood Effects On Adult Economic Self-Sufficiency And Health From A Randomized Housing Voucher Experiment." Princeton University Industrial Relations Section Working Paper 481.

Kling, Jeffrey R.; Jeffrey B. Liebman; and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." Econometrica, 75(1):83-119.

Lee, Lung-Fei. 1992. "Amemiya's Generalized Least Squares and Tests of Overidentification in Simultaneous Equation Models with Qualitative or Limited Dependent Variables." Econometric Reviews, 11(3): 319-328.

Lucchetti, Riccardo and Claudia Pigini. 2011. "Conditional Moment Tests for Normality in Bivariate Limited Dependent Variable Models: a Monte Carlo Study." Quaderni Di Ricerca N. 357.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

McCarthy, Ian M. and Rusty Tchernis. 2010. "On the Estimation of Selection Models when Participation is Endogenous and Misclassified." Working Paper.

Murphy, Anthony. 2007. "Score Tests of Normality in Bivariate Probit Models." *Economics Letters,* 95(3): 374-379.

Newey, Whitney K. 1987. "Efficient Estimation of Limited Dependent Variable Models with Endogeneous Explanatory Variables". Journal of Econometrics, 36: 231-250.

Newey, Whitney K.; James L Powell; and Francis Vella. 1999. "Nonparametric estimation of triangular simultaneous equations models." *Econometrica,* 67(3)

Nichols, Austin. 2007. "Causal inference with observational data." Stata Journal 7(4): 507-541.

Nichols, Austin. 2008. "Erratum and discussion of propensity-score reweighting." Stata Journal 8(4):532-539.

Nord, Mark, and Anne Marie Golla. 2009. "Does SNAP Decrease Food Insecurity? Untangling the Self-Selection Effect." Washington, DC: USDA, Economic Research Service, Economic Research Report Number 85, October.

Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage.

Quigley, John, and Steven Raphael. 2008. "Neighborhoods, economic self-sufficiency, and the MTO." Brookings-Wharton Papers on Urban Economics Affairs, 8. Washington, DC: Brookings Institution.

Ratcliffe, Caroline and Signe-Mary McKernan. 2010. "How Much Does SNAP Reduce Food Insecurity?" Washington, DC: Urban Institute [http://www.urban.org/publications/412065.html]

Roodman, David. 2009. "Mixed-process models with cmp." Presentation at Stata Conference, DC 2009.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66: 688-701.

Shaikh, Azeem M. and Edward J. Vytlacil. 2011. "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables." Econometrica 79(3):949955, May 2011

Staiger, Douglas and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65, 557-586.

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." Ch. 5 in J.H. Stock and D.W.K. Andrews (eds), *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Cambridge University Press. Originally published 2001 as NBER Technical Working Paper No. 284; newer version (2004) available at Stock's website.

Stock, James H.; Jonathan H. Wright; and Motohiro Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business and Economic Statistics*, 20, 518-529. Available from Yogo's website.

Wilde, Joachim. 2008. "A note on GMM estimation of probit models with endogenous regressors." Statistical Papers 49(3):471484.

Wilde, Parke, and Mark Nord. 2005. "The Effect of Food Stamps on Food Security: A Panel Data Approach." Review of Agricultural Economics 27(3): 425-432.

Wooldridge, Jeffrey. 2008. "Inference for partial effects in nonlinear panel-data models using Stata." Presentation at Summer 2008 Stata Meetings.