ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Chesher, Andrew; Rosen, Adam

Working Paper

What do instrumental variable models deliver with discrete dependent variables?

cemmap working paper, No. CWP10/13

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Chesher, Andrew; Rosen, Adam (2013) : What do instrumental variable models deliver with discrete dependent variables?, cemmap working paper, No. CWP10/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.1013

This Version is available at:
https://hdl.handle.net/10419/79505

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

cemmap
centre for microdata methods and practice

What do instrumental variable models deliver with discrete dependent variables?

Andrew Chesher
Adam Rosen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP10/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

What do instrumental variable models deliver with discrete dependent variables?*

Andrew Chesher and Adam M. Rosen
CeMMAP & UCL

March 2013

KEYWORDS: Discrete endogenous variables, Endogeneity, Incomplete models, Instrumental variables, Set Identification, Structural econometrics.
JEL CODES: C10, C14, C50, C51.

1 Introduction

We study models with discrete endogenous variables and compare the use of two stage least squares (2SLS) in a linear probability model with bounds analysis using a nonparametric instrumental variable model.

2SLS has the advantage of providing an easy to compute point estimator of a slope coefficient which can be interpreted as a local average treatment effect (LATE). However, the 2SLS estimator does not measure the value of other useful treatment effect parameters without invoking untenable restrictions.

The nonparametric instrumental variable (IV) model has the advantage of being weakly restrictive, so more generally applicable, but it usually delivers set identification. Nonetheless it can be used to consistently estimate bounds on many parameters of interest including, for example, average treatment effects. We illustrate using data from Angrist & Evans (1998) and study the effect of family size on female employment.

The instrumental variable model takes the form:

$$Y_1 = h(Y_2, W, U), \tag{1.1}$$

where Y_1 is an outcome of interest, Y_2 is an endogenous variable, and W is a vector of exogenous variables. U denotes unobserved heterogeneity. In addition to (Y_1, Y_2, W), realizations of an instrumental variable (IV) Z which does not enter (1.1) are available. IV models limit the dependence of U and Z imposing for example stochastic or mean independence restrictions. The models are *incomplete* in the sense that details of the process generating values of endogenous Y_2 are not specified. Structural equation (1.1) may for example be one of a system of simultaneous equations, the structure of the others being left unspecified.

A linear specification, $Y_1 = \alpha + Y_2\beta + W\gamma + U$, is used in many applications. If the required expectations exist, then a mean independence restriction $E[U|w, z] = 0$ delivers the relationship $E[Y_1|w, z] = \alpha + E[Y_2|w, z]\beta + w\gamma$, and point identification of the parameters follows under a simple rank condition.

In this paper we focus on models with discrete Y_1 and Y_2, henceforth referred to as "the endogenous variables", in particular binary outcome models with dummy endogenous variables. In such cases the 2SLS or equivalently linear IV estimator retains the quality of being easy to compute. Yet the assumptions

*University College London and Centre for Microdata Methods and Practice, Department of Economics, Gower Street, London WC1E 6BT, United Kingdom, andrew.chesher@ucl.ac.uk, adam.rosen@ucl.ac.uk. We thank Amit Gandhi for insightful comments and discussion. We acknowledge financial support through a grant (RES-589-28-0001) to the ESRC Research Centre, CeMMAP, through the funding of the "Programme Evaluation for Policy Analysis" node of the UK National Centre for Research Methods and from the European Research Council grant ERC-2009-StG-240910-ROMETA. A version of this paper is forthcoming in the May 2013 issue of the *American Economic Review: Papers and Proceedings*.

required for its justification are incompatible with the discrete nature of the endogenous variables. It is well-known for instance that linear IV estimators often lead to fitted probabilities that lie outside the unit interval for certain covariate values. Control function approaches that respect the binary nature of Y_1, e.g. those discussed by Blundell & Powell (2003) and Lewbel, Dong & Yang (2012), generally require Y_2 to be continuously distributed.

Nonetheless, 2SLS is often used in empirical work with discrete endogenous variables. Two main justifications are used in support of this. The first is that the 2SLS estimator is easy to compute, and the linear IV model should be thought of only as an approximation to the process generating outcomes. The second is that even with discrete endogenous variables, the 2SLS estimator is consistent for the local average treatment response (LATE) under the restrictions of Imbens & Angrist (1994).

The first of these arguments seems to us unsound, as computational ease does not justify reliance on a model which does not respect elementary properties of the data at hand. The second argument however does not hinge on the validity of the untenable assumption of the linear IV model. The assumptions under which the 2SLS estimator is consistent for the LATE allow for discrete endogenous variables and are readily interpretable. This argument, put forward in Angrist & Pischke (2009), is compelling if the LATE parameter is of interest.

There may be interest in other functionals of the distribution of treatment effects, such as average or quantile treatment effects. Unless treatment response is homogeneous, the LATE parameter, and thus the 2SLS estimator, will not in general coincide with these. A comprehensive analysis of the relation between various treatment effect parameters is provided by Heckman & Vytlacil (2005).

2 The Nonparametric Binary Response Model

We consider a nonparametric threshold-crossing model for binary Y_1,

$$Y_1 = 1\left[p\left(Y_2, W\right) < U\right], \tag{2.1}$$

and normalize the function p so that U is uniformly distributed on the unit interval, and impose the independence restriction $U \perp (W, Z)$. The model is a special case of the IV unordered choice models studied by Chesher, Rosen & Smolinski (forthcoming), hereafter CRS. It was shown in Chesher (2010) that IV models for discrete outcomes generally point identify neither the function h in (1.1) nor many interesting functionals of it. Yet the model can deliver non-trivial *set* identification of functionals of h, and the sets may be small enough to allow data to convey valuable information and the impact of a policy to be usefully bounded.

While we focus here on nonparametric modeling, CRS provides set identification results that accommodate further restrictions on the structural relation or the distribution of unobservables, for example parametric restrictions. Chesher (forthcoming) investigates the use of index and monotonicity restrictions in the context of binary outcomes.

Discrete outcomes are common in microeconometrics and it is good to have robust and flexible ways of handling endogeneity. Most attempts to date use complete models that place structure on the process determining Y_2. Leading examples are parametric models for which a likelihood function can be defined as used in Heckman (1978) and semi- and nonparametric models with a triangular recursive equation structure, see for example Blundell & Powell (2003), Chesher (2003), and Imbens & Newey (2009). In these models U and Y_2 vary independently conditional on the value of some control function depending on Y_2 and the covariates (W, Z).

These potentially point identifying complete models place strong restrictions on the process generating endogenous explanatory variables. They are typically nested in our nonparametric IV model, in which case they identify points in our IV model's identified set. If *only* the IV model's restrictions are held to be plausible, then the choice of one complete model over another is arbitrary.

The identified set delivered by IV models for discrete outcomes is characterized for a general class of multiple discrete choice models in CRS using powerful tools developed in random set theory, see Molchanov (2005). Here we set out the implications of these results for binary outcome threshold crossing models of

which leading parametric examples are probit and logit models, our contribution being the extension of these models to cases with endogenous explanatory variables and IV restrictions.

In this paper we focus on set identification of an average treatment effect (ATE) with a single endogenous variable. Set identification of the ATE under alternative nonparametric restrictions has been previously considered by e.g. Manski (1990), Balke & Pearl (1997), and Shaikh & Vytlacil (2011). The characterizations of CRS that we invoke are for identified sets of structural functions as well as the distribution of unobserved heterogeneity, and can be used to trace out the identified set for any functional of these quantities, not just the ATE. Moreover these characterizations easily accommodate further semiparametric and parametric restrictions. We illustrate the application of the results using data on female employment and family size employed in Angrist & Evans (1998).

3 Identification Analysis

Key in the analysis of CRS are level sets, $\mathcal{U}_h(y_1,y_2,w)$, of the function h, defined as follows.

$$\mathcal{U}_h(y_1,y_2,w) = \{u : y_1 = h(y_2,w,u)\} \tag{3.1}$$

In continuous outcome models with scalar latent U these sets are usually singleton residuals, but in discrete outcome models they are typically non-singleton sets. Any particular function h along with any observed probability distribution for $Y = (Y_1, Y_2)$ given $(W, Z) = (w, z)$, denoted $F_{Y|WZ}$ delivers a set-valued random variable, $\mathcal{U}_h(Y_1,Y_2,w)$, whose realizations are the level sets defined in (3.1). Let $\mathbb{P}$ indicate probabilities calculated using a distribution $F_{Y|WZ}$. Let $\mathcal{R}_W$ denote the support of W and $\mathcal{R}_Z(w)$ the support of Z when $W = w$. The objects of interest are the structural function h and the conditional distribution of U given $(W, Z) = (w, z)$, denoted $G_{U|WZ}$.

CRS show the identified set of functions $(h, G_{U|WZ})$ supported by a model $\mathfrak{M}$ and a probability distribution $F_{Y|WZ}$ comprises all pairs $(h, G_{U|WZ})$ that: (a) are admitted by the model $\mathfrak{M}$ and (b) satisfy the following inequalities

$$\forall \mathcal{S} \in \mathsf{C}(w,z),\ G_{U|WZ}(\mathcal{S}|w,z) \geq \mathbb{P}[\mathcal{U}_h(Y_1,Y_2,w) \subseteq \mathcal{S}|w,z], \tag{3.2}$$

for almost every (w, z) in the support of (W, Z). Here $\mathsf{C}(w, z)$ is the collection of all unions of the sets on the support of the random set $\mathcal{U}_h(Y_1,Y_2,w)$ given $Z = z$, and $G_{U|WZ}(\mathcal{S}|w, z)$ is the probability mass placed on the set $\mathcal{S}$ by the distribution $G_{U|WZ}$ when $(W, Z) = (w, z)$. Essential elements of the model $\mathfrak{M}$ are restrictions limiting the dependence of U on (W, Z).

In the context of the binary threshold-crossing model (2.1) the only relevant elements of the collection of sets $\mathsf{C}(w, z)$ are the pairs of intervals, $[0, p(y_2, w)]$ and $(p(y_2, w), 1]$, generated as y_2 takes all values on the support of Y_2. The distribution $G_{U|WZ}$ is uniform and does not depend on (W, Z), so $G_{U|WZ}(\mathcal{S}|w, z)$ is just the length of the interval, either $p(y_2, w)$ or $1 - p(y_2, w)$.

Define the set $\mathcal{A}_p(y_2, w)$:

$$\mathcal{A}_p(y_2,w) \equiv \{y_2^* : p(y_2^*,w) \leq p(y_2,w)\},$$

let $\mathcal{A}_p(y_2, w)^c$ denote its complement and consider a model $\mathfrak{M}$ which may place parametric restrictions on the threshold functions p, for example specifying a probit functional form and an index restriction. Applying the result just set out we find that the identified set of threshold crossing functions comprises the functions p admitted by the model $\mathfrak{M}$ and satisfying, for almost every (y_2, w), the following inequalities.

$$\sup_{z\in\mathcal{R}_Z(w)} \left(\mathbb{P}\left[Y_1 = 0 \wedge Y_2 \in \mathcal{A}_p(y_2,w)|w,z\right]\right) \leq p(y_2,w) \leq \inf_{z\in\mathcal{R}_Z(w)} \left(1 - \mathbb{P}\left[Y_1 = 1 \wedge Y_2 \in \mathcal{A}_p(y_2,w)^c|w,z\right]\right). \tag{3.3}$$

4 Family Size and Employment

We apply these results in an analysis of the identifying power of a model for, and data on, female employment used in Angrist & Evans (1998). The element of this study on which we focus uses data from the US Census Public Use Microsamples on $254,654$ married mothers aged 21-35 in 1980 with at least 2 children and oldest child less than 18. The dependent variable in the model, Y_1, is binary, equal to 1 if a woman worked for pay in 1979. The potentially endogenous variable, Y_2, is binary, equal to 1 for women having three or more children. The instrumental variables are also binary, one taking the value 1 if the second birth of a woman was twins, the other taking the value 1 if the first two children have the same sex. We consider models in which the only explanatory variable is Y_2, so $p(y_2, w)$ becomes $p(y_2)$.

Define parameters as follows.

- $\rho_0 \equiv 1 - p(0)$: the mean potential Y_1 outcome when Y_2 is zero. In this binary outcome case this is a counterfactual probability.
- $\rho_1 \equiv p(0) - p(1)$: the difference in mean potential Y_1 outcomes comparing $Y_2 = 1$ and $Y_2 = 0$. This is the Average Treatment Effect (ATE) in the IV model.

One of the approaches advocated in Angrist & Pischke (2009), hereafter AP09, who use this data, employs a linear probability model

$$Y_1 = \alpha_{lpm} + \beta_{lpm} Y_2 + \varepsilon,$$

estimating the parameters by 2SLS. Under the normalization $E[\varepsilon] = 0$ there is $\rho_0 = \alpha_{lpm}$ and the ATE is $\rho_1 = \beta_{lpm}$.

The 2SLS and ordinary least squares (OLS) estimators, under mild conditions, converge to the following probability limits.

$$\beta_{2sls} = \frac{cov(Z, Y_1)}{cov(Z, Y_2)},\ \alpha_{2sls} = E[Y_1] - \beta_{2sls} E[Y_2],$$

$$\beta_{ols} = \frac{cov(Y_1, Y_2)}{var(Y_2)},\ \alpha_{ols} = E[Y_1] - \beta_{ols} E[Y_2],$$

In general these depend on the probability distribution of the instrumental variable, in the latter case through dependence of Y_2 on Z. The value of β_{2sls} is the LATE introduced in Imbens & Angrist (1994).

In Figures 2-9 of Appendix A we exhibit identified sets for (ρ_0, ρ_1) delivered by the IV model when probabilities are generated by a triangular model like that used in Heckman (1978) with

$$\begin{aligned} Y_1 &= 1\left[-\alpha_\Delta - \beta_\Delta Y_2 < U^*\right] \\ Y_2 &= 1\left[-\gamma - \delta Z < V^*\right] \end{aligned}$$

and correlated zero-mean, unit-variance Gaussian errors (U^*, V^*). In this structure the value of the ATE is $\Phi(\beta_\Delta + \alpha_\Delta) - \Phi(\alpha_\Delta)$. We show the locations of the probability limits of 2SLS and OLS estimators under different distributions of the instrumental variable. The results show that there can be substantial instrument sensitivity. The LATE, that is β_{2sls}, tends to lie close to the value of the ATE delivered by the Gaussian unobservable triangular structure, but can take values outside the identified set delivered by the nonparametric instrumental variable model.

Linear probability model estimates of β_{lpm} using the aforementioned 1980 census data from Angrist & Evans (1998) are as follows.

IV: Same-sex	IV: Twins	OLS
-0.138	-0.084	-0.115
(.029)	(.017)	(.002)

The quarter of a million observations deployed in the estimation deliver small standard errors (in parentheses).

Figure 1 shows estimated sharp identified sets delivered by the IV model and the 2SLS estimates obtained using each instrument one-at-a-time and the OLS estimate. In this model with a binary instrument Z and

no exogenous covariates W the identified set represented in (3.3) is simply those pairs of (ρ_0, ρ_1) such that $\rho_1 \leq 0$ and for each $z \in \{0, 1\}$

$$\begin{aligned} f_{10}(z) + f_{11}(z) &\leq \rho_0 \leq 1 - f_{00}(z), \\ f_{11}(z) &\leq \rho_0 + \rho_1 \leq f_{10}(z) + f_{11}(z), \end{aligned}$$

or, $\rho_1 \geq 0$ and for each $z \in \{0, 1\}$

$$\begin{aligned} f_{10}(z) &\leq \rho_0 \leq f_{10}(z) + f_{11}(z), \\ f_{10}(z) + f_{11}(z) &\leq \rho_0 + \rho_1 \leq 1 - f_{01}(z), \end{aligned}$$

where $f_{ij}(z) \equiv \mathbb{P}[Y_1 = i \wedge Y_2 = j | z]$.

The same-sex instrument delivers an identified set comprised of two parallelograms (drawn in blue) one on each side of the $\rho_1 = 0$ axis. The 2SLS estimate lies in the lower parallelogram, close to the OLS estimate. The identified set is very large and *disconnected*. The simple IV model using the same-sex instrument and these data has little to say about the effect (ATE) of a third child on female labour force participation. It is not possible to sign the effect on the counterfactual employment probability of advancing to a third child. This happens because the IV model is only partially identifying and the same-sex instrument is not an accurate predictor of advancing to a third child - the probability of having three or more children conditional on the first two children having the same (different) sex is 0.41 (0.35).

With the twins instrument the situation is very different. The identified set is a one-dimensional line, drawn in red in Figure 1. The twins instrument is in one sense a very precise predictor of advancement to a third child because if the second birth is a twin birth then necessarily the mother has three children. The probability of having three or more children conditional on the second birth being (not being) a twin birth is 1.00 (0.38). It can be shown that in this situation the IV model point identifies neither ρ_0 nor ρ_1 individually but it does point identify their sum, so the identified set (the red line in Figure 1) lies on a line with slope -1 and intercept $\rho_0 + \rho_1$. Using the twins instrument the IV model identifies the sign of ρ_1 but the estimate of the interval in which it lies is quite large, $[-.25, -.05]$. Employing both instruments only marginally tightens this bound given the weakness of the same-sex instrument.

In this binary endogenous variable case, associated with each point in an identified set for (ρ_0, ρ_1) there is a unique pair of conditional distributions of U given $Y_2 = y_2 \in \{0, 1\}$ that generates the probability distributions of (Y_1, Y_2) given Z that deliver the set. Examples for a variety of values of (ρ_0, ρ_1) in the estimated same-sex identified set are shown in Figures 10-18 in Appendix B. The results show that at the 2SLS estimate the endogeneity of Y_2 is very weak but extreme points in the identified set are associated with widely varying degrees and directions of endogeneity.

5 Concluding Remarks

We have applied results from the recent literature on partial identification to the study of treatment effect parameters in binary response models with binary endogenous variables, or treatments. We compared a fully nonparametric, set identifying analysis to that delivered by the commonly used 2SLS estimator. The 2SLS estimator is simple to use but limited in application, delivering a point estimator of the LATE parameter but not of other treatment effect parameters unless there is treatment effect homogeneity. We used a set identifying nonparametric IV framework to characterize the sharp identified set for an average treatment effect, and compared it to the 2SLS and OLS estimands, both numerically (in the Appendix) and in an application using data on female labor supply from Angrist & Evans (1998).

The results indicate that practitioners should use the 2SLS estimator with caution. If researchers are interested in measuring a LATE, they can safely stop with 2SLS. If other parameters are of interest, 2SLS, equivalently linear IV, does not suffice. Our analysis should caution applied researchers who might wish to interpret 2SLS estimators as approximations for causal effects or treatment effect parameters other than LATE, by showing that at least one such parameter, namely the ATE, can in fact be quite far from the

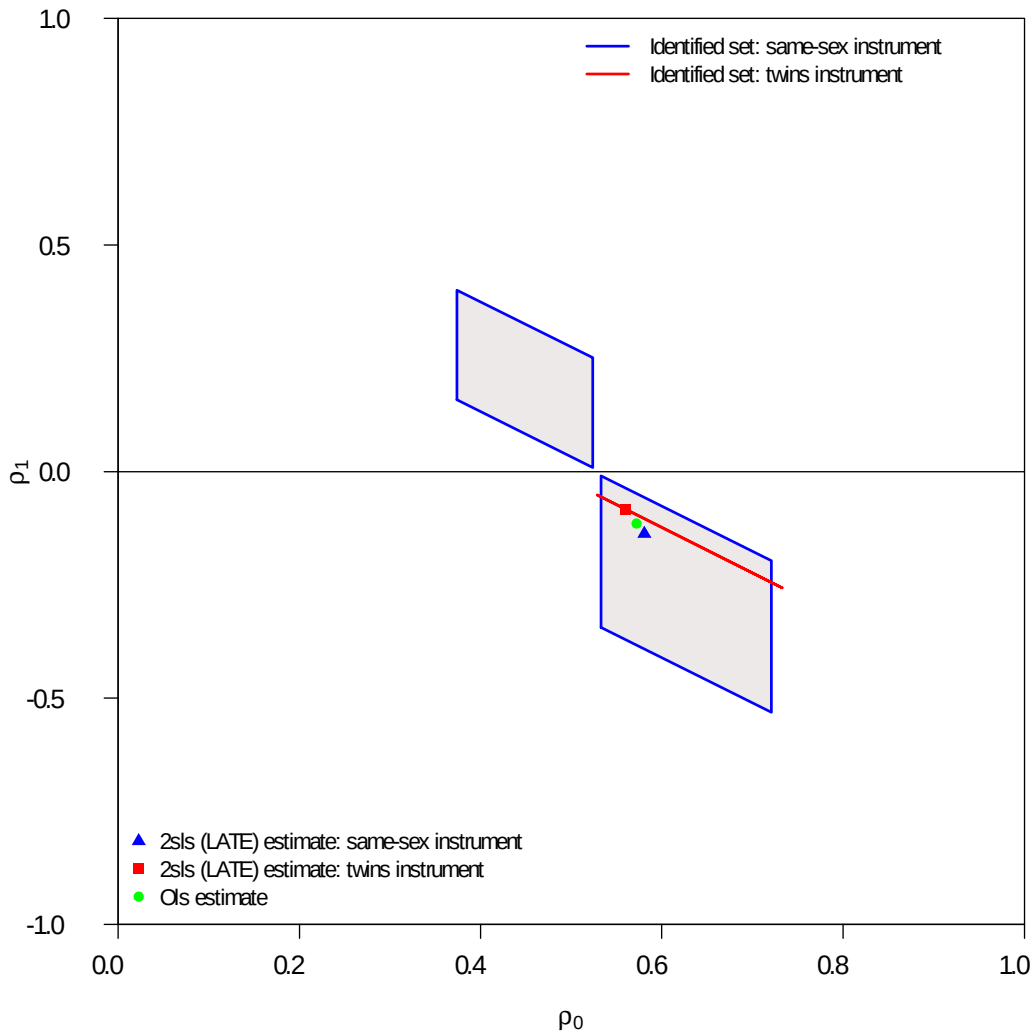

Figure 1: Identified sets for (ρ_0, ρ_1) and 2SLS and OLS estimates.

LATE. If one is interested in the ATE then bounds analysis may be preferable. Nonetheless, if one wishes to use 2SLS then bounds analysis provides a valuable sensitivity analysis.

Here we have focused on a nonparametric model with a simple binary instrument. Even in this case, the bounds can be useful, although they may be wider than hoped in some circumstances. If this is all that data and plausible assumptions deliver then this is important to learn. On the other hand, many applications feature instruments with much richer support. Further, there are more restrictive models than the fully nonparametric one used here that likewise respect the discrete nature of endogenous variables, and which in general deliver narrower bounds. Our results on the set identifying power of the incomplete IV model apply to a much wider range of problems than studied here.

There are both a variety of useful weakly restrictive models shown to deliver bounds on interesting quantities, and a growing number of inference methods available, for example Chernozhukov, Hong & Tamer (2007), Andrews & Shi (forthcoming), Chernozhukov, Lee & Rosen (forthcoming) and references therein. These tools expand the range of models, and parameters of interest, which can be studied in applications by practitioners, whether or not there is point identification.

References

Andrews, Donald W. K., and Xiaoxiao Shi. forthcoming. "Inference for Parameters Defined by Conditional Moment Inequalities." *Econometrica.*

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics.* Oxford, UK:Princeton University Press.

Angrist, Joshua, and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review*, 88(3): 450–487.

Balke, Alexander, and Judea Pearl. 1997. "Bounds on Treatment Effects From Studies With Imperfect Compliance." *Journal of the American Statistical Association*, 92(439): 1171–1176.

Blundell, Richard, and James L. Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." In *Advances in Economics and Econometrics.* , ed. Mathias Dewatripont, Lars Peter Hansen and Stephen J. Turnovsky, 312–357. Cambridge University Press.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica*, 75(5): 1243–1284.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen. forthcoming. "Intersection Bounds, Estimation and Inference." *Econometrica.*

Chesher, Andrew. 2003. "Identification in Nonseparable Models." *Econometrica*, 71(5): 1405–1441.

Chesher, Andrew. 2010. "Instrumental Variable Models for Discrete Outcomes." *Econometrica*, 78(2): 575–601.

Chesher, Andrew. forthcoming. "Semiparametric Structural Models of Binary Response: Shape Restrictions and Partial Identification." *Econometric Theory.*

Chesher, Andrew, Adam Rosen, and Konrad Smolinski. forthcoming. "An Instrumental Variable Model of Multiple Discrete Choice." *Quantitative Economics.*

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

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Imbens, Guido, and Whitney Newey. 2009. "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." *Econometrica*, 77(5): 1481–1512.

Lewbel, Arthur, Yingying Dong, and Thomas Tao Yang. 2012. "Comparing Features of Convenient Estimators for Binary Choice Models with Endogenous Regressors." working paper, Boston College.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review Papers and Proceedings*, 80(2): 319–323.

Molchanov, Ilya S. 2005. *Theory of Random Sets.* London:Springer Verlag.

Shaikh, Azeem, and Edward Vytlacil. 2011. "Partial Identification of Triangular Systems of Equations with Binary Dependent Variables." *Econometrica*, 79(3): 949–955.

Appendices:

These Appendices contain:

1. examples of identified sets delivered by an instrumental variable threshold crossing model for a binary outcome,
2. displays of conditional probability distributions of structural equation errors given endogenous explanatory variables.

A Examples of identified sets delivered by the IV model

We illustrate the identified sets delivered by the IV threshold crossing model using probability distributions created by us, generated using a Gaussian triangular structure, one of the cases studied in Heckman (1978), as follows.

$$Y_1 = 1[-\alpha_\Delta - \beta_\Delta Y_2 < U^*] \qquad Y_2 = 1[-\gamma - \delta Z < V^*]$$

$$\begin{bmatrix} U^* \\ V^* \end{bmatrix} | Z \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \right) \qquad Z \in \mathcal{R}_Z \equiv \{-1, 0, 1\}$$

Here $\rho_0 = \Phi(a)$ and the average treatment effect is $\rho_1 = \Phi(b + a) - \Phi(a)$ where Φ is the standard normal distribution function.

Figures 2-9 show identified sets (grey) for (ρ_0, ρ_1) at a variety of parameter settings. Throughout $\alpha_\Delta = \gamma = 0$ and $\beta_\Delta - 1$.

The value of (ρ_0, ρ_1) in the triangular Gaussian structure is marked in green. The points marked red and blue are discussed shortly. The IV model is partially identifying so other structures could have generated the probability distributions that delivered these sets. Example include non-triangular structures and non-Gaussian structures and in those cases the value of (ρ_0, ρ_1) would be at different points in the identified set.

In this simple case the sets comprise either one, or two disconnected, parallelograms. When the endogenous variable (here binary) has K points of support identified sets are unions of up to $K!$ convex polytopes which may be disconnected, see Chesher (forthcoming) for an illustration. In Figures 2-6 the identified sets are connected, the strength of the instrument steadily increases (the coefficient δ is increasing) and as it does the identified set shrinks. There would be point identification if δ increased without limit. In Figures 7-9 the correlation r is positive and the identified set is disconnected until the instrument becomes quite strong.

The points marked red and blue in Figures 2-9 show pseudo-true values

$$\beta_{2sls} = \frac{cov(Z, Y_1)}{cov(Z, Y_2)} \qquad \alpha_{2sls} = E[Y_1] - \beta_{2sls} E[Y_2]$$

$$\beta_{ols} = \frac{cov(Y_1, Y_2)}{var(Y_2)} \qquad \alpha_{ols} = E[Y_1] - \beta_{ols} E[Y_2]$$

of 2SLS (red) and OLS (blue) estimators obtained using a linear probability model and 500 randomly chosen distributions of the instrumental variable.[1]

The pseudo-true value of the OLS estimator is quite sensitive to the distribution of the instrumental variable. The pseudo-true value of the 2SLS estimator (LATE) is less variable and is often close to the triangular model value but sometimes it falls outside the identified set. When the set is large it is of course far from many of the observationally equivalent values of (ρ_0, ρ_1).

[1] The instrumental variable has support on $\{-1, 0, 1\}$. To generate a random distribution on this support, three independent uniform pseudo-random variates (B_1, B_2, B_3) were generated and probabilities generated as $B_i / \Sigma_j B_j$ for $i \in \{1, 2, 3\}$.

B Distribution of latent U conditional on endogenous explanatory variable Y_2

In this binary endogenous variable case there is associated with each point in the identified set for (ρ_0, ρ_1) a unique pair of conditional distributions of U given $Y_2 = y_2 \in \{0, 1\}$ that generates the probability distributions of (Y_1, Y_2) given Z that deliver the identified set. Figures 10-18 show some of these distributions for the Angrist & Evans (1998) analysis (described in the main text) using the same-sex instrument. Figure 10 shows the distributions at the AP09 2SLS estimate. There is hardly any endogeneity here. The remaining Figures show the distributions (right hand panes) associated with some extreme points in identified set (the points are highlighted in the left hand panes). Different points in the identified set are associated with widely varying degrees and directions of endogeneity.

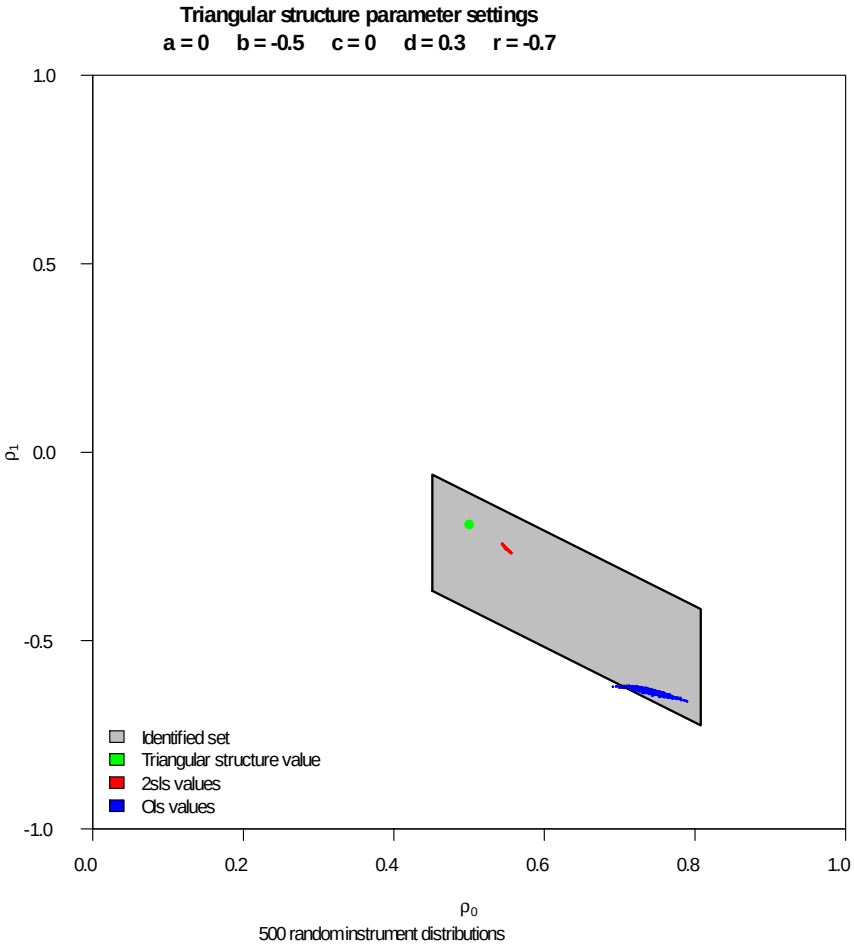

Figure 2: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

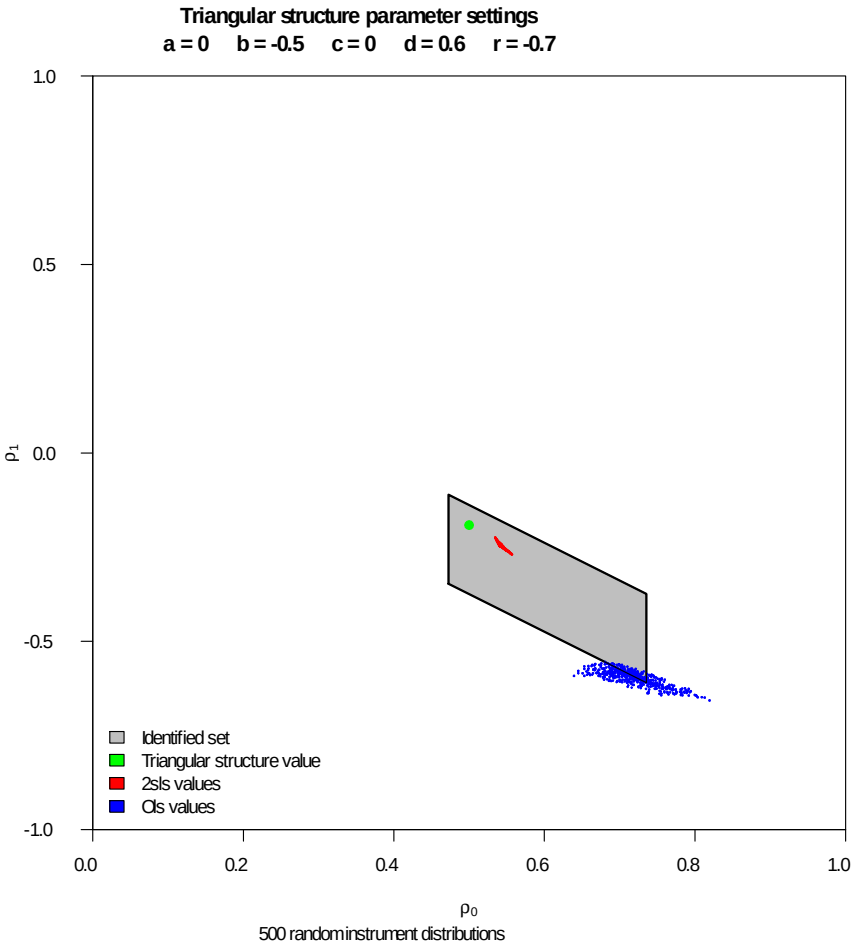

Figure 3: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

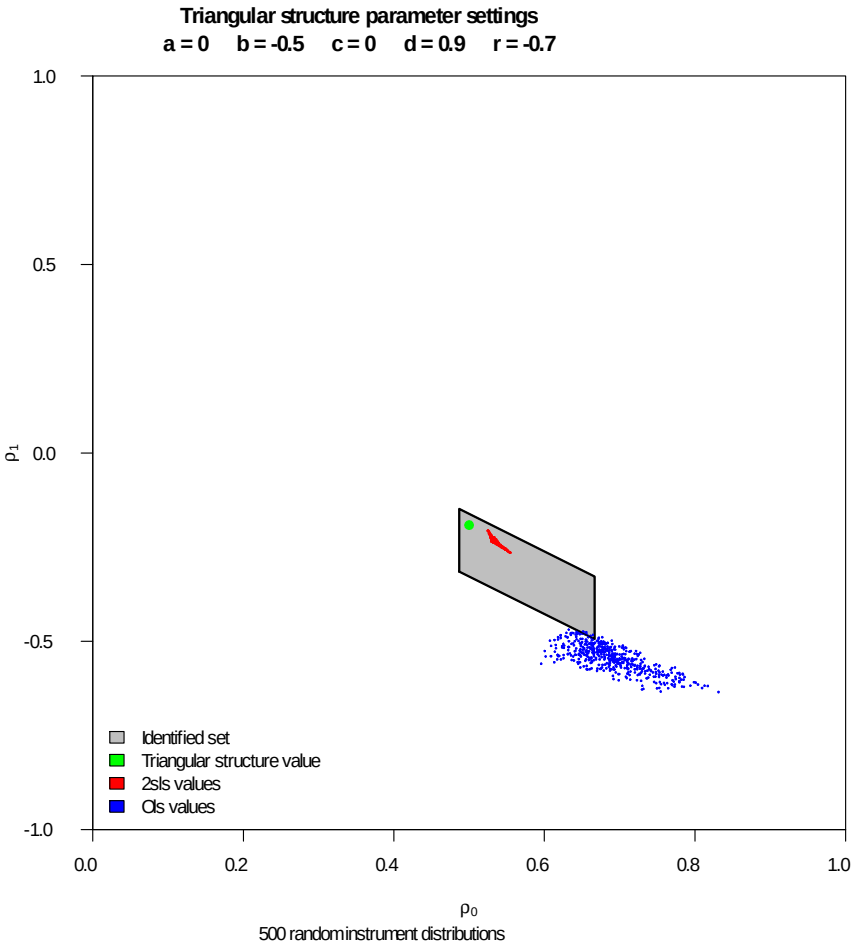

Figure 4: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

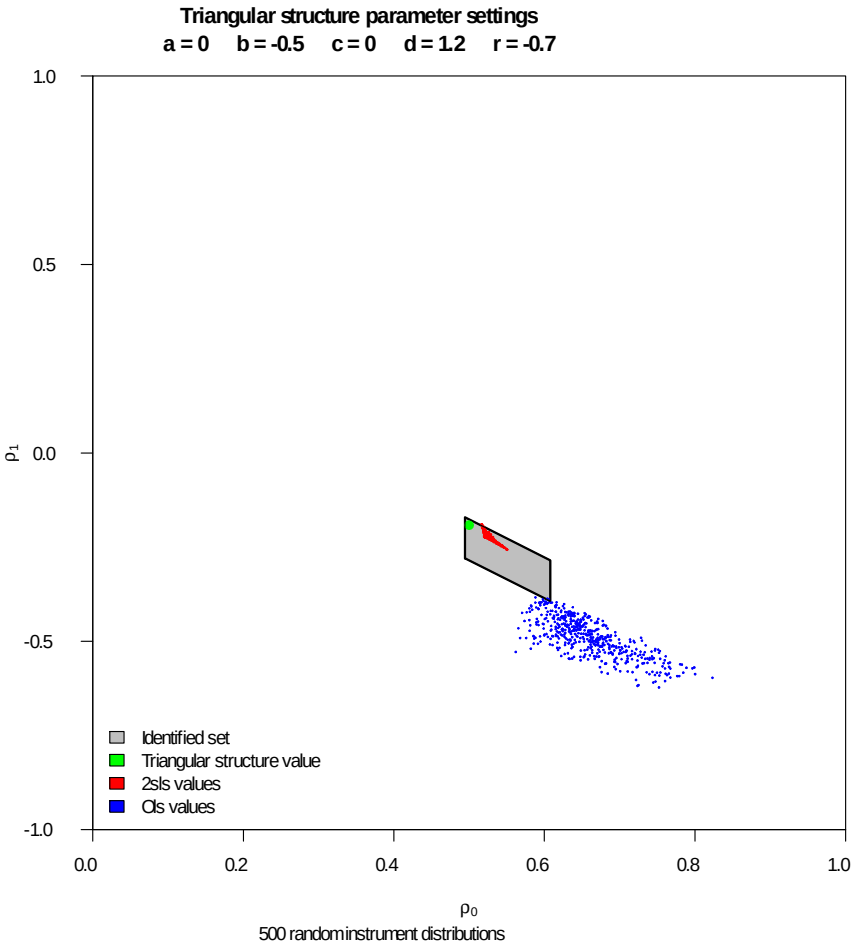

Figure 5: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

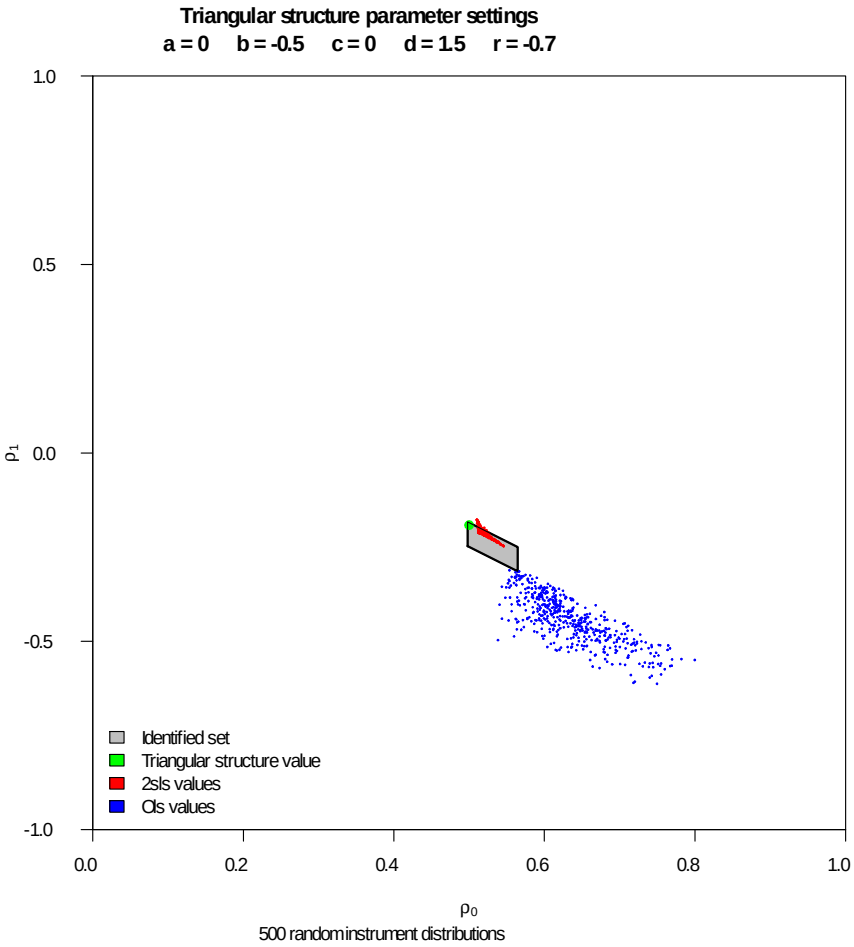

Figure 6: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

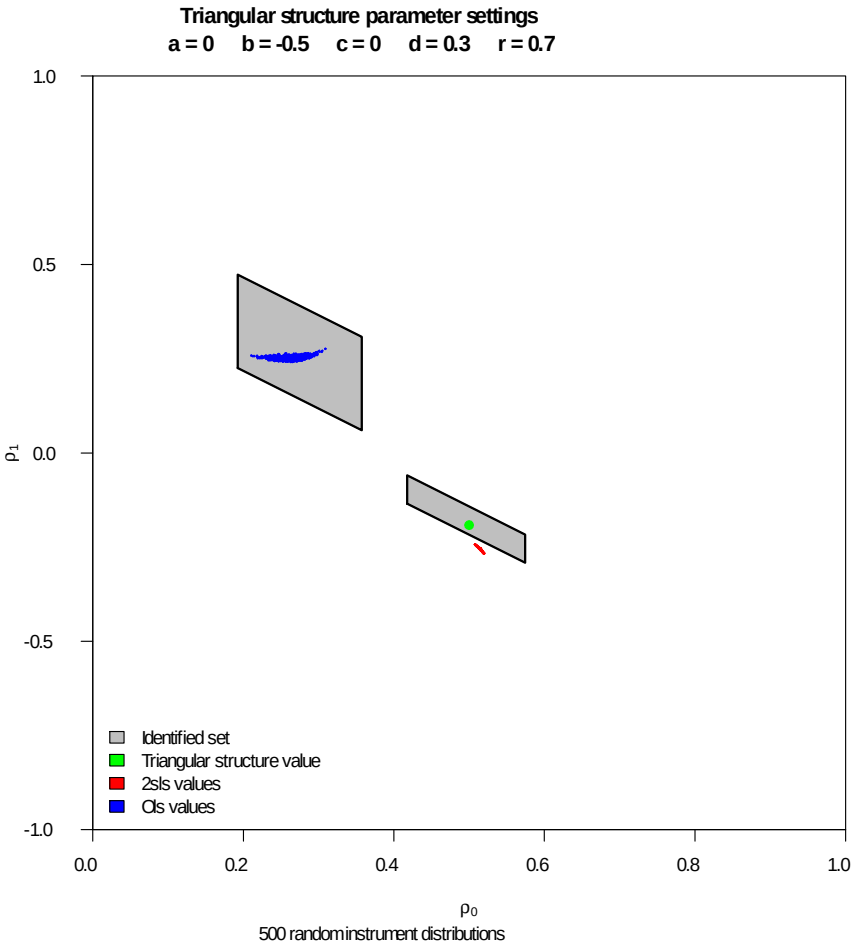

Figure 7: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

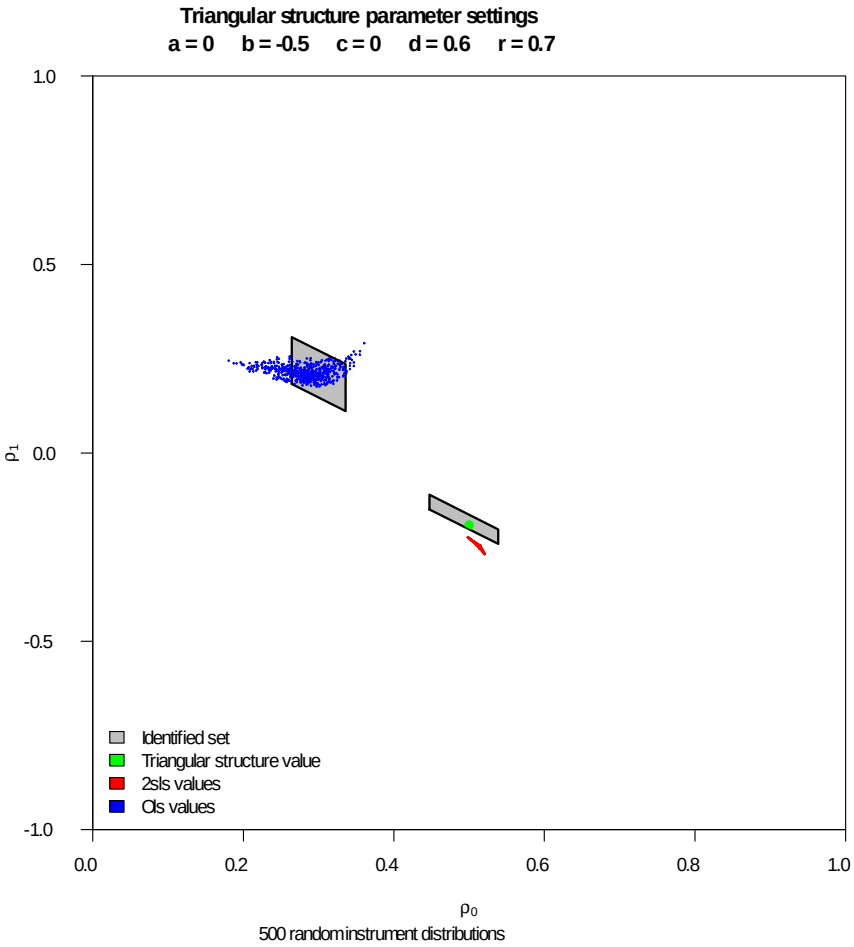

Figure 8: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

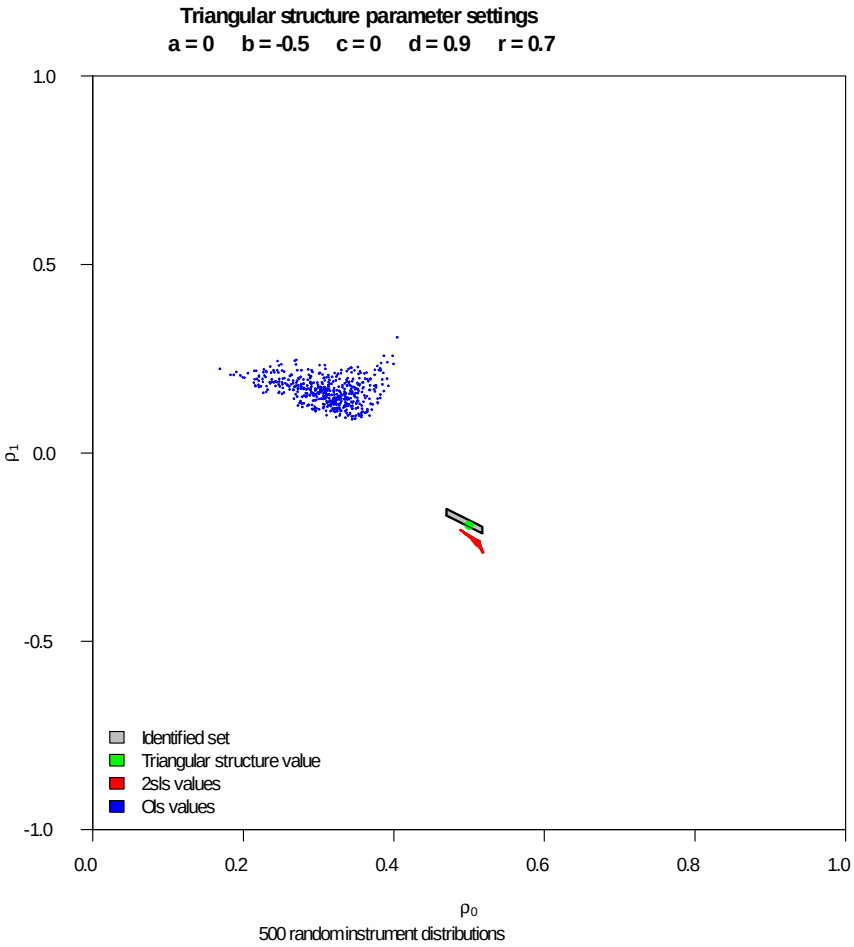

Figure 9: Identified set for ρ_0 and ρ_1 using probabilities generated by a triangular structure.

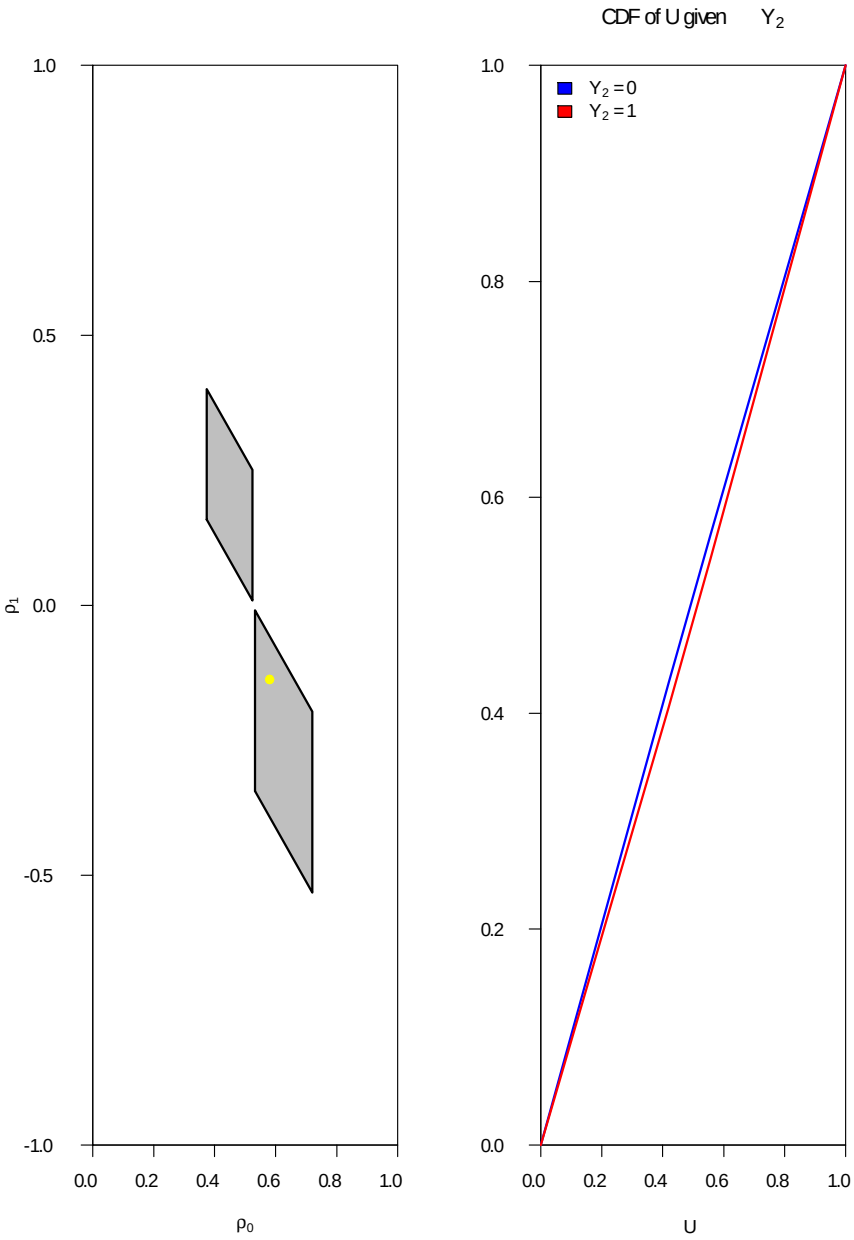

Figure 10: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane. This Figure shows the 2SLS estimator.

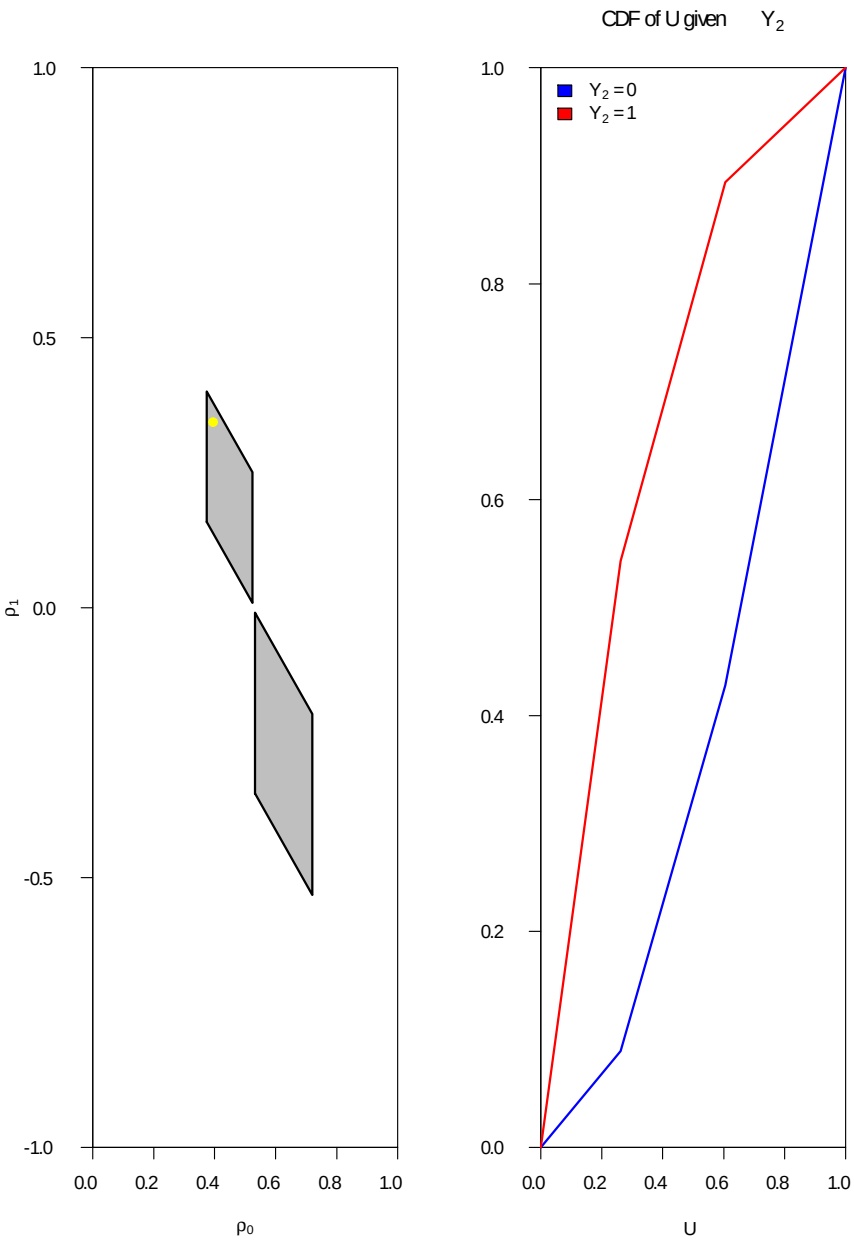

Figure 11: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

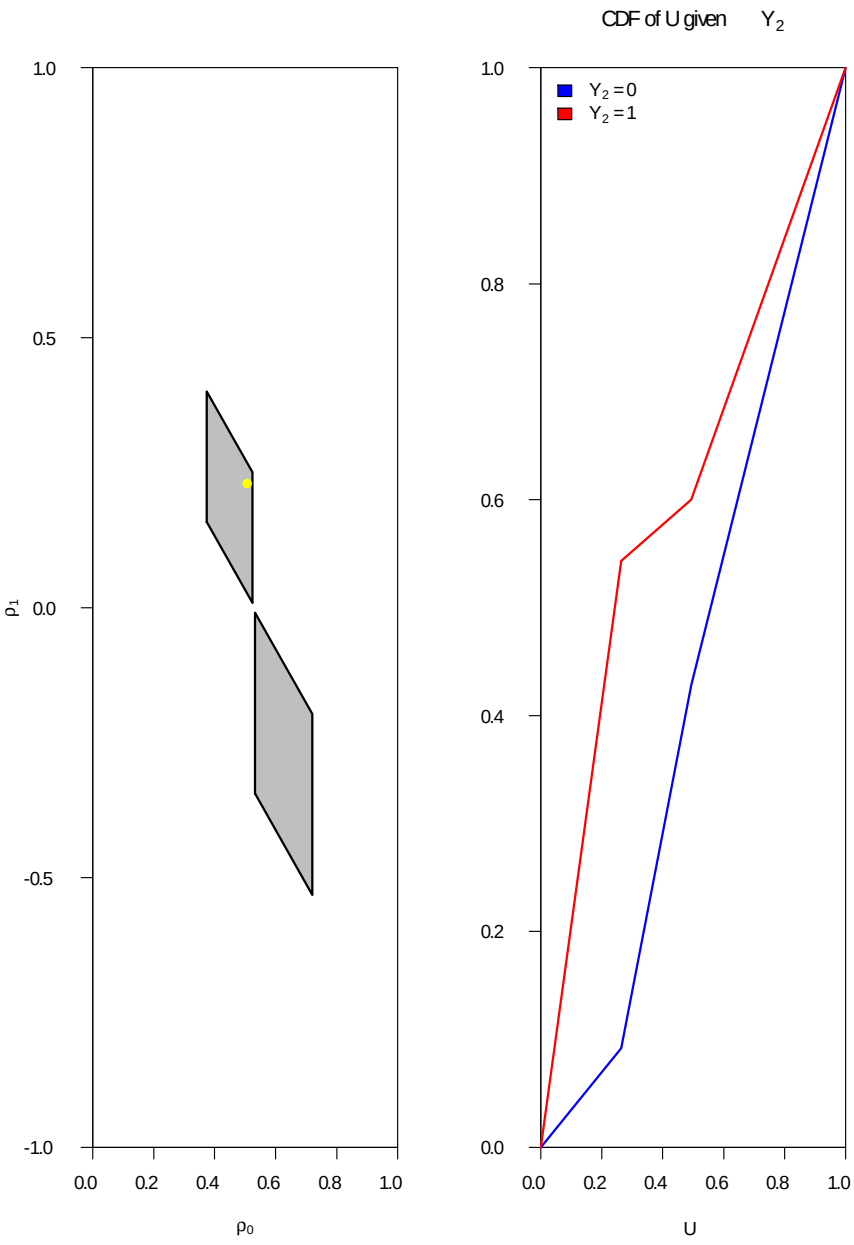

Figure 12: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

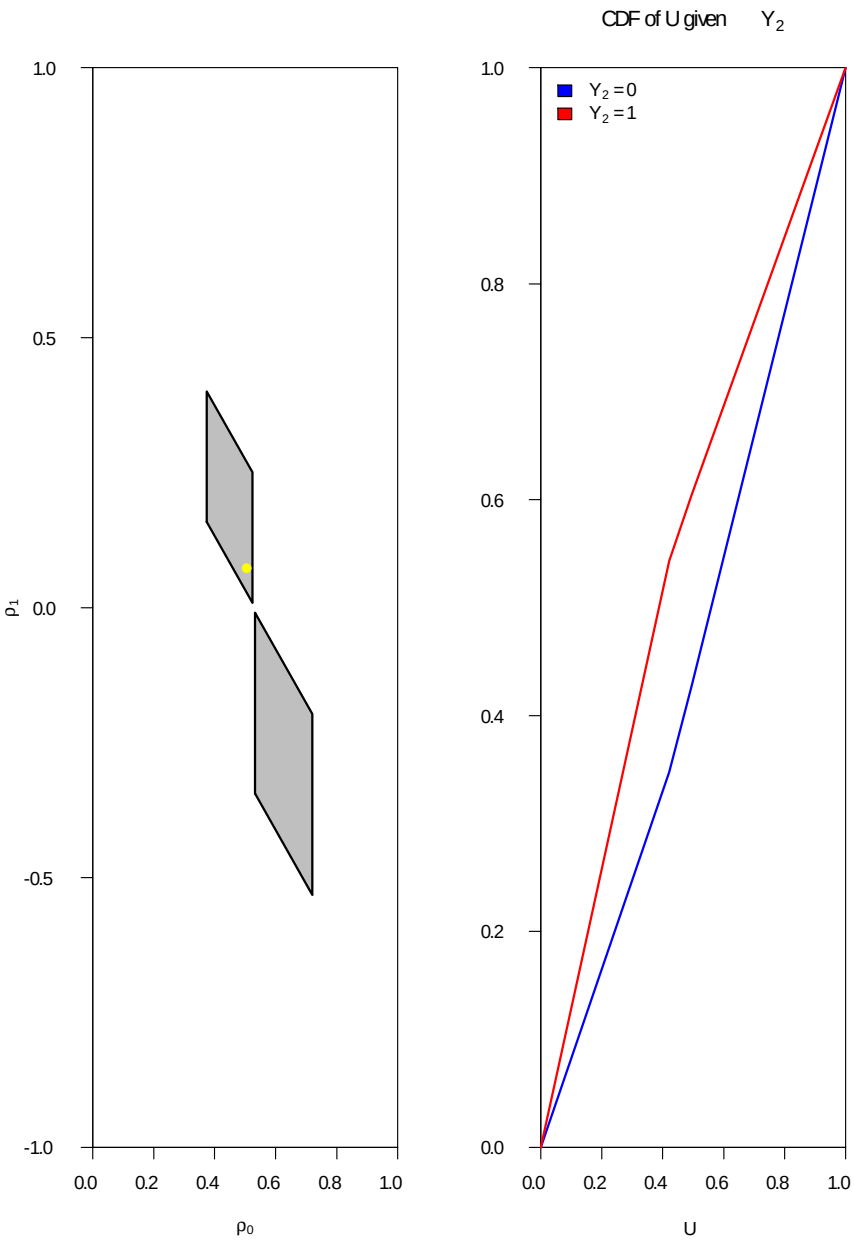

Figure 13: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

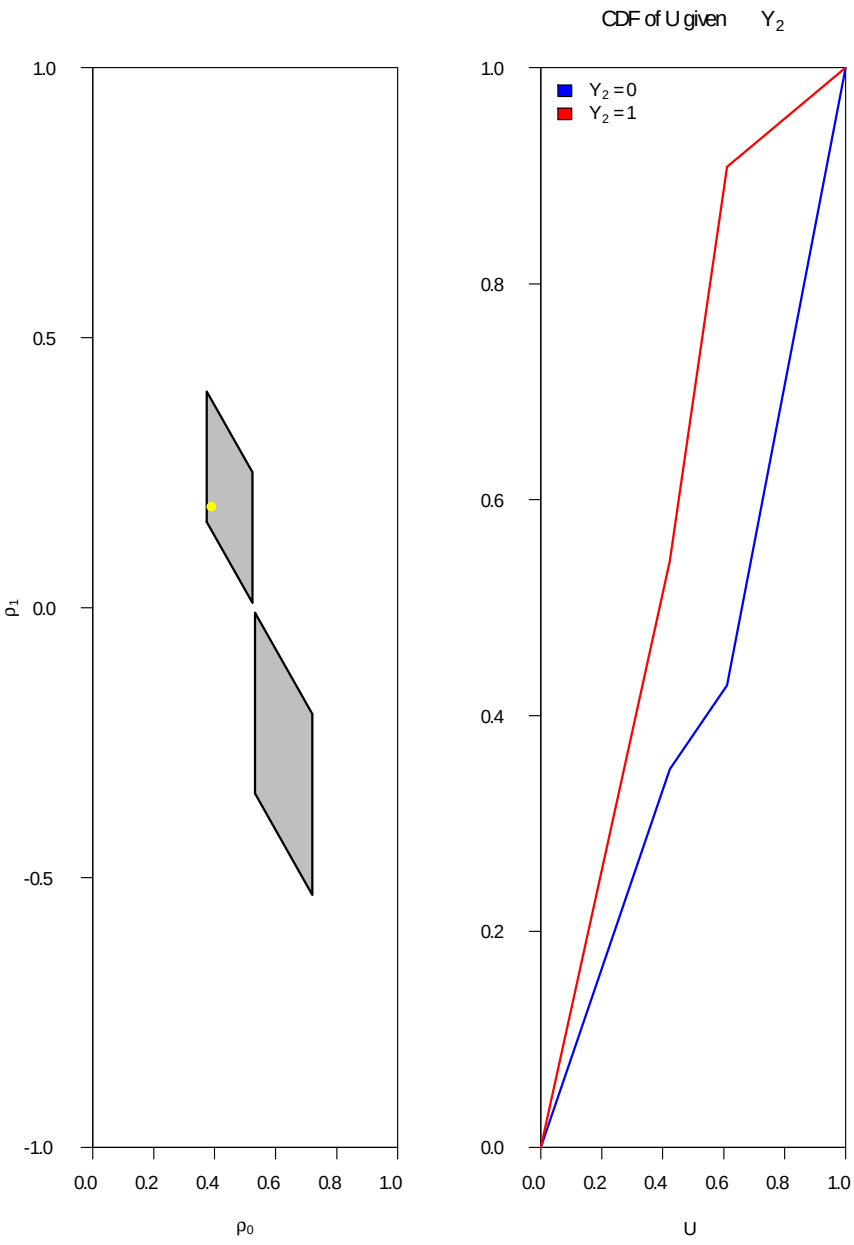

Figure 14: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

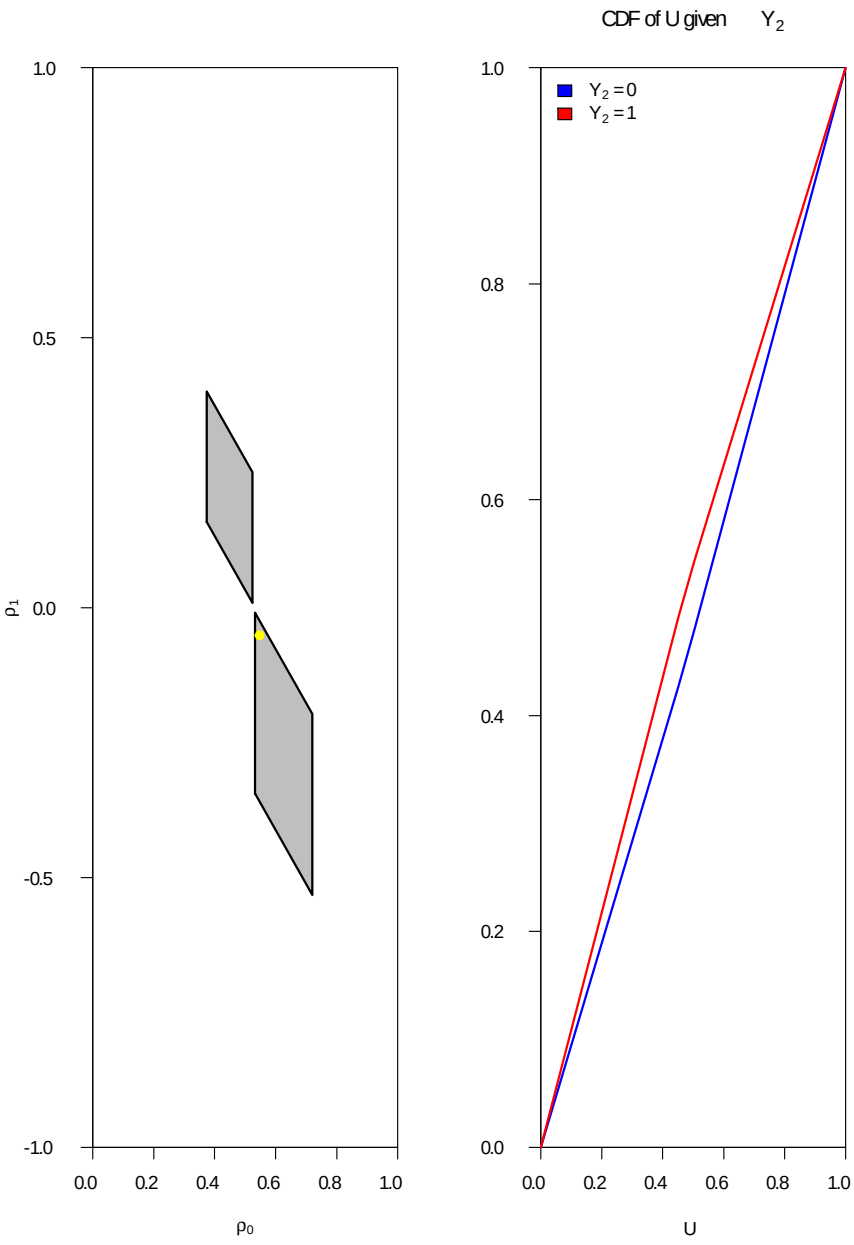

Figure 15: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

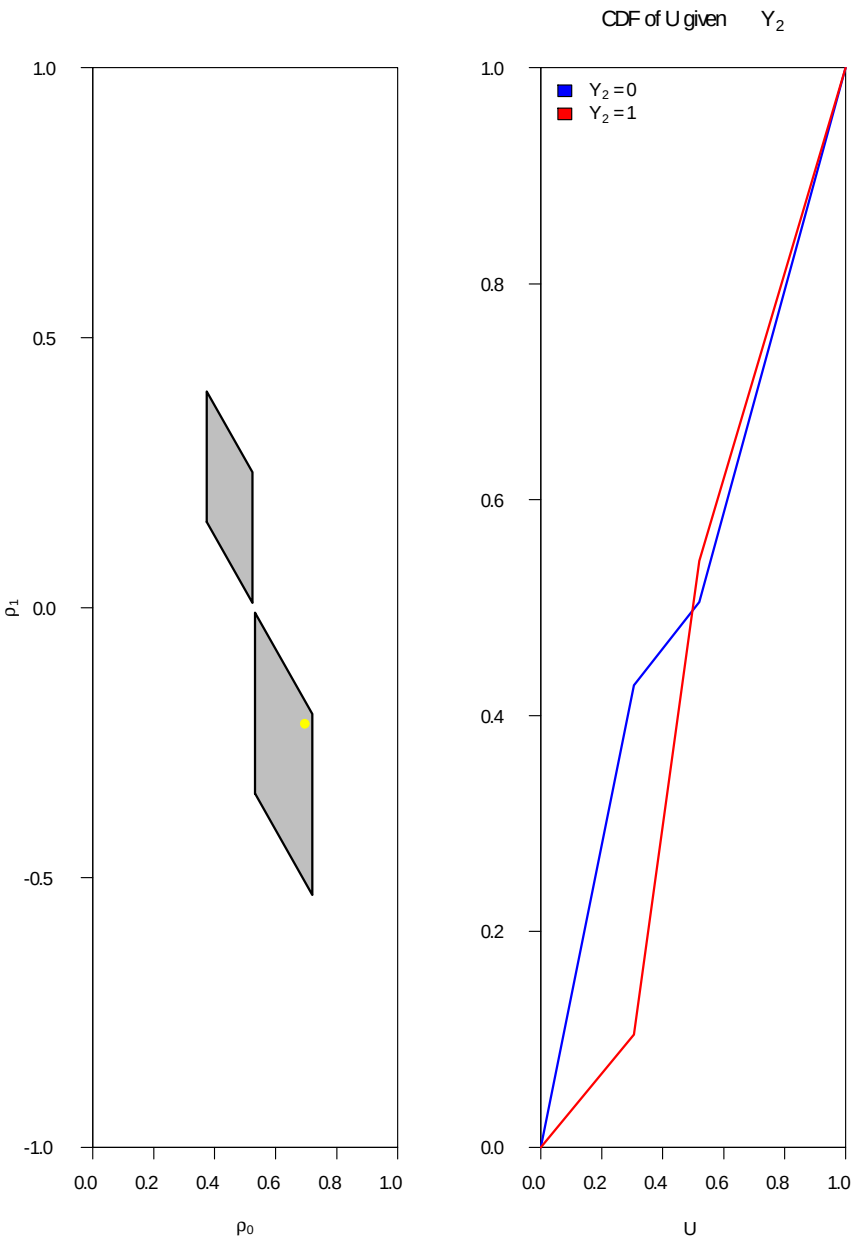

Figure 16: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

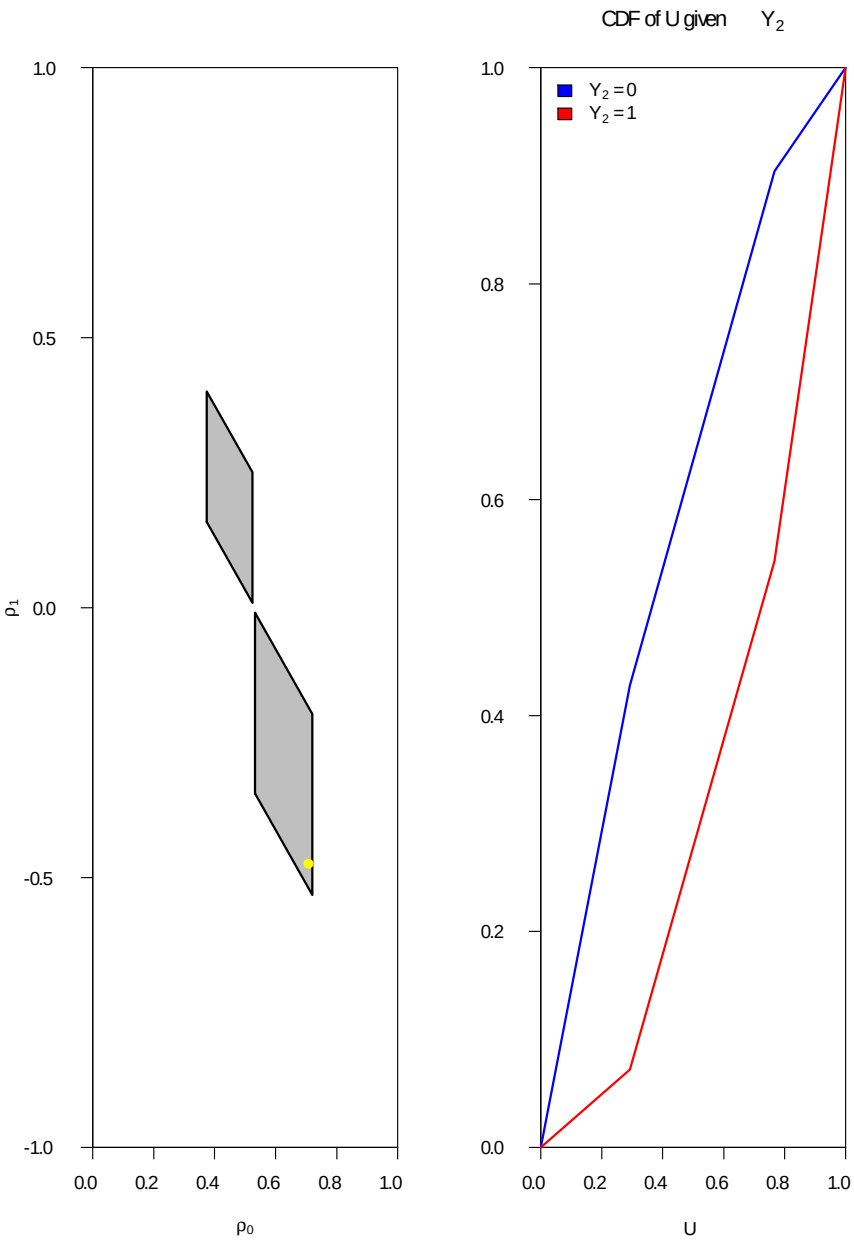

Figure 17: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.

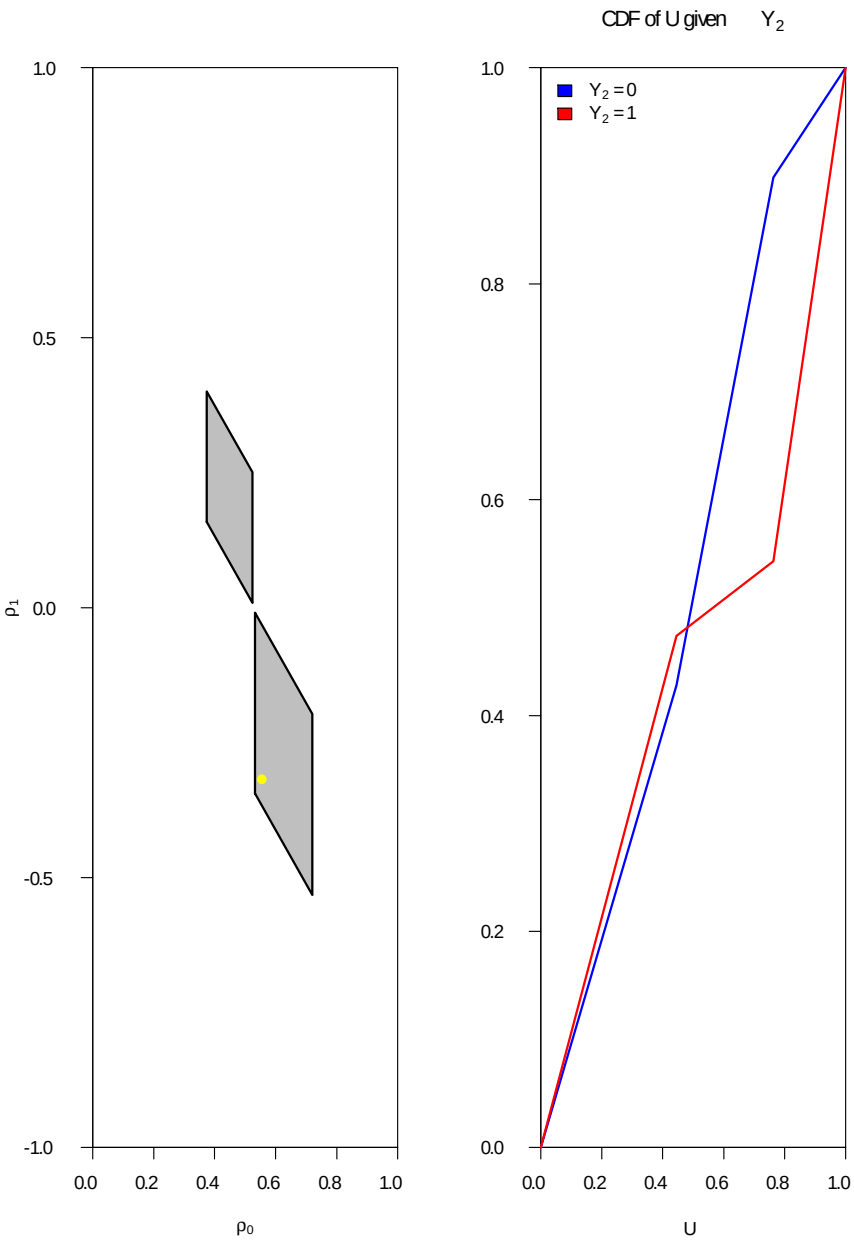

Figure 18: Right hand pane shows conditional distribution functions of U given Y_2 associated with the highlighted point in the estimated identified set shown in the left hand pane.