IVTREATREG: A NEW STATA ROUTINE FOR ESTIMATING BINARY TREATMENT MODELS WITH HETEROGENEOUS RESPONSE TO TREATMENT UNDER OBSERVABLE AND UNOBSERVABLE SELECTION

Cerulli Giovanni
Ivtreatreg: a new STATA routine for estimating binary treatment models with heterogeneous response to treatment under observable and unobservable selection

Giovanni Cerulli
National Research Council of Italy
Ceris-CNR
Institute for Economic Research on Firms and Growth
Via dei Taurini, 19 - 00185 Roma
Phone: +39.06.4993.7885
E-mail: g.cerulli@ceris.cnr.it

ABSTRACT: This paper presents a new user-written STATA command called ivtreatreg for the estimation of five different (binary) treatment models with and without idiosyncratic (or heterogeneous) average treatment effect. Depending on the model specified by the user, ivtreatreg provides consistent estimation of average treatment effects both under the hypothesis of “selection on observables” and “selection on unobservables” by using Ordinary Least Squares (OLS) regression in the first case, and Instrumental-Variables (IV) and Selection-model (à la Heckman) in the second one. Conditional on a pre-specified subset of exogenous variables – thought of as driving the heterogeneous response to treatment – ivtreatreg calculates for each model the Average Treatment Effect (ATE), the Average Treatment Effect on Treated (ATET) and the Average Treatment Effect on Non-Treated (ATENT), as well as the estimates of these parameters conditional on the observable factors, i.e., ATE(x), ATET(x) and ATENT(x). The five models estimated by ivtreatreg are: Cf-ols (Control-function regression estimated by OLS), Direct-2sls (IV regression estimated by direct two-stage least squares), Probit-2sls (IV regression estimated by Probit and two-stage least squares), Probit-ols (IV two-step regression estimated by Probit and ordinary least squares), and Heckit (Heckman two-step selection model). An extensive treatment of the conditions under which previous methods provide consistent estimation of ATE, ATET and ATENT can be found, for instance, in Wooldgrige (2002, Chapter 18). The value added of this new STATA command is that it allows for a generalization of the regression approach typically employed in standard program evaluation, by assuming heterogeneous response to treatment.

Keywords: Micro-econometrics, Treatment models, Instrumental variables, STATA routine

JEL Codes: C21, C87, D04
CONTENTS

INTRODUCTION .............................................................................................................5
1. TREATMENT EFFECT: DEFINITION AND STATISTICAL SET-UP .........................6
2. RANDOM AND NON-RANDOM ASSIGNMENT ...................................................8
3. SELECTION ON OBSERVABLES AND SELECTION ON UNOBSERVABLES .................9
   3.1 Selection on observables ..................................................................................9
   3.2 Selection on unobservables ..............................................................................10
4. ESTIMATION METHODS ......................................................................................11
   4.1 Control-function regression ..........................................................................11
   4.2 Instrumental variables ....................................................................................13
   4.3 Selection model ..............................................................................................18
5. THE STATA COMMAND IVTREATREG .................................................................20
6. USING IVTREATREG IN PRACTICE: AN APPLICATION TO THE RELATION BETWEEN EDUCATION AND FERTILITY ........................................23
REFERENCES ...........................................................................................................30
INTRODUCTION

It is nowadays common practice, especially at policymaking level, to perform ex-post evaluation of economic and social programs via evidence-based statistical analysis. This effort is mainly devoted to measure “causal effects” of an intervention on the part of an external authority (generally, local or national Government) on a set of subjects (individuals, firms, etc.) undergoing the program. But also in an environment not characterized by a formal policy intervention, rethinking usual causal relations in a counterfactual stance is becoming popular in the modern micro-econometric practice. In this regard, several new user-written STATA commands to accomplish the task of enlarging the set of statistical tools to perform counterfactual causal analysis have been recently realized (some examples are: treatreg that is an in-build STATA command; pscore by Becker and Ichino, 2002; psmatch2 by Leuven and Sianesi, 2003; nnmatch by Abadie et al. 2004).

This paper develops on this wake by presenting a new user-written STATA routine called ivtreatreg for the estimation of five different (binary) treatment models with and without idiosyncratic (or heterogeneous) average treatment effect. To our knowledge no previous STATA commands addressed this objective. Depending on the model specified by the user, ivtreatreg provides consistent estimation of average treatment effects both under the hypothesis of “selection on observables” and “selection on unobservables” by using Ordinary Least Squares (OLS) regression in the first case, and Instrumental-Variables (IV) and Selection-model (à la Heckman) in the second one. Conditional on a pre-specified subset of exogenous variables $x$ – thought of as driving the heterogeneous response to treatment – ivtreatreg calculates for each model the Average Treatment Effect (ATE), the Average Treatment Effect on Treated (ATET) and the Average Treatment Effect on Non-Treated (ATENT), as well as the estimates of these parameters conditional on the observable factors $x$, i.e., $\text{ATE}(x)$, $\text{ATET}(x)$ and $\text{ATENT}(x)$. The five models estimated by ivtreatreg are: Cf-ols (Control-function regression estimated by OLS), Direct-2sls (IV regression estimated by direct two-stage least squares), Probit-2sls (IV regression estimated by Probit and two-stage least squares), Probit-ols (IV two-step regression estimated by Probit and ordinary least squares), and Heckit (Heckman two-step selection model). An extensive treatment of the conditions under which previous methods provide consistent estimation of ATE, ATET and ATENT can be found, for instance, in Wooldgrige (2002, Chapter 18). The value added of this new STATA command is that it allows for a generalization of the
regression approach typically employed in standard program evaluation, by assuming heterogeneous response to treatment.

Section 2, 3 and 4 put forward a brief account of definitions and statistical background needed to present in section 5 the five treatment models estimated by ivtreatreg. Section 6 presents and discusses the “help” of this routine, while section 7 ends the paper by providing a didactic application of ivtreatreg on real data for studying the relation between education and fertility on a set of women living in a developing country.

1. TREATMENT EFFECT: DEFINITION AND STATISTICAL SET-UP

From a statistical point of view, our background is that of an analyst interested in the estimation of the so-called “treatment effect” of a given policy program in a “non-experimental” set-up, where the treatment variable w (taking value 1 for treated and 0 for untreated units) is expected to affect a specific target variable y (that can have a variety of forms: binary, count, continuous, etc.). In this context, we define the unit i’s Treatment Effect (TE) as:

$$\text{TE}_i = y_{1i} - y_{0i}$$

where $y_{1i}$ is the outcome of unit i when it is treated, and $y_{0i}$ is the outcome of unit i when it is not treated. Identifying TE$_i$ is not possible: in fact, as this quantity refers to the same individual at the same time, it goes without saying that the analyst can observe just one of the two quantities feeding into TE$_i$ (i.e. $y_{1i}$ or $y_{0i}$) but never both. For instance, it might be the case that we can observe the investment behavior of a supported company, but we cannot know what the investment of this company would have been if this firm had not been supported, and vice versa. The analyst faces a fundamental missing observation problem (Holland, 1986) that needs to be overcome to recover reliably the causal effect (Rubin, 1974; 1977). What on the contrary is observable to the analyst is the actual status of unit i, that is:

$$y_i = y_{0i} + w_i (y_{1i} - y_{0i})$$

This relation, called Potential Outcome Model, links together the treatment binary indicator, the observable and non observable outcomes. For identification purposes, the
treatment evaluation literature suggests to see at a specific effect called the Average Treatment Effect (ATE) of a given policy intervention, defined (in the population) as:\(^1\):

Average Treatment Effect = ATE = E(\(y_1 - y_0\))

Nevertheless, a policymaker might be interested also in knowing what is the effect on the subset of units actually treated. In this case, the parameter of interest is the so called Average Treatment Effect on Treated (ATET), defined as:

Average Treatment Effect on Treated = ATET = E(\(y_1 - y_0\) | \(w=1\))

Similarly, it is also possible to define the Average Treatment Effect on Non Treated (ATENT) that is the average treatment effect calculated within the subsample of untreated units:

Average Treatment Effect on Non Treated = ATENT = E(\(y_1 - y_0\) | \(w=0\))

The combined knowledge of ATE, ATET and ATENT can provide relevant information on how the causal relation between \(w\) and \(y\) actually behaves. Furthermore an interesting relation links these parameters, as it can be proved that:

\[
ATE = ATET P(w=1) + ATENT P(w=0)
\]

where \(P(w=1)\) is the probability of being treated, and \(P(w=0)\) of being untreated. But another important ingredient is needed to go on with the analysis of program evaluation. Indeed, for each individual, besides the observation on \(y\) and \(w\), analysts (normally) have access also to a certain number of observable covariates that can be collected in a row vector \(x\). Usually, the \(x\)-variables represent various individual characteristics such as: age, gender, income, etc.. The knowledge of \(x\)-variables, as we will see, is of primary usefulness in the estimation phase of previous parameters, as they represent essential observable confounding conditioning factors. It is then worth stressing that, under the knowledge of \(x\), we can also define the previous parameters “conditional on \(x\)”, as:

\[
ATE(x) = E(y_1 - y_0 | x)
\]
\[
ATET(x) = E(y_1 - y_0 | w=1, x)
\]
\[
ATENT(x) = E(y_1 - y_0 | w=0, x)
\]

\(^1\) For the sake of simplicity we avoid to write the subscript referring to unit \(i\) when we refer to the population parameters.
These quantities are, by definition, no more single values as before but functions of \( x \). It means that they can also be seen as “individual specific average treatment effects” as each individual owns a different and specific value of \( x \). Furthermore, it comes from the Law of Iterated Expectations that:

\[
\begin{align*}
\text{ATE} & = E_x\{\text{ATE}(x)\} \\
\text{ATET} & = E_x\{\text{ATET}(x)\} \\
\text{ATENT} & = E_x\{\text{ATENT}(x)\}
\end{align*}
\]

The aim of the econometrician involved into program evaluation is to recover consistent (and, when possible efficient) estimators of the previous parameters from observational data, that is from an i.i.d. sample of observed variables for each individual \( i \):

\[
\{y_i, w_i, x_i\} \text{ with } i = 1, \ldots, N
\]

Observe that, according to this set-up, we exclude the possibility that the treatment of one unit affects the outcome of another unit. In the literature this is called SUTVA (or stable unit treatment value assumption), and we will assume the validity of this hypothesis throughout this paper.

2. RANDOM AND NON-RANDOM ASSIGNMENT

If the sample were drawn at random (random assignment to program), it can be showed that \( \text{ATE} = \text{ATET} = \text{ATENT} \) and, more importantly, it is possible to estimate ATE as the difference between the sample mean of treated and the sample mean of untreated units: this is the well-known “difference-in-mean estimator” of classical statistics. Indeed, under random assignment, the so-called Independence Assumption (IA), stating that “\((y_1; y_0)\) are independent of \( w \)”, does hold and the “difference-in-mean” estimator is consistent, efficient and asymptotically normal.

When the sample of treated and untreated units, as it is often the case, is not randomly drawn, but it depends on either individual observable as well as unobservable to analyst characteristics, the difference-in-mean estimator is no longer a consistent estimation strategy. In this case, in fact, it occurs that “\((y_1; y_0)\) are dependent on \( w \)” so that a selection bias arises and it can be also proved that \( \text{ATE} \neq \text{ATET} \neq \text{ATENT} \).

What determines selection bias in program evaluation settings are basically to mechanisms: (i) the self-selection of individuals on the one hand, and (ii) the selection procedure from an external actor, on the other hand. Under “selection on observables”
the knowledge of \( x \) may be sufficient to identify previous causal parameters. Self-selection regards the choice of the individuals to apply for a specific program. This entails a cost-benefit calculus, as applying for a policy program can be costly to some extent. This choice may not be assumed to be done at random, as firms are endogenously involved in this decision. The selection mechanism is more intuitively following a non-random assignment, as a public agency is generally characterized by the pursuit of various objectives, such as direct (on the target-variable) and indirect (welfare) objectives. For instance, in order to maximize the final effect of an investment supporting program, the agency could apply the principle of “picking-the-winner”, that is choosing to support those units having an already high propensity to succeed. This is a sufficient condition to make the sample of beneficiaries far from being randomly built.

3. SELECTION ON OBSERVABLES AND SELECTION ON UNOBSERVABLES

3.1 Selection on observables

On the part of the evaluator, the factors affecting the non-random assignment of beneficiaries could have an observable or an unobservable nature. In the first case the analyst knows with precision what are the elements driving the self-selection of individuals and the selection of the agency. In this case the knowledge of \( x \), the structural variables that are supposed to drive the non-random assignment to treatment, are sufficient to identify, as we will see later, the actual effect of the policy in question. Nevertheless, when other factors driving the non-random assignment are impossible or difficult to observe, then the only knowledge of the vector \( x \) is not sufficient to identify the effect of the policy.

These two situations faced by the evaluator are known in the literature as the case of “selection on observable” and “selection on unobservables”: they ask for different methodologies to identify the actual effect of policy programs, and the greatest effort of past and current econometric literature has been that of dealing with these two situations and provide suitable solutions in both cases.

Under selection on observables the knowledge of \( x \), the factors driving the non-random assignment, may be sufficient to identify the causal parameters defined above. Of course, since the missing observation problem still holds, we need to rely on an assumption (or hypothesis) able to overcome that problem. Rosenbaum and Rubin (1983), introduced the so-called Conditional Independence Assumption (CIA), stating that - conditional on the knowledge of \( x \) - \( y_1 \) and \( y_0 \) are independent of \( w \), formally:
This assumption means that, once the knowledge of the factors affecting the sample selection are taken into account by the analyst, then the condition of randomization is restored. This assumption can be restricted to the so-called Conditional Mean Independence (CMI), stating that:

\[ E(y_1 | x, w) = E(y_1 | x) \quad \text{and} \quad E(y_0 | x, w) = E(y_0 | x) \]

that restricts the independence only on the mean. The CMI is the basis for (consistent) estimation of ATE, ATET and ATENT by parametric and non parametric methods. Within the parametric approaches the regression analysis is the most known and applied, while within the non-parametric ones the Matching methods and Reweighting are the most popular. But also the Sharp Regression Discontinuity Design brings to consistent estimation under CMI.

### 3.2 Selection on unobservables

When the selection into program is governed not only by observable-to-analyst factors, but also by unobservable variables, the CMI is not sufficient to identify causal parameters. Other assumptions are needed. Two classes of models are particularly suitable in this case: Selection model (à la Heckman) also known as Heckit Model and the Instrumental Variables (IV) approach. Before going on, it is worth to distinguish between “genuine unobservables” and “contingent unobservables”: the first type refers to factors that are intrinsically unknowable to the analyst as, for instance, some individual specific characteristics such as personal ability, propensity to bear risk, etc.; the second type refers to factors that, in principle, would be knowable, but that the available set of information prevents to employ. In many policy contexts the presence of contingent unobservables could be very problematic, as many (potentially observable) elements driving the selection into program could be overlooked, thus leaving the selection bias still present.

---

2 Furthermore, also the Fuzzy Regression Discontinuity Design can deal with selection on unobservables, as it can be proved that it is a particular kind of IV estimator. Finally, also the Difference-In-Differences (DID) estimator is able to treat unobservable selection, but it needs the availability of longitudinal data.

3 In the case of firm R&D and fixed investment support, for instance, many studies have a lot of information about firm characteristics, but very little about R&D projects’ quality. Since the selection is led by both these aspects, these studies run the risk to be severely biased (Cerulli, 2010).
4. ESTIMATION METHODS

The new STATA routine `ivtreaterg` implements the estimation of five models, where three of them are particular IV estimators. These methods are called: **cf-ols** (Control-function regression estimated by OLS), **direct-2sls** (IV regression estimated by direct two-stage least squares), **probit-2sls** (IV regression estimated by Probit and two-stage least squares), **probit-ols** (IV two-step regression estimated by Probit and ordinary least squares), and **heckit** (Heckman two-step selection model). Each of these can be estimated either by assuming homogenous or heterogeneous response to treatment (for a total of ten models). Before presenting how `ivtreaterg` actually works, the identification conditions, procedures and formulas of each model are briefly set out.

4.1 Control-function regression

By assuming Conditional Mean Independence (CMI), we have seen that:

\[ E(y_1 \mid x, w) = E(y_1 \mid x) \hspace{1cm} \text{and} \hspace{1cm} E(y_0 \mid x, w) = E(y_0 \mid x) \]

that restricts the independence only on the mean. Suppose to modeling the potential outcomes as follows:

(a) \( y_0 = \mu_0 + v_0 \), \( E(v_0) = 0 \), \( \mu_0 = \text{parameter} \)
(b) \( y_1 = \mu_1 + v_1 \), \( E(v_1) = 0 \), \( \mu_1 = \text{parameter} \)
(c) \( y = y_0 + w (y_1 - y_0) \)
(d) **CMI holds**

By substituting (a) and (b) into (c) we get:

\[ y = \mu_0 + w (\mu_1 - \mu_0) + v_0 + w (v_1 - v_0) \]

By assuming \( E(v_0 \mid x) = g_0(x) = x \beta_0 \) and \( E(v_1 \mid x) = g_1(x) = x \beta_1 \) we can distinguish two case:

a) **Case 1. Homogenous reaction function of** \( y_0 \) **and** \( y_1 \) **to** \( x \): \( E(v_1 \mid x) = E(v_0 \mid x) \)

In Case 1 we can show that:

1. \( E(y \mid w, x) = \mu_0 + w \text{ATE} + x \beta \)
2. \( \text{ATE} = \text{ATE}(x) = \text{ATET} = \text{ATET}(x) = \text{ATENT} = \text{ATENT}(x) = \mu_1 - \mu_0 \)
Thus, no heterogeneous average treatment effect (over \( x \)) does exist.

b) **Case 2. Heterogeneous reaction function of** \( y_0 \) and \( y_1 \) to \( x \): \( E(v_i | x) \neq E(v_i | x) \)  

In this second case it can be showed that:

1. \( E(v_i | w, x) = \mu_0 + w \ \text{ATE} + x^T \beta_0 + w (x - \mu_x) \beta \)
2. \( \text{ATE} \neq \text{ATET} \neq \text{ATENT} \)

where an estimator for \( \mu_x = E(x) \) can be the simple sample mean of \( x \). In this case, heterogeneous average treatment effects (over \( x \)) exist and the population causal parameters take on the following form:

\[
\begin{align*}
\text{ATE} &= (\mu_1 - \mu_0) + \mu_x \beta \\
\text{ATE}(x) &= \text{ATE} + (x - \mu_x) \beta \\
\text{ATET} &= \text{ATE} + E_x \{x - \mu_x | w=1\} \beta \\
\text{ATET}(x) &= [\text{ATE} + (x - \mu_x) \beta | w=1] \\
\text{ATENT} &= \text{ATE} + E_x \{x - \mu_x | w=0\} \beta \\
\text{ATENT}(x) &= [\text{ATE} + (x - \mu_x) \beta | w=0]
\end{align*}
\]

whose sample equivalents are:

\[
\begin{align*}
\hat{\text{ATE}} &= \hat{\alpha} \\
\hat{\text{ATE}}(x) &= \hat{\alpha} + (x - \bar{x}) \hat{\beta} \\
\hat{\text{ATET}} &= \hat{\alpha} + \frac{1}{N} \sum_{i=1}^{N} w_i (x - \bar{x}) \hat{\beta} \\
\hat{\text{ATET}}(x) &= \left[ \hat{\alpha} + (x - \bar{x}) \hat{\beta} \right]_{(w=1)} \\
\hat{\text{ATENT}} &= \hat{\alpha} + \frac{1}{\sum_{i=1}^{N} (1 - w_i)} \sum_{i=1}^{N} (1 - w_i) (x_i - \bar{x}) \hat{\beta} \\
\hat{\text{ATENT}}(x) &= \left[ \hat{\alpha} + (x_i - \bar{x}) \hat{\beta} \right]_{(w=0)}
\end{align*}
\]

Operationalizing regression in Case 2 is fairly straightforward:

1. estimate \( y_i = \mu_0 + w_i \alpha + x_i \beta_0 + w_i (x_i - \mu_x) \beta + \text{error}_i \) by OLS, thus getting consistent estimates of \( \mu_0 \), \( \alpha \), \( \beta_0 \) and \( \beta \);
2. plug these estimated parameters into the sample formulas and recover all the causal effects.
3. Obtain standard errors for ATET and ATENT via bootstrapping.
4.2 Instrumental variables

When the CMI hypothesis does not hold, Control-function regression brings to biased estimates of ATE, ATET and ATENT. This happens when the selection-into-treatment is due not only to observable, but also “unobeservable-to-analyst” factors. In this case, \( w \) becomes endogenous, that is, correlated with the regression error term. Instrumental-variables estimation (hereinafter, IV) solves this problem by restoring costincency also under the hypothesis of selection on unobservables. Nevertheless, the application of IV requires the availability of at least one variable \( z \), called “instrumental variable”, assumed to have the following two properties:

1. \( z \) is (directly) correlated with treatment \( w \)
2. \( z \) is (directly) uncorrelated with outcome \( y \).

This means that the selection into program depends on the same factors affecting the outcome plus \( z \) that does not affect directly the outcome (but only indirectly via its effect on \( w \)). This is the basic exclusion restriction under which IV is able to identify casual parameters.

Now, consider again the switching random coefficient model:

\[
y = \mu_0 + w (\mu_1 - \mu_0) + v_0 + w (v_1 - v_0)
\]

Here, when CMI does not hold we have that: \( E(v_1| w, x) \neq E(v_1| x) \) and \( E(v_0| w, x) \neq E(v_0| x) \). As in the case of control-function, we can distinguish these two cases.

a) Case 1. \( v_1 = v_0 \) (homogenous case)

In this case \( v_1 = v_0 \) so that \( y = \mu_0 + w (\mu_1 - \mu_0) + v_0 \) implying that:

\[\text{ATE} = \text{ATET} = \text{ATENT} = \mu_1 - \mu_0.\]

Suppose to have access to a variable \( z \) (instrumental variable) having these two properties:

1. \( E(v_0| x, z) = E(v_0| x) \iff z \) is uncorrelated with \( v_0 \)
2. \( E(w| x, z) \neq E(w| x) \iff z \) is correlated with \( w \)

Taking (1), we assume that: \( E(v_0| x, z) = E(v_0| x) = g(x) = x\beta \) meaning means that \( E(v_0| x, z) \neq 0 \). After simple manipulations, we get a regression model having a error term with zero unconditional mean of this type:

\[
y = \mu_0 + w \text{ATE} + x\beta + u_0
\]
that is a regression model in which \((x, z)\) are uncorrelated with the error term \(u_0\) (i.e., \((x, z)\) are exogenous) but the error term \(u_0\) is correlated with \(w\). These conditions bring to the following \emph{Structural System of (two) Equations}:

\[
\begin{align*}
(a) & \quad y_i = \mu_0 + w_i \text{ATE} + x_i \beta + u_{0i} \\
(b) & \quad w_i^* = \eta + q_i \delta + \epsilon_i \\
(c) & \quad w_i = \begin{cases} 
1 & \text{if } w_i^* \geq 0 \\
0 & \text{if } w_i^* < 0 
\end{cases} \\
(d) & \quad q_i = (x_i, z_i)
\end{align*}
\]

where \text{ATE} cannot be consistently estimated by OLS because conditions \(\text{Cov}(u_{0i} ; \epsilon_i) \neq 0\) i.e., \(w\) is \emph{endogenous} in equation (a). Equation (a) is known as the \emph{outcome equation}, equation (b) and (c) is known as the \emph{selection equation} and relation (d) is the \emph{exclusion restriction}. How can we estimate consistently \text{ATE} in System (11)? We may rely on three (consistent, but differently efficient) methods:

1. Direct Two-Stage-Least-Squares (2SLS)
2. Probit-2SLS
3. Probit-OLS

1. \textit{Direct Two-Stage-Least-Squares (Direct-2SLS)}

By using direct-2SLS the analyst does not consider at all the binary nature of \(w\). It follows two steps:

1. run an OLS regression of \(w\) on \(x\) and \(z\) of the type: 
\[w_i = \eta + x_i \delta_x + z_i \delta_z + \text{error}_i,\] thus getting the “predicted values” of \(w_i\), that we indicate with \(w_{p,i};\)
2. run a second OLS of \(y\) on \(x\) and \(w_{p,i}\). The coefficient of \(w_{p,i}\) is a consistent estimation of \text{ATE}.

2. Probit-2SLS

In this case, the analyst exploits suitably the \emph{binary} nature of \(w\): first he applies a Probit of \(w\) on \(x\) and \(z\), getting the “predicted probability of \(w\)”, and then he uses these probabilities by applying a 2SLS with predicted probabilities as instrument for \(w\).

Probit-2SLS is generally \emph{more efficient} than Direct-2SLS. Among all the possible instruments for \(w\), the optimal one is the \emph{orthogonal projection} of \(w\) in the vector space...
generated by \((x, z)\). Why and which is this projection? \(E(w \mid x, z)\) is the orthogonal projection of \(w\) in the vector space generated by \((x, z)\). Among all the projections, the orthogonal one produces the “smallest error”. But we know that \(E(w \mid x, z) = P(w=1 \mid x, z)\) = Probit selection equation. It means that the “probabilities of getting treated” (i.e., the propensity scores) estimated from the Selection Equation is the best instrument for \(w\) (because it generates the smallest projection error). Operationally, Probit-2SLS follows these three steps:

1. apply a Probit of \(w\) on \(x\) and \(z\), getting \(p_w\), i.e., the “predicted probability of \(w\)”;
2. run OLS of \(w\) on \((1, x, p_w)\), thus getting the fitted values \(w_{2fv,i}\);
3. run a second OLS of \(y\) on \((1, x, w_{2fv,i})\).

The coefficient of \(w_{2fv,i}\) is the most efficient estimator of ATE in the class of linear instruments for \(w\). Furthermore, this procedure does not require for consistency that the process generating \(w\) is correctly specified.

3. Probit-OLS

This method exploits the previous relation \(E(w \mid x, z) = P(w=1 \mid x, z)\). By tacking the expectation of \(y\) conditional on \((x, z)\), we get:

\[
E( y \mid x, z) = \mu_0 + \text{ATE} \cdot E(w \mid x, z) + x\beta
\]

Since we saw that \(E(u_0 \mid x, z) = 0\). By plug-in (12) into the previous equation we have:

\[
E( y \mid x, z) = \mu_0 + \text{ATE} \cdot P(w = 1 \mid x, z) + x\beta
\]

This relation suggests to estimate consistently ATE with a simple OLS regression of \(y\) on \((1, p_w, x)\). This model, however, is less efficient than Probit-2SLS and requires for consistency that the Probit is “correctly” specified. Standard errors have to be corrected for the presence of a “generated regressor” and “heteroscedasticity”.

From a technical point of view, in order to identify \((\mu_0, \text{ATE}, \beta)\) in equation (1), it not necessary to introduce \(z\) in the selection equation (2). It is sufficient that the selection equation (2) contains just \(x\). Indeed, since \(G(x, \delta)\) is a non-linear function of \(x\), then it is not perfectly collinear with \(x\). Therefore, \(G(x, \delta)\) can be used as instrument besides \(x\), as it does not produce problems of collinearity (as it occurs, conversely, if \(G\) is a linear probability model). Nevertheless, since \(x\) and \(G(x, \delta)\) are strongly correlated and are used jointly as instruments, it can be proved that the IV estimator gets larger variances, thereby becoming more imprecise.
b) Case 2. \( v_1 \neq v_0 \)

Consider now the case in which: \( v_1 \neq v_0 \) so that: \( y = \mu_0 + w(\mu_1 - \mu_0) + v_0 + w(v_1 - v_0) \).

As in the case of Control-Function, it implies that \( \text{ATE} \neq \text{ATET} \neq \text{ATENT} \). We are in the case of observable heterogeneity and \( \text{ATE}(x) \), \( \text{ATET}(x) \) and \( \text{ATENT}(x) \) can be defined and estimated. Suppose that \( v_1 \) and \( v_0 \) are independent on \( z \): it means that \( z \) is assumed to be endogenous in this model, that is:

\[
E(v_0 | x, z) = E(v_0 | x) = g_0(x) \\
E(v_1 | x, z) = E(v_1 | x) = g_1(x)
\]

It is equivalent to write:

\[
v_0 = g_0(x) + e_0 \quad \text{with} \quad E(e_0 | x, z) = 0 \\
v_1 = g_1(x) + e_1 \quad \text{with} \quad E(e_1 | x, z) = 0
\]

By substituting these expressions for \( v_0 \) and \( v_1 \) into the previous switching regression for \( y \), we get:

\[
y = \mu_0 + \alpha w + g_0(x) + w[g_1(x) - g_0(x)] + e_0 + w(e_1 - e_0)
\]

Now, by assuming in the previous equation: \( g_0(x) = x\beta_0 \), \( g_1(x) = x\beta_1 \), \( \varepsilon = e_0 + w(e_1 - e_0) \) and by applying the same procedure of case 1, we finally get:

\[
y = \mu_0 + \text{ATE}w + x\beta_0 + w(x - \mu_x)\beta + \varepsilon
\]

In this model we have two endogenous variables: \( w \) and \( w(x - \mu_x) \). Intuitively, if \( q = q(x, z) \) is an instrument for \( w \), then a suitable instrument for \( w(x - \mu_x) \) is: \( q(x - \mu_x) \). Nevertheless, before applying IV, we have to distinguish two sub-cases:

- Case 2.1: \( e_1 = e_0 \) (only observable heterogeneity)
- Case 2.2: \( e_1 \neq e_0 \) (both observable and unobservable heterogeneity)

In what follows we examine the two cases separately.

- Case 2.1: \( e_1 = e_0 \) (only observable heterogeneity)

In this case we have \( \varepsilon = e_0 \). By remembering that \( E(e_0 | x, z) = 0 \) we can conclude that:

\[
y = \mu_0 + aw + x\beta_0 + w(x - \mu_x)\beta + e_0 \quad , \quad \text{with} \quad E(e_0 | x, z, w) = E(e_0 | w)
\]

meaning that what is remaining is just the endogeneity due to \( w \).
Therefore, the following procedure provides consistent and efficient estimation:

1. apply a Probit of \( w \) on \( x \) and \( z \), getting \( p_w \), i.e., the “predicted probability of \( w \)”
2. estimate the following equation:
   \[
   y_i = \mu_0 + \alpha w_i + x_i \beta_0 + w_i (x_i - \mu_x) \beta + error_i
   \]
   using as instruments: \( 1, p_w, x_i, p_w(x_i - \mu_x) \).

Moreover, various functions and interactions of \((x, z)\) can be used to generate additional instruments, in order to get over-identification, and thus test the (joint) exogeneity of instruments.

- **Case 2.2:** \( e_1 \neq e_0 \) (both observable and unobservable heterogeneity)
  In this case, as seen, the full (and more general) model is:

\[
y = \mu_0 + aw + g_0(x) + w[g_1(x) - g_0(x)] + e_0 + w(e_1 - e_0)
\]

and we have to find a condition to restore consistent estimation. A possible condition could be: \( E[w(e_1 - e_0) | x, z] = E[w(e_1 - e_0)] \). Given this condition, and by applying previous procedures, we arrive to the following parametric equation for \( y \):

\[
y = \mu_0 + aw + x \beta_0 + w(x - \mu_x) \beta + e_0 + w(e_1 - e_0)
\]

By defining:

\[
r = w(e_1 - e_0) - E[w(e_1 - e_0)]
\]

and by adding and subtracting \( E[w(e_1 - e_0)] \) in the previous equation for \( y \), we get:

\[
y = \eta + aw + x \beta_0 + w(x - \mu_x) \beta + e_0 + r
\]

where \( \eta = \mu_0 + E[w(e_1 - e_0)] \). It is immediate to see that \( E(e_0 + r | x, z) = 0 \). It means that any function of \((x, z)\) can be used as instrument in the \( y \)-equation. It brings to apply the IV procedure identical to that for Case 2.1, that is, estimate:

\[
y_i = \eta + aw_i + x_i \beta_0 + w_i (x_i - \mu_x) \beta + error_i
\]

using as instruments: \( 1, p_w, x_i, p_w(x_i - \mu_x) \). This IV estimator is consistent, but not efficient. To get an efficient estimation it needs to introduce additional hypotheses. There are recent contributions, using more or less parametric approaches, to restore efficiency. In what follow we focus on the Heckit model with unobservable heterogeneity. It is a strong parametric model, but it may be useful sometimes to get efficient estimation. We will treat this model in the part on “Selection Models”.
c) Problems with IV

The main drawback of IV approaches regards the availability of good instruments. To be good an instrument has to be:

1. exogenous for the outcome $y$
2. sufficiently well correlated with $w$

If one of these two conditions is not met, the correctness of IV estimation is questionable. Usually, it is fairly difficult to find a variable that explains the selection-into-program having, at the same time, no (direct) relation with the outcome. When such a variable is available, anyways, its *exogeneity* is not easily testable. Indeed, testing instruments’ exogeneity requires to rely on an *over-identified* setting, that is, to get access to more than one instrument for $w$ (at least two). Observe that, in this case, the analyst can test only the *joint exogeneity* of all the instruments used and not that of each single instrument. In the case of *just-identified* settings (only one instrument for $w$), testing instrument’s exogeneity is not possible, and analysts normally have to discuss very carefully the suitability of the instrument adopted.

4.3 Selection model

From the IV-estimation section, in the Case 2.2, we had that:

$$y = \mu_0 + \alpha w + g_0(x) + w[g_1(x) - g_0(x)] + e_0 + w(e_1 - e_0)$$

and after some manipulations:

$$y = \mu_0 + \alpha w + x\beta_0 + w(x - \mu_x)\beta + e_0 + w(e_1 - e_0)$$

This model, as said, presents both observable and unobservable heterogeneity, and a consistent estimation in this case requires strong hypotheses (see the IV section). Nevertheless, we can use a generalized Heckit model to estimate consistently and efficiently such a model. The prize is that of relying on some distributional hypotheses.
The model is made of these assumptions:

1. \( y = \mu_0 + \alpha w + x \beta_0 + w (x - \mu_x) \beta + u \)
2. \( E(e_1 | x, z) = E(e_0 | x, z) = 0 \)
3. \( w = 1[0_0 + \theta_1 x + \theta_2 z + a \geq 0] \)
4. \( E(a | x, z) = 0 \)
5. \( (a, e_0, e_1) \sim \text{N} \)
6. \( a \sim \text{N}(0,1) \Rightarrow \sigma_a = 1 \)
7. \( u = e_0 + w(e_1 - e_0) \)

Given these starting conditions, we can directly calculate to what is equal \( E(y | x, z, w) \). To that end, write the \( y \)-equation as \( y = A + u \), with \( A = \mu_0 + \alpha w + x \beta_0 + w (x - \mu_x) \beta \) and \( u = e_0 + w(e_1 - e_0) \).

It can be proved that:

\[
E(y | x, z, w) = \mu_0 + \alpha w + x \beta_0 + w (x - \mu_x) \beta + \rho_w \phi(q\theta) \frac{\phi(q\theta)}{\Phi(q\theta)} + \rho_0 (1-w) \frac{\phi(q\theta)}{1-\Phi(q\theta)}
\]

where \( \rho_w = \sigma_{e_1} \sigma_{a_i} \) and \( \rho_0 = \sigma_{e_0} \sigma_{a_i} \). For the estimation of this equation a two-step procedure can be performed:

1. run a Probit regression of \( w_i \) on \((1, x_{ii}, z_i)\) and gets: \( (\hat{\rho}_i, \hat{\Phi}_i) \);
2. run an OLS of \( y_i \) on \( \left[1, w_i, x_i, w_i(x_i - \mu_x), w_i \frac{\phi\left(q\theta_i\right)}{\Phi\left(q\theta_i\right)}, (1-w_i) \frac{\phi\left(q\theta_i\right)}{1-\Phi\left(q\theta_i\right)} \right] \)

The previous two-step procedure produces \textit{consistent} and \textit{efficient} estimations. Given estimations, we can also test the hypothesis:

\( H_0: \rho_w = \rho_0 = 0 \)

that, if accepted, brings to the conclusion of \textit{no selection on unobservables}. Finally, by putting:

\[
\hat{\lambda}_i(q\theta) = \frac{\phi(q\theta)}{\Phi(q\theta)} \quad \text{and} \quad \hat{\lambda}_0(q\theta) = \frac{\phi(q\theta)}{1-\Phi(q\theta)}
\]

we can write the regression as:

\[
E(y | x, z, w) = \mu_0 + \alpha w + x \beta_0 + w (x - \mu_x) \beta + \rho_w \hat{\lambda}_i(q\theta) + \rho_0 (1-w) \hat{\lambda}_0(q\theta)
\]
Given the two-step estimation of the previous equation, once recovered all the parameters, it is possible to calculate the usual causal parameters. It is immediate to see, that:

\[ ATE = \alpha \]
\[ ATE(x) = \alpha + (x - \bar{x})\beta \]

Since it follows the same procedure as seen in the case of Control-function Case 2. Nevertheless, \( ATET(x) \), \( ATET \), \( ATENT(x) \) and \( ATENT \) assume a different form compared to Control-function Case 2. It is not difficult to show that:

\[ ATET(x) = \left[ \alpha + (x - \bar{x})\beta + (\rho_1 + \rho_0) \cdot \lambda_1(q\theta) \right] \]
\[ ATET = \alpha + \frac{1}{\sum w_i} \sum_{i=1}^{N} w_i (x_i - \bar{x})\beta + (\rho_1 + \rho_0) \cdot \frac{1}{\sum w_i} \sum_{i=1}^{N} w_i \cdot \lambda_1(q\theta) \]

and:

\[ ATENT(x) = \left[ \alpha + (x - \bar{x})\beta + (\rho_1 + \rho_0) \cdot \lambda_0(q\theta) \right] \]
\[ ATENT = \alpha + \frac{1}{\sum (1 - w_i)} \sum_{i=1}^{N} (1 - w_i)(x_i - \bar{x})\beta - (\rho_1 + \rho_0) \cdot \frac{1}{\sum (1 - w_i)} \sum_{i=1}^{N} (1 - w_i) \cdot \lambda_0(q\theta) \]

Given the estimation of \( \alpha, \rho_1, \rho_0, \beta, \lambda_1, \lambda_0 \) from the previous two-step procedure, all these causal effects can be calculated. Standard errors for \( ATET \) and \( ATENT \) can be obtained by bootstrapping.

5. THE STATA COMMAND IVTREATREG

The STATA routine ivtreatreg estimates the five binary treatment models presented above, with and without idiosyncratic (or heterogeneous) average treatment effect. Depending on the model specified, ivtreatreg provides consistent estimation of Average Treatment Effects either under the hypothesis of "selection on observables" (using the Control-function regression) or "selection on unobservables" (by using one of the three Intrumental-Variables (IV) models or the Heckman’s Selection-Model). Conditional on a pre-specified subset of exogenous variables - thought of as those driving the heterogeneous response to treatment - ivtreatreg calculates for each specific model the Average Treatment Effect (ATE), the Average Treatment Effect on Treated (ATET) and the Average Treatment Effect on Non-Treated (ATENT), as well
as the estimates of these parameters conditional on the observable factors \( x \) (i.e., ATE(\( x \)), ATET(\( x \)) and ATENT(\( x \))).

The syntax of the command is fairly simple and takes on this form:

\[
\text{Syntax of ivtreatreg} \\
\text{ivtreatreg outcome treatment [varlist] [if] [in] [weight], model(modeltype) [hetero(varlist_h) iv(varlist_iv) conf(number) graphic vce(robust) const(noconstant) head(noheader)]} \\
fweights, iweights, and pweights are allowed; see weight.
\]

where:

- **outcome** specifies the target variable that is the object of the evaluation.
- **treatment** specifies the binary (i.e. taking 0=treated or 1=untreated) treatment variable.
- **varlist** defines the list of exogenous variables that are considered as observable confounders.

The present routine allows for specifying a series of convenient options of different importance:

**Options of ivtreatreg**

- **model(modeltype)** specifies the treatment model to be estimated, where modeltype must be one of the following (and abovementioned) five models: "cf-ols", "direct-2sls", "probit-2sls", "probit-ols", "heckit". It is always required to specify one model.

<table>
<thead>
<tr>
<th>modeltype_options</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modeltype</strong></td>
<td></td>
</tr>
<tr>
<td>cf-ols</td>
<td>Control-function regression estimated by OLS</td>
</tr>
<tr>
<td>direct-2sls</td>
<td>IV regression estimated by direct two-stage least squares</td>
</tr>
<tr>
<td>probit-2sls</td>
<td>IV regression estimated by Probit and two-stage least squares</td>
</tr>
<tr>
<td>probit-ols</td>
<td>IV two-step regression estimated by Probit and OLS</td>
</tr>
<tr>
<td>heckit</td>
<td>Heckman two-step selection model</td>
</tr>
</tbody>
</table>

- **hetero(varlist_h)** specifies the variables over which to calculate the idiosyncratic Average Treatment Effect ATE(\( x \)), ATET(\( x \)) and ATENT(\( x \)), where \( x = \text{varlist\_h} \). It is optional for all models. When this option is not specified, the command estimates the specified model without heterogeneous average effect. Observe that \( \text{varlist\_h} \) should be the same set or a subset of the variables specified in \text{varlist}.

- **iv(varlist_iv)** specifies the variable(s) to be used as instruments. This option is strictly required only for "direct-2sls", "probit-2sls" and "probit-ols", while it is optional for "heckit".

- **graphic** allows for a graphical representation of the density distributions of ATE(\( x \)), ATET(\( x \)) and ATENT(\( x \)). It is optional for all models and gives an outcome only if variables into \text{hetero()} are specified.

- **vce(robust)** allows for robust regression standard errors. It is optional for all models.

- **beta** reports standardized beta coefficients. It is optional for all models.
const(noconstant) suppresses regression constant term. It is optional for all models.

conf(number) sets the confidence level equal to the specified number. The default is number=95.

The routine creates also a number of variables that can be fruitfully used to inspect further into data:

_wvarname_h are the additional regressors used in model’s regression when hetero(varlist_h) is specified. They are created for all models.

_zvarname_h are the instrumental-variables used in model’s regression when hetero(varlist_h) and iv(varlist_iv) are specified. They are created only in IV models.

ATE(x) is an estimate of the idiosyncratic Average Treatment Effect.

ATET(x) is an estimate of the idiosyncratic Average Treatment Effect on treated.

ATENT(x) is an estimate of the idiosyncratic Average Treatment Effect on Non-Treated.

G_{f v} is the predicted probability from the Probit regression, conditional on the observable confounders used.

_wL0, wL1 are the Heckman correction-terms.

Interestingly, ivtreatreg returns also some useful scalars:

r(N_tot) is the total number of (used) observations.

r(N_treated) is the number of (used) treated units.

r(N_untreated) is the number of (used) untreated units.

r(ate) is the value of the Average Treatment Effect.

r(atet) is the value of the Average Treatment Effect on Treated.

r(atent) is the value of the Average Treatment Effect on Non-treated.

Finally, some remarks are useful before using this routine:

The treatment has to be a 0/1 binary variable (1 = treated, 0 = untreated).

The standard errors for ATET and ATENT may be obtained via bootstrapping.

When option hetero() is not specified, ATE(x), ATET(x) and ATENT(x) are singleton numbers equal to ATE=ATET=ATENT.

Since when hetero is not specified in model “heckit” ivtreatreg uses the in-built command treatreg, the following has to be taken into account: (i) option beta and option head(noheader) are not allowed; (ii) option vce takes this syntax: vce(vcetype), where vcetype may be “conventional”, “bootstrap”, or “jackknife”.

Please remember to use the update query command before running this program to make sure you have an up-to-date version of Stata installed.
6. USING IVTREATREG IN PRACTICE: AN APPLICATION TO THE RELATION BETWEEN EDUCATION AND FERTILITY

In order to see how ivtreatreg actually works, we consider an instructional dataset called FERTIL2.DTA accompanying the manual *Introductory Econometrics: A Modern Approach*, by Wooldridge (2000) collecting cross-sectional data on 4,361 women of childbearing age in Botswana. It is freely downloadable at http://fmwww.bc.edu/ec-p/data/wooldridge/FERTIL2.dta and a description of this dataset is presented below.

Table 1. Description of the dataset FERTIL2.DTA.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>mnthborn</td>
<td>month woman born</td>
</tr>
<tr>
<td>yearborn</td>
<td>year woman born</td>
</tr>
<tr>
<td>age</td>
<td>age in years</td>
</tr>
<tr>
<td>electric</td>
<td>=1 if has electricity</td>
</tr>
<tr>
<td>radio</td>
<td>=1 if has radio</td>
</tr>
<tr>
<td>tv</td>
<td>=1 if has tv</td>
</tr>
<tr>
<td>bicycle</td>
<td>=1 if has bicycle</td>
</tr>
<tr>
<td>educ</td>
<td>years of education</td>
</tr>
<tr>
<td>ceb</td>
<td>children ever born</td>
</tr>
<tr>
<td>agefbrth</td>
<td>age at first birth</td>
</tr>
<tr>
<td>children</td>
<td>number of living children</td>
</tr>
<tr>
<td>knowmeth</td>
<td>=1 if know about birth control</td>
</tr>
<tr>
<td>usemeth</td>
<td>=1 if ever use birth control</td>
</tr>
<tr>
<td>monthfm</td>
<td>month of first marriage</td>
</tr>
<tr>
<td>yearfm</td>
<td>year of first marriage</td>
</tr>
<tr>
<td>agefm</td>
<td>age at first marriage</td>
</tr>
<tr>
<td>idlnchld</td>
<td>'ideal' number of children</td>
</tr>
<tr>
<td>heduc</td>
<td>husband's years of education</td>
</tr>
<tr>
<td>agesq</td>
<td>age^2</td>
</tr>
<tr>
<td>urban</td>
<td>=1 if live in urban area</td>
</tr>
<tr>
<td>urb Educ</td>
<td>urban*educ</td>
</tr>
<tr>
<td>spirit</td>
<td>=1 if religion == spirit</td>
</tr>
<tr>
<td>protest</td>
<td>=1 if religion == protestant</td>
</tr>
<tr>
<td>catholic</td>
<td>=1 if religion == catholic</td>
</tr>
<tr>
<td>frsthalf</td>
<td>=1 if mnthborn &lt;= 6</td>
</tr>
<tr>
<td>educ0</td>
<td>=1 if educ == 0</td>
</tr>
<tr>
<td>evermarr</td>
<td>=1 if ever married</td>
</tr>
<tr>
<td>educ7</td>
<td>=1 if educ &gt;= 7</td>
</tr>
</tbody>
</table>
This dataset contains 28 variables on various woman and family characteristics. In this exercise, we are in particular interested in evaluating the impact of the variable `educ7` (taking value 1 if a woman has more than or exactly seven years of education and 0 otherwise) on the number of family children (`children`). Several conditioning (or confounding) observable factors are included in the dataset, such as: the age of the woman (`age`), whether or not the family owns a TV (`tv`), whether or not the woman lives in a city (`urban`), and so forth. In order to inquiry into the relation between education and fertility and according to Wooldridge (2002, example 18.3, p. 624) we estimate the following specification for each of the five models implemented by `ivtreatreg`:

```
set more off
xi: ivtreatreg children educ7 age agesq evermarr urban electric tv , ///
    hetero(age agesq evermarr urban) iv(frsthalf) model(modeltype)
    graphic
```

As proposed by Wooldridge (2002) this specification adopts - as instrumental variable - the covariate `frsthalf` taking value 1 if the woman was born in the first six month of the year and zero otherwise. This variable is (partially) correlated with `educ7`, but should not have any direct relation with the number of family children.

The simple difference-in-mean estimator (the mean of children in the group of more educated women, the treated ones, minus the mean of children in the group of less educated women, the untreated ones) is -1.77 with a t-value of -28.46. It means that more educated women show – without ceteris paribus conditions – about two children less than lower educated ones. By adding confounding factors in the regression specification, we get the OLS estimate of ATE that, again in absence of heterogeneous treatment, is -0.394 with a t-value of -7.94: it is still significant, but the magnitude, as expected, reduces considerably compared to the difference-in-mean estimation thus showing that confounders are relevant. When we consider OLS estimation with heterogeneity, we get an ATE equal to 0.67 still significant at 1% (see column on CF-OLS in Table 4).

When we consider IV estimation, results change dramatically. As working example of how to use `ivtreatreg`, we estimate previous specification in the case of `probit-2sls` (with heterogeneous treatment response). The main outcome is reported in Table 2, where both results from the probit first-step and the IV regression of the second-step are set out. Results on the probit show that `frsthalf` is partially fairly correlated with `educ7`, thus it can be reliably used as instrument for this variable. Step 2 shows that the ATE (again, the coefficient of `educ7`) is no more significant and, above all, it changes the sign by becoming positive and equal to 0.30.
Table 2. Results from ivtreateg when probit-2SLS is the specified model and treatment heterogeneous response is assumed.

<table>
<thead>
<tr>
<th>Step 1. Probit regression</th>
<th>Number of obs = 4358</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR chi2(7)</td>
<td>1130.84</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-2428.384</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.1889</td>
</tr>
</tbody>
</table>

**educ7 | Coef. Std. Err. z P>|z| [95% Conf. Interval]**

| frsthalf | -0.2206627 | 0.0418563 | -5.27 0.000 | -0.3026995 | -0.1386259 |
| age      | -0.0150337 | 0.0174845 | -0.86 0.390 | -0.0493027 | 0.0192354  |
| agesq    | -0.007325  | 0.002897  | -2.53 0.011 | -0.013003  | -0.0001647 |
| evermarr | -0.2972879 | 0.0486734 | -6.11 0.000 | -0.392686  | -0.2018988 |
| urban    | 0.2998122  | 0.0432321 | 6.93 0.000 | 0.2150789  | 0.3845456  |
| electric | 0.4246668  | 0.0751255 | 5.65 0.000 | 0.2774235  | 0.57191    |
| tv       | 0.9281707  | 0.0977462 | 9.50 0.000 | 0.7365915  | 1.11975    |
| _cons    | 1.13537    | 0.2440057 | 4.65 0.000 | 0.6571273  | 1.613612   |

<table>
<thead>
<tr>
<th>Step 2. Instrumental variables (2SLS) regression</th>
<th>Number of obs = 4358</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>SS</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>children</td>
<td>21510.0321</td>
</tr>
<tr>
<td>-ws_age</td>
<td>0.011469</td>
</tr>
<tr>
<td>-ws_agesq</td>
<td>-0.8979833</td>
</tr>
<tr>
<td>-ws_evermarr</td>
<td>0.4167504</td>
</tr>
<tr>
<td>-ws_urban</td>
<td>0.859302</td>
</tr>
<tr>
<td>age</td>
<td>0.010030</td>
</tr>
<tr>
<td>agesq</td>
<td>1.253709</td>
</tr>
<tr>
<td>evermarr</td>
<td>-0.531325</td>
</tr>
<tr>
<td>urban</td>
<td>-0.2392104</td>
</tr>
<tr>
<td>electric</td>
<td>-0.234937</td>
</tr>
<tr>
<td>_cons</td>
<td>-13.7584</td>
</tr>
</tbody>
</table>

**Instrumented:** educ7 ws_age ws_agesq ws_evermarr ws_urban
**Instruments:** age agesq evermarr urban electric tv g_fv z_age z_agesq z_evermarr z_urban

This result is in line with the IV estimation obtained by Wooldridge. Nevertheless, having assumed heterogeneous response to treatment allows now to calculate also the ATET and ATENT and to inspect into the cross-unit distribution of these effects. First of all, ivtreateg returns these parameters as scalars (along with treated and untreated sample size):
In order to get the standard errors for testing ATET and ATENT significance, a bootstrap procedure can be easily implemented as follows:

```
xi: bootstrap atet=r(atet) atent=r(atent), rep(100): ///
  ivtreatreg children educ7 age agesq evermarr urban electric tv, ///
  hetero(age agesq evermarr urban) iv(frsthalf) model(probit-2sls)
```

Table 3 shows the result. As it can be immediate to see, both ATET and ATENT are not significant and show values quite different but not too much far from that of ATE. Furthermore, a simple check should show that $ATE = ATET P(w=1) + ATENT P(w=0)$:

```
.dl "ATE= (r(N_treat)/r(N_tot))*r(atet)+(r(N_untreat)/r(N_tot))*r(atent)"
ATE= .30040086
```

that confirms the expected result. Finally, we may analyze the distribution of $ATE(x)$, $ATET(x)$ and $ATENT(x)$ in this case and Figure 2 shows the result.

### Table 3. Bootstrap standard errors for $ATET(x)$ and $ATENT(x)$ using ivtreateg with model probit-2sls.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Bootstrap</th>
<th>Normal-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>atet</td>
<td>.89829</td>
<td>.5488267</td>
</tr>
<tr>
<td>atent</td>
<td>-.4468834</td>
<td>.4124428</td>
</tr>
</tbody>
</table>
What emerges is that ATET(x) shows a substantially uniform distribution, while both ATE(x) and ATENT(x) a distribution more concentrated on negative values. In particular ATENT(x) shows the highest modal value around -2.2 children, thus predicting that less educated women would have been less fertile if they had been more educated.

Table 4 shows ATE results for all the five models, and also for the simple “Difference-in-Mean” (t-test). The ATE obtained by IV methods is always not significant, but it has a positive sign only for probit-2sls. The rest of ATEs present always negative sign: it means that more educated women would have been more fertile if they had been less educated. The case of heckit is a little more puzzling as the result is significant and very close to the difference-in-mean estimation that is highly suspected to be bias. This could be due to the fact that the identification condition of heckit are not met in this dataset.
Table 4. Estimation of the ATE for the five models estimated by ivtreatreg.

<table>
<thead>
<tr>
<th>Variable</th>
<th>T-TEST</th>
<th>CF-OLS</th>
<th>PROBIT-OLS</th>
<th>DIRECT-2SLS</th>
<th>PROBIT_2SLS</th>
<th>HECKIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>educ7</td>
<td>-1.770***</td>
<td>-0.372***</td>
<td>-1.044</td>
<td>0.300</td>
<td>-1.915***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.06219</td>
<td>0.05020</td>
<td>0.66626</td>
<td>0.49956</td>
<td>0.39871</td>
<td></td>
</tr>
<tr>
<td>G_fv</td>
<td></td>
<td></td>
<td>-0.11395</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.50330</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2, finally, shows the plot of the average treatment effect distribution for each method. By and large, these distributions follow a similar pattern, although direct-2sls and heckit show some appreciable differences. The heckit, in particular, shows a pattern very different with a strong demarcation between the plot of treated and untreated units. As such, it seems not to a reliable estimation procedure and this should deserve further inspection.

Observe, finally, that the distributions for direct-2sls are largely more uniform than in the other cases where a strong left-side inflation dominates with the ATENT(x) more concentrated on negative values that ATET(x) on positive ones. What this might mean? It seems that the counterfactual condition of these women is not the same: on average, if a less educated woman became more educated, then their fertility would decrease more than the increase in fertility of more educated women becoming (in a virtual sense) less educated.
Figure 2. Distribution of $ATE(x)$, $ATET(x)$ and $ATENT(x)$ for the five models estimated by $ivtreatreg$. 
REFERENCES


