2012 UK Stata Users Group Meeting
Cass Business School, London, UK
September 13–14, 2012

`ctreatreg`: a STATA module for estimating *Dose Response Treatment Models* under (continuous) treatment endogeneity and heterogeneous response to observable confounders

Giovanni Cerulli

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

Introduction

Consider a policy program where a binary treatment is ***assigned not randomly*** (i.e., according to a *structural* rule).

The program provides a different **“level” of treatment** (or **dose *t***) to treated - ranging from 0 (absence of treatment) to 100 (maximum treatment level).

Two groups of units are thus formed:

(i) *untreated*, whose level of treatment (or “dose”) is zero, and
(ii) *treated*, whose level of treatment is greater than zero.

⇨ We are interested in estimating the causal effect of the treatment variable *t* on an outcome *y* by assuming that treated and untreated units may ***respond differently to observable confounders*** (**x**)

⇨ We wish to estimate a ***Dose-Response-Function of y on t***.

The STATA routine “ctreatreg” estimates a **Dose-Response-Function (DRF) for such a model**. It is shown to be equal to the “Average Treatment Effect (ATE), given the level of treatment *t*” (i.e. ATE(*t*)), along with other “causal” parameters of interest (ATE, ATET, ATENT).

Compared with similar models - as the one proposed by **Hirano and Imbens (2004)** implemented in STATA by **Bia and Mattei (2008)** – the present model:

1. does not need a full normality assumption, and
2. is well-suited when many individuals have a zero-level of treatment.
3. may account for treatment "endogeneity", by exploiting an Instrumental-Variables (IV) estimation.

The Dose-Response-Function is estimated by a **third degree polynomial approximation.**

Both **OLS** and **IV** estimation are available. In particular:

IV is based on two steps:

- ⇨ **STEP 1**. *Heckman bivariate selection model* of *w* (the yes/no decision to treat a given unit) and *t* (the level of the treatment provided) in the first step,
- ⇨ **STEP 2**. 2SLS estimation for the outcome (*y*) equation.

The routine provides also a graphical representation of results. An empirical application to real data will be set out.

The model

- □ Two exclusive outcomes => when a unit is treated: y_1 ; when the <u>*same*</u> unit is untreated: y_0;
- □ w is the treatment indicator, taking value 1 for treated and 0 for untreated units;
- □ $g_1(\mathbf{x})$ and $g_0(\mathbf{x})$ are the unit responses to the vector of *confounding variables* $\mathbf{x}$ when the unit is treated and untreated;
- □ μ_1 and μ_0 are two scalars;
- □ e_1 and e_0 two random variables with zero unconditional mean and constant variance;
- □ $h(t)$ is the *response function* to the level of treatment t.

The model takes on this form:

$$\begin{cases} w=1 \quad => \quad y_1 = \mu_1 + g_1(\mathbf{x}) + h(t) + e_1 \\ w=0 \quad => \quad y_0 = \mu_0 + g_0(\mathbf{x}) + e_0 \end{cases}$$

where:

$$\begin{cases} h(t) = 0 \quad if \quad w=0 \\ h(t) \neq 0 \quad if \quad w=1 \end{cases}$$

By assuming a parametric form of:

$$g_1(\mathbf{x}) = \mathbf{x}\boldsymbol{\delta}_1$$
$$g_0(\mathbf{x}) = \mathbf{x}\boldsymbol{\delta}_\mathbf{0}$$

Define the Average Treatment Effect (ATE) conditional on **x** as:

$$\text{ATE}(\mathbf{x};t)=\text{E}(y_1-y_0\,|\,\mathbf{x},t)=\begin{cases}(\mu_1-\mu_0)+\mathbf{x}(\boldsymbol{\delta}_1-\boldsymbol{\delta}_0)\ +h(t) & \text{if}\quad t>0\\ (\mu_1-\mu_0)+\mathbf{x}(\boldsymbol{\delta}_1-\boldsymbol{\delta}_0) & \text{if}\quad t=0\end{cases}$$

$$=\begin{cases}\mu+\mathbf{x}\boldsymbol{\delta}+h(t) & \text{if}\quad t>0\\ \mu+\mathbf{x}\boldsymbol{\delta} & \text{if}\quad t=0\end{cases}$$

thereby getting:

$$\text{ATE}(\mathbf{x},t,w)=\begin{cases}\text{ATE}(\mathbf{x},t>0) & \textit{if}\quad w=1\\ \text{ATE}(\mathbf{x},t=1) & \textit{if}\quad w=0\end{cases}=\text{I}(t>0)[\mu+\mathbf{x}\boldsymbol{\delta}+h(t)]+\text{I}(t=0)[\mu+\mathbf{x}\boldsymbol{\delta}]=$$

$$=w[\mu+\mathbf{x}\boldsymbol{\delta}+h(t)]+(1-w)[\mu+\mathbf{x}\boldsymbol{\delta}]$$

By averaging on $(\mathbf{x},t,w)$, the previous formula becomes:

$$\text{ATE}=N_T\,/\,N\cdot(\mu+\overline{\mathbf{x}}_{t>0}\boldsymbol{\delta}+\overline{h}_{t>0})+N_{NT}\,/\,N\cdot(\mu+\overline{\mathbf{x}}_{t=0}\boldsymbol{\delta})$$

Since by definition ATE = $p(w{=}1)$ · ATET + $p(w{=}0)$ · ATENT, we can get from the last row of the previous formula that:

$$\begin{cases}\text{ATE}=p(w=1)(\mu+\overline{\mathbf{x}}_{t>0}\boldsymbol{\delta}+\overline{h}_{t>0})+p(w=0)(\mu+\overline{\mathbf{x}}_{t=0}\boldsymbol{\delta})\\ \text{ATET}=\mu+\overline{\mathbf{x}}_{t>0}\boldsymbol{\delta}+\overline{h}_{t>0}\\ \text{ATENT}=\mu+\overline{\mathbf{x}}_{t=0}\boldsymbol{\delta}\end{cases}\qquad [1]$$

and by simple algebra:

$$\mathrm{ATE}(\mathbf{x},t,w) = w \cdot [\mathrm{ATET} + (\mathbf{x}_{t>0} - \overline{\mathbf{x}}_{t>0})\boldsymbol{\delta} + (h(t) - \overline{h}_{t>0})] + (1-w) \cdot [\mathrm{ATENT} + (\mathbf{x}_{t=0} - \overline{\mathbf{x}}_{t=0})\boldsymbol{\delta}]$$

It means that:

$$\begin{cases} \mathrm{ATET}(\mathbf{x},t) = \mathrm{ATE}(\mathbf{x},t,w=1) = \mathrm{ATET} + (\mathbf{x}_{t>0} - \overline{\mathbf{x}}_{t>0})\boldsymbol{\delta} + (h(t) - \overline{h}_{t>0}) \\ \mathrm{ATE}(\mathbf{x},t) = \mathrm{ATE}(\mathbf{x},t,w=0) = \mathrm{ATENT} + (\mathbf{x}_{t=0} - \overline{\mathbf{x}}_{t=0})\boldsymbol{\delta} \end{cases} \qquad [2]$$

where:

$$\begin{cases} \mathrm{ATET} = \mu + \overline{\mathbf{x}}_{t>0}\boldsymbol{\delta} + \overline{h}_{t>0} \\ \mathrm{ATENT} = \mu + \overline{\mathbf{x}}_{t=0}\boldsymbol{\delta} \end{cases} \qquad [3]$$

We can define the **Dose-Response-Function (DRF)** simply by averaging ATE(**x**, *t*) on **x**:

$$\mathrm{ATE}(t,w) = \mathrm{E}_{\mathbf{x}}\{\mathrm{ATE}(\mathbf{x},t,w)\} = w \cdot [\mathrm{ATET} + (h(t) - \overline{h}_{t>0})] + (1-w) \cdot \mathrm{ATENT}$$

that is:

$$\boxed{\mathrm{ATE}(t) = \begin{cases} \mathrm{ATET} + (h(t) - \overline{h}_{t>0}) & \text{if} \quad t > 0 \\ \mathrm{ATENT} & \text{if} \quad t = 0 \end{cases} \qquad [4]}$$

The estimation of [4] is main purpose of this paper.

The Regression Approach

From the Potential Outcome Model (POM), the observable outcome is $y = y_0 + w(y_1 - y_0)$ is:

$$y = \mu_0 + w \cdot [(\mu_1 - \mu_0) + \bar{\mathbf{x}}\boldsymbol{\delta} + \bar{h}] + \mathbf{x}\boldsymbol{\delta}_\mathbf{0} + w \cdot (\mathbf{x} - \bar{\mathbf{x}})\boldsymbol{\delta} + w \cdot (h(t) - \bar{h}) + e_0 + w \cdot (e_1 - e_0) \qquad [5]$$

By assuming ***Conditional Mean Independence*** (CMI), namely that – given **x** both *w* and *t* are *exogenous* in equation [5], we can write the regression line of *y* as:

$$\mathrm{E}(y \,|\, \mathbf{x}, w, t) = \mu_0 + \mathbf{x}\boldsymbol{\delta}_\mathbf{0} + w\underbrace{[(\mu_1 - \mu_0) + \bar{\mathbf{x}}\boldsymbol{\delta} + \bar{h}]}_{\text{ATE}} + w[\mathbf{x} - \bar{\mathbf{x}}]\boldsymbol{\delta} + w[h(t) - \bar{h}] \qquad [6]$$

In equation [6] we show (see the proof in the paper) that:

$$\text{ATE} = (\mu_1 - \mu_0) + \bar{\mathbf{x}}\boldsymbol{\delta} + \bar{h}.$$

This leads to the estimation of this regression equation:

$$\boxed{\mathrm{E}(y \,|\, \mathbf{x}, w, t) = \mu_0 + \mathbf{x}\boldsymbol{\delta}_\mathbf{0} + w\text{ATE} + w[\mathbf{x} - \bar{\mathbf{x}}]\boldsymbol{\delta} + w[h(t) - \bar{h}] \qquad [7]}$$

where the term $[h(t) - \bar{h}]$ can be estimated by *polynomial regression*.

Estimation of the Dose-Response-Function under CMI

By supposing a three degree polynomial form for the function $h(t)$ of this form:

$$h(t) = at + bt^2 + ct^3$$

We get that equation [7] becomes:

$$\mathrm{E}(y \mid \mathbf{x}, w, t) = \mu_0 + \mathbf{x}\boldsymbol{\delta}_{\mathbf{0}} + w\mathrm{ATE} + w[\mathbf{x} - \overline{\mathbf{x}}]\boldsymbol{\delta} + a[t - \mathrm{E}(t)]w + b[t^2 - \mathrm{E}(t^2)]w + c[t^3 - \mathrm{E}(t^3)]w \qquad [8]$$

Under CMI, an OLS estimation of equation [8] leads to ***consistent estimates*** of the parameters.

The ***Dose-Response-Function***, is estimated by:

$$\boxed{\widehat{\mathrm{ATE}}(t_i) = w[\widehat{\mathrm{ATET}} + \hat{a}(t_i - \frac{1}{N}\sum_{i=1}^{N} t_i) + \hat{b}(t_i^2 - \frac{1}{N}\sum_{i=1}^{N} t_i^2) + \hat{c}(t_i^3 - \frac{1}{N}\sum_{i=1}^{N} t_i^3)] + (1-w)\,\widehat{\mathrm{ATENT}} \qquad [9]}$$

A simple graph of the curve $\widehat{\mathrm{ATE}}(t_i)_{t_i>0}$ as function of *t*, returns the form of the DRF.

Estimation of the Dose-Response-Function under treatment “endogeneity”

When *w* (and thus *t*) are endogenous (i.e., CMI hypothesis does not hold) **OLS are *biased***. Nevertheless, an **Instrumental-Variables (IV)** estimation procedure may be implemented to restore consistency.

Express the model in extensive form:

$$y = \mu_0 + \mathbf{x}\boldsymbol{\delta}_\mathbf{0} + w\mathrm{ATE} + w[\mathbf{x} - \overline{\mathbf{x}}]\boldsymbol{\delta} + b[t^2 - \mathrm{E}(t^2)]w + c[t^3 - \mathrm{E}(t^3)]w_3 + \varepsilon$$

$$w = \begin{cases} 1 & \text{if} \quad w^* > 0 \\ 0 & \text{if} \quad w^* \le 0 \end{cases}$$

$$t = \begin{cases} t' & \text{if} \quad w^* > 0 \\ t^* & \text{if} \quad w^* \le 0 \end{cases}$$

- w^* is he latent unobservable counterpart of the binary treatment w;
- *t* is fully observed only when *w*=1 (and *t*=*t’*); otherwise it is unobserved (and put equal to zero).

By defining T_1=t-E(t), T_2=t^2-E(t^2) and T_3= t^3-E(t^3), the previous model may be re-written as follows:

$$\begin{cases} y = \mu_0 + \mathbf{x}\boldsymbol{\delta}_{\mathbf{0}} + w\mathrm{ATE} + w[\mathbf{x} - \overline{\mathbf{x}}]\boldsymbol{\delta} + awT_1 + bwT_1 + wT_3 + \varepsilon_y & [10-1] \\ w^* = \mathbf{x}_1\boldsymbol{\beta}_1 + \varepsilon_w & [10-2] \\ t' = \mathbf{x}_2\boldsymbol{\beta}_2 + \varepsilon_t & [10-3] \end{cases}$$

where:

✓ Error terms ε_w, ε_t and ε_y are supposed to be freely correlated with zero mean

✓ Equation [10-2] is the *selection* equation

✓ Vector of covariates $\mathbf{x}_1$ are the selection criteria

✓ Equation [10-3] is the *treatment-level* equation

✓ The vector of covariates $\mathbf{x}_2$ are exogenous variables determining the treatment level.

IV estimation

In equation [10-1], both *w* and T_1, T_2 and T_3 are endogenous. To estimate consistently the parameters of that system we may proceed in two steps:

1. ***First***: Estimate the last two equations [10-2]-[10-3] jointly by a Heckman two-step "bivariate sample-selection model" (Heckman, 1979).

   □ The Heckman two-step procedure performs a probit of *w* on $\mathbf{x}_1$ in the first step and a OLS regression of *t'* on $\mathbf{x}_2$ augmented with the *Mills' ratio* obtained from the probit in the second step.

2. ***Second***: Take the all sample predicted values of *w* (i.e. $\hat{p}_w$) and *t* (i.e. $\hat{t}$) from the previous Heckman estimation, and then we perform a 2SLS for equation [10-1] using as instruments the following exogenous variables $(\mathbf{x}, \hat{p}_w, \hat{p}_w[\mathbf{x}-\bar{\mathbf{x}}], \hat{p}_w\hat{T}_1, \hat{p}_w\hat{T}_2, \hat{p}_w\hat{T}_3)$, thus getting a consistent estimation of the coefficients $(\mu_0, \boldsymbol{\delta}_\mathbf{0}, \text{ATE}, \boldsymbol{\delta}, a, b, c)$.

   □ Observe that the instruments used are based on the orthogonal projection of *w* and *t* on the vector space generated by the all exogenous variables of the model.

IDENTIFICATION

The problem of this procedure is with ***parameters' identification***. To get precise estimation, we need at least one instrumental variable (*z*) appearing only in equation [10-2], that is, only able to explain directly the selection process (*exclusion restriction*). Thus, we run under the following identification assumption:

$$\mathbf{x}_1 = [\mathbf{x}; z]$$
$$\mathbf{x}_2 = [\mathbf{x}]$$

so that a full specified model (all the equations depend on the same exogenous **x**) is considered, where *z* is the ***instrumental variable*** directly correlated with the selection, but directly uncorrelated with the level of the subsidization as well as the level of the outcome. This procedure indentifies correctly the parameters of interest.

Estimation of comparative Dose-Response-Functions

The model allows also for calculating the ***average comparative response*** at different level of treatment (as in Hirano and Imbens, 2004). This quantity takes this formula:

$$\mathrm{ATE}(t,\Delta) = \mathrm{E}[y(t+\Delta) - y(t)] \qquad [11]$$

Equation [11] identifies the average treatment effect between two states (or levels of treatment): *t* and $t+\Delta$.
Given a certain level of:

$$\Delta = \overline{\Delta}$$

we can get a particular:

$$\mathrm{ATE}(t,\overline{\Delta})$$

i.e, the “treatment function at $\overline{\Delta}$”.

An estimation is given by (see paper):

$$\boxed{\mathrm{A\hat{T}E}(t,\Delta) = \hat{a}(t+\Delta) + \hat{b}(t+\Delta)^2 + \hat{c}(t+\Delta)^3 - [\hat{a}t + \hat{b}t^2 + \hat{c}t^3]}$$

We can use a *bootstrap* of $A\hat{T}E(t,\Delta)$ over $(\hat{a},\hat{b},\hat{c})$ to get the standard errors of $\mathrm{A\hat{T}E}(t,\Delta)$ and then its statistical significance at various level of *t*.

The STATA routine `ctreatreg`

```
help ctreatreg
-------------------------------------------------------------------------------

Title

    ctreatreg -  Dose-Response model with "continuous" treatment, endogeneity and heterogeneous response to
                   observable confounders


Syntax

        ctreatreg outcome treatment [varlist] [if] [in] [weight], model(modeltype) ct(treat_level)
                 [hetero(varlist_h) iv(instrument) delta(number) graphic conf(number) vce(robust) const(noconstant)
                 head(noheader) beta]


    fweights, iweights, and pweights are allowed; see weight.



Description

    ctreatreg estimates the Dose-Response-Function (DRF) of a given treatment on a specific target variable, within
    a model where units are treated with different levels. The DRF is defined as the “average treatment effect,
    given the level of the treatment t” (i.e. ATE(t)).  The routine also estimates other “causal” parameters of
    interest, such as the average treatment effect (ATE), the average treatment effect on treated (ATET), the
    average treatment effect on non-treated (ATENT), and the same effects conditional on t and on the vector of
    covariates x.The DRF is approximated by a third degree polynomial function.  Both OLS and IV estimation are
    available, according to the case in which the treatment is not or is endogenous. In particular, the implemented
    IV estimation is based on a Heckman bivariate selection model for w (the yes/no decision to treat a given unit)
    and t (the level of the treatment provided) in the first step, and a 2SLS estimation for the outcome (y)
    equation in the second step.  The routine allows also for a graphical representation of results.
```

Options

model(modeltype) specifies the treatment model to be estimated, where modeltype must be one of the following two models: "ct-ols", "ct-iv". it is always required to specify one model

ct(treat_level) specifies the treatment level (or dose). This variable takes values in the [0;100] interval, where 0 is the treatment level of non-treated units. The maximun dose is thus 100.

hetero(varlist_h) specifies the variables over which to calculate the idiosyncratic Average Treatment Effect ATE(x), ATET(x) and ATENT(x), where x=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 varlist_h should be the same set or a subset of the variables specified in varlist. Observe however that only numerical variables may be considered.

iv(instrument) specifies the variable to be used as instrument in the Heckman bivariate selection model. This option is required only for "ct-iv".

delta(number) identifies the average treatment effect between two states: t and t+delta. For any reliable delta, we can obtain the response function ATE(t;delta)=E[y(t)-y(t+delta)].

graphic allows for a graphical representation of the density distributions of ATE(x;t), ATET(x;t) and ATENT(x;t). It is optional for all models and gives an outcome only if variables into 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.

modeltype_options	description
Model	
ct-ols	Control-function regression estimated by ordinary least squares
ct-iv	IV regression estimated by Heckman bivariate selection model and 2SLS

ctreatreg creates a number of variables:

_ws_varname_h are the additional regressors used in model's regression when hetero(varlist_h) is specified.

_ps_varname_h are the additional instruments used in model's regression when hetero(varlist_h) is specified in model "ct-iv".

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

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

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

ATE(t) is an estimate of the Dose-Response-Function.

ATET(t) is the value of the Dose-Response-Function in t>0.

ATENT(t) it is the value of the Dose-Response-Function in t=0.

probw is the predicted probability from the Heckman selection model (estimated only for model "ct-iv").

mills is the predicted Mills' ratio from the Heckman selection model (estimated only for model "ct-iv").

t is a copy of the treatment level variable, but only in the sample considered.

t_hat is the prediction of the level of treatment from the Heckman bivariate selection model (estimated only for model "ct-iv").

der_ATE_t is the estimate of the derivative of the Dose-Response-Function.

std_ATE_t is the standardized value of the Dose-Response-Function.

std_der_ATE_t is the standardized value of the derivative of the Dose-Response-Function.

Tw, T2w, T3w are the three polynomial factors of the Dose-Response-Function.

T_hatp, T2_hatp, T3_hatp are the three instruments for the polynomial factors of the Dose-Response-Function when model "ct-iv" is used.

ctreatreg returns the following 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.

Remarks

The variable specified in 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.

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.

Following the Help-file this routine is rather straightforward to use and provides suitable graphical representation of results. In particular it provides a graph for the DRF and a combined graph for the densities of ATE(**x**,*t*), ATET(**x**,*t*) and ATENT(**x**,*t*).

References

- Hirano, K., and Imbens, G. (2004). The propensity score with continuous treatments. In Gelman, A. & Meng, X.L. (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (73-84). New York: Wiley.

- Cerulli (2012). “ivtreatreg: a new STATA routine for estimating binary treatment models with heterogeneous response to treatment under observable and unobservable selection”, CNR-Ceris Working Papers, No. 03/12. Available at: http://econpapers.repec.org/software/bocbocode/s457405.htm

- Bia, M., and Mattei, A. (2008). A Stata package for the estimation of the dose–response function through adjustment for the generalized propensity score, *The Stata Journal*, 8, 3, 354–373.