The Stata Journal (2014)
14, Number 1, pp. 141–158

Estimating the dose–response function through a generalized linear model approach

Barbara Guardabascio
Istat, Italian National Institute of Statistics
Rome, Italy
guardabascio@istat.it

Marco Ventura
Istat, Italian National Institute of Statistics
Rome, Italy
mventura@istat.it

Abstract. In this article, we revise the estimation of the dose–response function described in Hirano and Imbens (2004, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 73–84) by proposing a flexible way to estimate the generalized propensity score when the treatment variable is not necessarily normally distributed. We also provide a set of programs that accomplish this task. To do this, in the existing `doseresponse` program (Bia and Mattei, 2008, *Stata Journal* 8: 354–373), we substitute the maximum likelihood estimator in the first step of the computation with the more flexible generalized linear model.

Keywords: st0328, glmgpscore, glmdose, generalized propensity score, generalized linear model, dose–response, continuous treatment, bias removal

1 Introduction

How effective are policy programs with continuous treatment exposure? Answering this question essentially amounts to estimating a dose–response function as proposed in Hirano and Imbens (2004). Whenever doses are not randomly assigned but are given under experimental conditions, estimation of a dose–response function is possible using the generalized propensity score (GPS). The GPS for continuous treatment is an extension of the popular propensity-score methodology for binary treatment group assignments (Rosenbaum and Rubin 1983, 1984) and multivalued treatments (Imbens 2000; Lechner 2001). Indeed, Hirano and Imbens (2004) show that the GPS has a balancing property similar to the binary propensity score. Conditional on observable characteristics, the level of treatment can be considered random for units belonging to the same GPS strata. This means that adjusting for the GPS removes all biases associated with differences in the covariates.

Since its formulation, the GPS has been repeatedly used in observational studies, and programs have been provided for Stata users with `doseresponse.ado` and `gpscore.ado` by Bia and Mattei (2008), henceforth BM. However, many applied works (Fryges and Wagner 2008; Fryges 2009) remark that the treatment variable may not be

normally distributed. In this case, the BM programs should not be used, because they do not allow for distribution assumptions other than the normal density.

In this article, we overcome this problem. Building on BM programs, we provide a new set of Stata programs, `glmdose.ado` and `glmgpscore.ado`, that accommodates different distribution functions of the treatment variable. We accomplish this task in the first step by substituting the maximum likelihood (ML) estimator (from the existing program `doseresponse.ado`) with the generalized linear model (GLM).

To help compare the estimates—which we will present with those from the seminal work by Hirano and Imbens (2004)—we use the same dataset, originally collected by Imbens, Rubin, and Sacerdote (2001). The sample comprises individuals winning the Megabucks lottery in Massachusetts in the mid-1980s. The main source of potential bias is the unit and item nonresponse. Hirano and Imbens (2004) claim that it is possible to prove that the nonresponse was nonrandom. The missing data imply that the amount of the prize is potentially correlated with background characteristics and potential outcomes. We emphasize that by using these bias-reducing techniques, it is possible to reduce, but not to eliminate, the bias generated by unobservable heterogeneity. The extent to which unconfoundedness holds, namely, the extent to which the bias is reduced, depends on the quality of the database used to compute the GPS. This caveat is independent of the particular distribution function one is willing to assume for the treatment variable.

Notice also that the Stata command `teffects` estimates average treatment effects and average treatment effects on the treated by regression adjustment, inverse-probability weighting, and doubly robust methods, including inverse-probability weighted regression adjustment. This command is suitable for a binary or multinomial treatment variable. In contrast, `glmdose` does not impose this restriction on the treatment variable; in addition, it incorporates a test for the balancing not offered by `teffects`.

The remainder of the article proceeds as follows. Section 2 briefly reviews the estimation of the dose–response function. Section 3 introduces the GLM and explains how to use it to fit the GPS. Section 3.1 analyzes the fractional logit case, flogit, a special case of particular interest in economics. Section 4 describes how the programs work step by step. Sections 5 and 6 list the syntax and the options, respectively. Section 7 presents an application of the programs using a nonnormal distribution of the treatment variable. Section 8 concludes.

2 A brief review of the econometrics of the dose–response function

Let us define a set of potential outcomes $\{Y_i(t)\}$ for $t \in \mathcal{T}$, where $\mathcal{T}$ represents the continuous set of potential treatments defined over the interval $[t_0, t_1]$, and $Y_i(t)$ is referred to as the unit-level dose–response function.

Let us suppose we have a random sample of N units. For each unit i, we observe a $k \times 1$ vector of pretreatment covariates, X_i; the level of the treatment delivered, T_i; and the outcome corresponding to the level of the treatment received, $Y_i = Y_i(T_i)$. We are interested in the average dose–response function $\psi(t) = E\{Y_i(t)\}$.

Under some regularity conditions[1] of $\{Y_i(t)\}$, X_i, and T_i, Hirano and Imbens (2004) define the propensity function as the conditional density of the actual treatment given the covariates. In more detail, if we define $r(t, x) = f_{T|X}(t|x)$ as the conditional density function of the treatment given the covariates, then the GPS is

$$R = r(T|X)$$

The balancing property can be defined similarly to that defined by the binary case. That is, within strata with the same value of $r(t, x)$, the probability that $T = t$ does not depend on the value of X:

$$X \perp 1(T = t)|r(t, x)$$

This balancing property, along with unconfoundedness, implies that assignment to treatment is unconfounded given the GPS. If the weak unconfoundedness assumption holds, given the pretreatment variables X, we have

$$Y(t) \perp T|X \quad \forall \ t \in \mathcal{T}$$

Then for every t, we have

$$f_T\{t|r(t, X), Y(t)\} = f_T\{t|r(t, X)\}$$

This means that the GPS can be used to eliminate any bias associated with differences in the covariates (for a formal proof, see theorem 2.1 and 3.1 of Hirano and Imbens [2004]). Therefore, the dose–response function can be obtained as

$$\gamma(t, r) = E\{Y(t)|r(t, X) = r\} = E(Y|T = t, R = r) \quad (1)$$

$$\psi(t) = E[\gamma\{t, r(t, X)\}] \quad (2)$$

Practical implementation of the GPS is accomplished in three steps.[2]

In the first step, the score $r(t, x)$ is estimated. In the second step, the conditional expectation of the outcome as a function of two scalar variables, the treatment level T and the GPS R, is estimated: $E(Y|T = t, R = r)$. In the third step, the dose–response function, $\psi(t) = E[\{t, r(t, X)\}], t \in \mathcal{T}$, is estimated by averaging the estimated conditional expectation, $\widehat{\gamma}\{t, r(t, X)\}$, over the GPS at each level of the treatment in which one is interested.

1. For each i, $\{Y_i(t)\}$, X_i, and T_i are supposed to be defined on a common probability space; T_i is continuously distributed with respect to Lebesgue measure on $\mathcal{T}$; and $Y_i = Y_i(T_i)$ is a well-defined random variable.
2. In their seminal article, Hirano and Imbens (2004) use the notation μ instead of ψ and β instead of γ. We have slightly changed notation to avoid confusion in the following sections.

Because the second and the third steps in our programs replicate BM's program, we refer the reader to it for more details about these steps. Meanwhile, we will devote more attention to explaining how our programs implement the first step to compute the score $r(t, x)$.

3 Estimation of the score through the GLM

In many economic applications, T cannot be assumed to be normally distributed, and assuming a normal distribution of the treatment given the covariates $T_i|X_i \sim N(\beta' X_i, \sigma^2)$, where β is a $k \times 1$ vector of parameters, has several drawbacks. The problem is not new in the econometric literature; count, binomial, fractional, and survival data are a few examples (see Wooldridge [2010] for a comprehensive review of this topic). We aim to overcome these problems by presenting a possible solution to the estimation of the GPS in these cases. Our idea consists of replacing the linear regression[3] with the GLM developed by McCullagh and Nelder (1989) in the first step to estimate the dose–response and to retrieve the GPS from the exponential family distribution. By using the GLM, we differentiate the modeling from the ordinary regression in two important respects. First, the distribution of T is specified from the exponential family.[4] Thus the distribution may be explicitly nonnormal. Second, a nonidentity transformation of the mean of the treatment is linearly related to the explanatory variables. These two basic elements of the GLM can be formalized as follows:

$$f(T) = c(T, \phi)\exp\left\{\frac{T\theta - a(\theta)}{\phi}\right\} \tag{3}$$

$$g\{E(T)\} = \beta' X \tag{4}$$

Equation (3) specifies that the distribution of the treatment variable belongs to the exponential family. Equation (4) states that a transformation of the mean $g(\cdot)$ is linearly related to explanatory variables contained in X.

The choice of $a(\theta)$, commonly referred to as the family, is guided by the nature of the treatment variable. It determines the actual probability function, such as the binomial, Poisson, normal, gamma, inverse Gaussian, and negative binomial. Moreover, irrespective of the distribution chosen, the following relationships hold for the first and the second moment,

$$E(T) = \dot{a}(\theta), \qquad \mathrm{Var}(T) = \phi\ddot{a}(\theta)$$

where the dots represent the first and the second derivative with respect to θ.

The choice of $g(\cdot)$, a monotonic, differentiable function called a link function, is suggested by the functional form of the relationship between the treatment and the explanatory variables. It determines how the mean is related to the covariates X, while

3. Precisely, `gpscore` estimates the GPS assuming $T|X$ or some transformations of T, $g(T)|X$, normally distributed. The estimation of β is performed through the ML.
4. For a formal proof of how to obtain the different distributions from the exponential family, see de Jong and Heller (2008, chap. 3) or Rabe-Hesketh and Everitt (2007, chap. 7).

θ and ϕ represent the canonical parameter and the dispersion parameter, respectively. In this context, given X, μ is determined through $g(\mu)$. Given μ, θ is determined through $\dot{a}(\theta) = \mu$. Finally, given θ, T_i is determined as a draw from the exponential density specified in $a(\theta)$. When we compare our modeling with ordinary regression modeling, we clearly see that the extra steps are related to the choice of the `family()` and `link()` options: $a(\theta)$ and $g(\mu)$.

Indeed, by simply changing `family()` and `link()`, one can accommodate a very broad spectrum of distributions of T; however, not all combinations make sense (for a list of the feasible ones, we refer the reader to [SEM] **gsem family-and-link options**). In addition, and more importantly, Hirano and Imbens (2004) state: "In the first stage we use a normal distribution for the treatment given the covariates [...]. We may consider more general models such as mixtures of normals, or heteroskedastic normal distributions [...]". The GLM fully captures this point: because it allows T to be a member of the exponential family, the treatment can be heteroskedastic. Thus the variance will vary with the mean, which in turn varies with explanatory variables.

The GLM allows one to estimate β by maximizing the following quasi-maximum log likelihood (QML) for T_i independently distributed:

$$l(\beta) \equiv \sum_{i=1}^{N} l_i(\beta) \equiv \sum_{i=1}^{N} \log f(T_i;\beta) = \sum_{i=1}^{N} \left\{ \log c(T_i,\phi) + \frac{T_i\theta_i - a(\theta_i)}{\phi} \right\} \tag{5}$$

Because the GPS is the conditional density of the treatment received given the covariates, we can compute the GPS by using the exponential density function evaluated at $\widehat{\beta}$, given the covariates

$$R = r(T,X) = f\left(\widehat{\beta}\right)$$

where f is according to (3). Put another way, the GPS coincides with the vector of the likelihood evaluated at $\widehat{\beta}$, $L(\widehat{\beta})$, where $L(\widehat{\beta}) = \exp\{l(\widehat{\beta})\}$.

However, whenever T is discrete or fractional, a clarification is in order. In these cases, the ML in (5) is replaced by the Bernoulli QML, as seen in (6):

$$l^B(\beta) \equiv \sum_{i=1}^{N} l_i^B(\beta) = \sum_{i=1}^{N} T_i \log\{F(T_i;\beta)\} + (1-T_i)\log\{1-F(T_i;\beta)\} \tag{6}$$

If T is binary and (6) is estimated by setting binomial as family and logit (or probit) as link, (6) exactly reproduces the case of binary treatment. In this case, the probability of being assigned to treatment—that is, the p-score—is $F(T=1)$, which is the cumulative logit (or probit) evaluated at $\widehat{\beta}'X$ for $T=1$. By definition, this is not the cumulative logit (or probit) evaluated at the actual level of the treatment received, which can be either 0 or 1. Starting from this consideration, we extend this argument from the binary to the fractional case. Because a great part of the empirical literature has come across the necessity to estimate a dose–response function with fractional treatment data (Fryges and Wagner 2008; Fryges 2009), we believe this case deserves special attention. Thus we will treat it in more detail in the following subsection.

3.1 Flogit or fractional treatment data: A case of particular interest

In economics, it is quite common to come across a fractional dependent variable, in our setup, $T \in [0, 1]$. Some examples include the fraction of income contributed to charity, the fraction of weekly hours spent working, the proportion of total firm capitalization accounted for by debt capital, high school graduation rates, and export sales ratio. (See Hausman and Leonard [1997]; Liu et al. [1999]; Wagner [2001]; Fryges and Wagner [2008]; and Fryges [2009].) Papke and Wooldridge (1996) show that the problems of linear models for fractional data are analogous to those of the linear probability model for binary data. Thus if T is bounded, the effect of any particular covariate in X_i cannot be constant over its range. Augmenting the model with nonlinear functions of X_i does not overcome the problem, because the values from an ordinary least-squares regression can never be guaranteed to lie in the unit interval.

The common practice of regressing the log odds-ratio, that is, $\log\{T/(1-T)\}$ in the linear regression instead of T, generates problems whenever any observation T_i takes on the values 0 or 1 with positive probability. As a practice, when T_i are proportions from a fixed number of groups with known group size, the extreme values are adjusted before taking the transformation. However, the fraction T_i is not always a proportion from a discrete group size. In addition, if a large percentage is at the extremes, the adjustment mechanism is at least debatable. Papke and Wooldridge (1996) sidestep these problems by specifying a class of functional forms for $E(T|X)$ and show how to estimate the parameters using a Bernoulli QML estimator of β, namely, the GLM. In particular, they assume that for all i,

$$E(T_i|X_i) = F(\beta' X_i)$$

where $F(\cdot)$ is typically a logit or probit function, from here the name of flogit estimator.[5]

Analogously to the binary case, the estimation procedure defines the Bernoulli log-likelihood function as

$$l_i(\beta) \equiv T_i \log\{F(\beta' X_i)\} + (1 - T_i) \log\{1 - F(\beta' X_i)\} \tag{7}$$

and maximizes the sum of $l_i(\beta)$ over all N using the GLM. Because the GPS is the probability of the actual (that is, the observed) treatment received, $L_i^B(\beta)$ does not coincide with the GPS.[6] $\{1 - F(\beta' X_i)\}$ attains the probability of receiving $T = 1 - t$, which is not the actual treatment, that is, the observed one, but its complement. Hence, it must not enter the `gpscore()`. The estimated GPS based on the Bernoulli log-likelihood function in (7) is

$$R_i = F\left(\widehat{\beta}' X_i\right) \qquad \forall \ i$$

In this respect, the GPS and the p-score are computable exactly in the same way whenever the likelihood is Bernoulli.

5. Notice that in the notation of, for instance, (4), $F = g^{-1}$, if $g(\cdot)$ is the log-odds or logit transformation, $g(\mu) = \log\{\mu/(1-\mu)\}$, $F = \exp(\mu)/\{1 + \exp(\mu)\}$; that is, $F = \Lambda$, the logit distribution.
6. See Wooldridge (2010, 739–743).

Therefore, as a general rule, we can state that by using the GLM in the first step of the dose–response function to retrieve the GPS, one must

- take $L(\widehat{\beta})$ whenever the QML is not Bernoulli.
- take $F(\widehat{\beta}'X_i)$ whenever the QML is Bernoulli, where $F(\cdot)$ is the probability of succeeding, that is, of being assigned to treatment t. That is exactly what our programs automatically implement.[7]

4 The estimation algorithm

The implementation method can be broken down into three steps. In the first step, the program `glmgpscore.ado` estimates the GPS and tests the balancing properties for any family and link set. In the second step, the conditional expectation of the outcome is estimated as a function of the treatment level T and the GPS $R, \gamma(t,r) = E(Y|T = t, R = r)$. Finally, in the third step, the dose–response function, $\psi(t) = E[\gamma\{t, r(t,X)\}]$, is estimated by averaging the estimated conditional expectation, $\widehat{\gamma}\{t, r(t,X)\}$, over the GPS at each level of the treatment in which the user is interested.

In detail, the first step is implemented as follows:

1. Estimate the parameters θ and ϕ of the selected conditional distribution of the treatment given the covariates. Indeed, the distribution of T is specified from the exponential family through the `family()` and `link()` options.
2. If the family selected is normal, assess the validity of the assumed normal distribution model by one of the following user-specified goodness-of-fit tests: the Kolmogorov–Smirnov, the Shapiro–Francia, the Shapiro–Wilk, or the Stata skewness and kurtosis test for normality. The user can skip the test by specifying the `flag_b(2)` option. If the normal distribution model is not statistically supported, inform the user that the assumption of normality is not satisfied. The user is invited to use different `family()` and `link()` options or a different transformation of the treatment variable.
3. Estimate the GPS as

$$\widehat{R_i} = r(T,X) = c\left(T, \widehat{\phi}\right) \exp\left\{\frac{T\widehat{\theta} - a\left(\widehat{\theta}\right)}{\widehat{\phi}}\right\}$$

   where $\widehat{\theta}$ and $\widehat{\phi}$ are the estimated parameters in step 1.

7. The authors wish to thank K. Hirano for helping on this point in a private conversation. In contrast with our approach in a Bernoulli QML, Fryges and Wagner (2008) and Fryges (2009) take $L_i^B(\widehat{\beta})$.

4. Test the balancing property, and inform the user whether it is supported by the data and, if so, to what extent. Following Hirano and Imbens (2004), the program `glmgpscore.ado` tests for balancing of covariates according to the following scheme:

    a. Divide the sample into k groups according to a user-specified rule, which should be defined on the basis of the sample distribution of the treatment variable.

    b. In the first group, $k = 1$, compute the GPS at the user-specified representative point. For instance, compute the median of the group, and evaluate the GPS for each individual in the sample by setting $t =$ median of the group.[8]

    c. Take the GPS obtained in step b, and divide it into nq subintervals defined by its quantiles of order j/nq, $j = 1, \ldots, nq - 1$. We refer to these subintervals as blocks.

    d. Within each block, compare individuals who are treated—that is, who belong to group k (according to step a)—with individuals who are in the same block but belong to another group. Specifically, within each block, calculate the mean difference of each covariate between units belonging to group k and units not belonging to group k.

    e. Combine the nq mean differences, calculated in step d using a weighted average, with weights given by the number of observations in each GPS block.

    f. Go to step b, set $k = 2$, and go through steps b–e.

For each group, test statistics (the Student's t statistics or the Bayes factors) are calculated and shown in the Results window. Finally, the most extreme value of the test statistics (the highest absolute value of the Student's t statistics or the lowest value of the Bayes factors) is compared with reference values, and the user is told to what extent the balancing property is supported by the data. If adjustment for the GPS properly balances the covariates, we would expect all differences to be statistically not significant.

Notice that for binary treatments, although the GPS is correctly calculated, the dose–response function reduces to a point rather than a curve. For this standard case, we refer the user to `pscore.ado` by Becker and Ichino (2002) and to `psmatch2.ado` by Leuven and Sianesi (2003).[9]

8. Notice that this will generate a distribution of the GPS with N elements for each group.
9. When the family is binomial, the balancing mechanism is slightly different. Indeed, in this case, the GPS is independent of t because $r(t, x) = F(\beta' x)$. Therefore, going through step b, the algorithm will generate k times the same GPS vector. It means that step f becomes ineffective because the GPS does not change by changing the representative point of t.

In the second stage, the conditional expectation for the outcome Y_i, given T_i and R_i, is modeled as a flexible function of its two arguments. We use polynomial approximations of order not higher than three. Specifically, the most complex model we consider is

$$\begin{aligned}\varphi\{E(Y_i|T_i,R_i)\} &= \lambda(T_i,R_i;\alpha)\\ &= \alpha_0+\alpha_1T_i+\alpha_2T_i^2+\alpha_3T_i^3+\alpha_4R_i+\alpha_5R_i^2+\alpha_6R_i^3+\alpha_7T_iR_i\end{aligned}$$

where $\varphi(\cdot)$ is a function that relates the predictor, $\lambda(T_i,R_i;\alpha)$, to the conditional expectation $E(Y_i|T_i,R_i)$.

The last step consists of averaging the estimated regression function over the score function evaluated at the desired level of the treatment. Specifically, to obtain an estimate of the entire dose–response function, the program estimates the average potential outcome for each level of the treatments in which one is interested by applying the empirical counterpart of (1) and (2); that is,

$$\widehat{E\{Y(t)\}}=\frac{1}{N}\sum_{i=1}^{N}\widehat{\gamma}\left\{t,\widehat{r}(t,X_i)\right\}=\frac{1}{N}\sum_{i=1}^{N}\varphi^{-1}\left[\widehat{\lambda}\left\{t,\widehat{r}(t,X_i);\widehat{\alpha}\right\}\right]$$

Briefly, the program `glmdose.ado` estimates the dose–response function according to the following algorithm:

1. Estimate the GPS (according to the family and link specified by the user) through the GLM approach; check the normality, if required; and test the balancing property by using the routine `glmgpscore.ado`.
2. Estimate the conditional expectation of the outcome, given the treatment and the GPS, by calling the routine `doseresponse_model.ado`.[10]
3. Estimate the average potential outcome for each level of the treatment in which the user is interested.
4. Estimate the standard errors of the dose–response function via bootstrapping.[11]
5. Plot the estimated dose–response function and, if requested, its confidence intervals.

10. The `doseresponse_model` command (Bia and Mattei 2008) is required by the `glmdose` command. Type `search doseresponse_model`, and follow the links to install the latest package.
11. As in `glmdose.ado`, when bootstrapped standard errors are required, the bootstrap encompasses both the estimation of the GPS based on the specification given by the user and the estimation of the α parameter.

5 Syntax

This section presents the syntax of the programs, reporting all the potential options. The next section reports only the details of the options specific to `glmgpscore` and `glmdose` and refers the reader to `gpscore` and `doseresponse` (Bia and Mattei 2008) for the options in common.[12]

`glmgpscore` *varlist* [*if*] [*in*] [*weight*], `t(`*varname*`)` `gpscore(`*newvar*`)`
`predict(`*newvar*`)` `sigma(`*newvar*`)` `cutpoints(`*varname*`)` `index(`*string*`)`
`nq_gps(`#`)` `family(`*string*`)` `link(`*string*`)` [`t_transf(`*transformation*`)`
`normal_test(`*test*`)` `norm_level(`#`)` `test_varlist(`*varlist*`)` `test(`*type*`)`
`flag_b(`#`)` `opt_nb(`*string*`)` `opt_b(`*varname*`)` `detail`]

`glmdose` *varlist* [*if*] [*in*] [*weight*], `outcome(`*varname*`)` `t(`*varname*`)`
`gpscore(`*newvar*`)` `predict(`*newvar*`)` `sigma(`*newvar*`)` `cutpoints(`*varname*`)`
`index(`*string*`)` `nq_gps(`#`)` `dose_response(`*newvar*`)` `family(`*string*`)`
`link(`*string*`)` [`t_transf(`*transformation*`)` `normal_test(`*test*`)` `norm_level(`#`)`
`test_varlist(`*varlist*`)` `test(`*type*`)` `flag_b(`#`)` `cmd(`*regression_cmd*`)`
`reg_type_t(`*type*`)` `reg_type_gps(`*type*`)` `interaction(`#`)` `tpoints(`*vector*`)`
`npoints(`#`)` `delta(`#`)` `filename(`*filename*`)` `bootstrap(`*string*`)` `boot_reps(`#`)`
`analysis(`*string*`)` `analysis_level(`#`)` `graph(`*filename*`)` `opt_nb(`*string*`)`
`opt_b(`*varname*`)` `detail`]

Note that in the commands `glmgpscore` and `glmdose`, the argument *varlist* represents the control variables, which are used to estimate the GPS.

6 Options

The `glmdose` options include all the `doseresponse` options and some others strictly related to the GLM estimator. In what follows, we provide only a description of the options related to the `glmdose` command and not included in `doseresponse` or with a different content, referring the reader to BM for the others. In addition, we recall that the `glmdose` options include all the options for the `glmgpscore` command.

12. Type `search doseresponse` or `search gpscore`, and follow the links to install the latest package.

6.1 Required

`gpscore(`*newvar*`)` specifies the variable name for the estimated GPS via GLM.

`sigma(`*newvar*`)` creates a new variable containing the GLM fit of the conditional standard error of the treatment given the covariates, which are obtained from Pearson residuals.[13]

`family(`*string*`)` specifies the distribution family name of the treated variable.

`link(`*string*`)` specifies the link function for the treated variable. The default is the canonical link for the `family()` specified.[14]

6.2 Optional

`flag_b(#)` skips either the balancing or the normal test or both and takes as arguments `0`, `1`, or `2`. If `flag_b()` is not specified, the program estimates the GPS performing both the balancing and the normal tests. `flag_b(0)` skips both the balancing and the normal tests; `flag_b(1)` skips only the balancing test; `flag_b(2)` skips only the normal test.

`opt_nb(`*string*`)` specifies the negative binomial dispersion parameter. In the GLM approach, you specify `family(nb` $\#k$`)`, where $\#k$ is specified through the `opt_nb()` option. The GLM then searches for the $\#k$ that results in the deviance-based dispersion being 1. Instead, `nbreg` finds the ML estimate of $\#k$.

`opt_b(`*varname*`)` specifies the name of the variable that contains the number of binomial trials.

7 Stata output

We illustrate the details of our programs using the dataset collected by Imbens, Rubin, and Sacerdote (2001). Our choice of the dataset has been motivated by the need of comparing our results with those of Hirano and Imbens (2004). The aim of the original exercise was to estimate the effect of the Lottery prize amount on subsequent labor earnings, `year6`. Because our econometric exercise is simply motivated by the need to show the functioning of the programs, we have considered a different treatment variable (different from `prize`) that allows us to use the binary distribution function. In particular, the flogit case has been implemented by using the treatment variable `fraction`, which is obtained by normalizing the variable `prize` with respect to its highest value in the sample. Accordingly, the results of `glmgpscore` and `glmdose` are shown below.

13. The authors wish to thank J. Wooldridge for helping on this point in a private conversation. Recall that in the case of normal distribution, Pearson residuals coincide with usual residuals.
14. For the list of all the admissible family–link combinations, see [SEM] **gsem family-and-link options**.

7.1 Flogit glmgpscore output

In this case, the treatment variable is `fraction`, which by construction takes on values in the unit interval. The code is implemented by setting the cutpoints to divide the sample into three groups contained in the variable `cut1`. The link function is the canonical one, logit; however, other links are admissible. The output appears as follows:

```
. use lotterydataset
. egen max_p=max(prize)
. generate fraction= prize/max_p
. quietly generate cut1 = 23/max_p  if fraction<=23/max_p
. quietly replace cut1 = 80/max_p  if fraction>23/max_p & fraction<=80/max_p
. quietly replace cut1 = 485/max_p if fraction >80/max_p
. #delimit ;
delimiter now ;
. glmgpscore male ownhs owncoll tixbot workthen yearw yearm1 yearm2,
> t(fraction) gpscore(gpscore_fr) predict(y_hat_fr) sigma(sd_fr)
> cutpoints(cut1) index(mean) nq_gps(5) family(binomial) link(logit) detail
> ;
Generalized Propensity Score
**************************************************
Algorithm to estimate the generalized propensity score
**************************************************
Estimation of the propensity score
The treatment is fraction
                            T
-------------------------------------------------------------

      Percentiles      Smallest
 1%     .0103137       .0023495
 5%     .0202446       .0023495
10%     .0231977       .0103137       Obs                 237
25%     .0351369       .0110477       Sum of Wgt.         237
50%     .0654881                      Mean           .1138546
                        Largest       Std. Dev.       .127485
75%     .1299367       .5571485
90%      .270282        .629324       Variance       .0162524
95%     .3482539       .6669279       Skewness       2.888956
99%      .629324              1       Kurtosis       15.08626
note: T has noninteger values
Generalized linear models                         No. of obs      =       237
Optimization     : ML                             Residual df     =       228
                                                  Scale parameter =         1
Deviance         =  25.91237504                   (1/df) Deviance =  .1136508
Pearson          =  29.27315861                   (1/df) Pearson  =   .128391
Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = ln(u/(1-u))             [Logit]
                                                  AIC             =  .6036733
Log pseudolikelihood = -62.53528122               BIC             = -1220.805
```

T	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.6402121	.1694826	3.78	0.000	.3080323	.9723918
ownhs	-.1515907	.1086591	-1.40	0.163	-.3645586	.0613773
owncoll	.0401978	.0431132	0.93	0.351	-.0443026	.1246982
tixbot	.0202427	.0249659	0.81	0.417	-.0286895	.0691749
workthen	.1558366	.2139876	0.73	0.466	-.2635714	.5752446
yearw	-.0169543	.0603052	-0.28	0.779	-.1351503	.1012416
yearm1	-.0055257	.0131275	-0.42	0.674	-.0312552	.0202037
yearm2	.0089422	.0134262	0.67	0.505	-.0173726	.035257
_cons	-2.146518	.5413156	-3.97	0.000	-3.207477	-1.085559

robust standard errors reported

Estimated generalized propensity score

Percentiles	Smallest		
1% .0556678	.0537445		
5% .0600808	.0547833		
10% .0659206	.0556678	Obs	237
25% .0749973	.0563906	Sum of Wgt.	237
50% .1254999		Mean	.1138546
	Largest	Std. Dev.	.0387714
75% .1413647	.217338		
90% .1541515	.2175611	Variance	.0015032
95% .167948	.2198188	Skewness	.2804054
99% .2175611	.2256652	Kurtosis	2.511468

```
*******************************************
End of the algorithm to estimate the gpscore
*******************************************
```

```
*******************************************************************************
The set of the potential treatment values is divided into 3 intervals
The values of the gpscore evaluated at the representative point of each
treatment interval are divided into 5 intervals
*******************************************************************************
```

```
*********************************************************
Summary statistics of the distribution of the GPS evaluated
at the representative point of each treatment interval
*********************************************************
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gps_1	237	.1138546	.0387714	.0537445	.2256652
Variable	Obs	Mean	Std. Dev.	Min	Max
gps_2	237	.1138546	.0387714	.0537445	.2256652
Variable	Obs	Mean	Std. Dev.	Min	Max
gps_3	237	.1138546	.0387714	.0537445	.2256652

```
*******************************************************************************
Test that the conditional mean of the pre-treatment variables given the generalized
propensity score is not different between units who belong to a particular
treatment interval and units who belong to all other treatment intervals
*******************************************************************************
```

```
Treatment Interval No 1 - [.0023494709748775, .0474060922861099]
              Mean        Standard
              Difference  Deviation   t-value
        male   .07032       .03214     2.1881
       ownhs   .27061       .13368     2.0244
     owncoll   .14939       .21863      .6833
      tixbot   .09136       .43645     .20931
    workthen  -.01029       .05015    -.20523
       yearw   .15477       .18022     .85879
      yearm1   1.4991       1.7217      .8707
      yearm2    1.823       1.5597     1.1688
Treatment Interval No 2 - [.0476247407495975, .1631902456283569]
              Mean        Standard
              Difference  Deviation   t-value
        male  -.06435       .02183    -2.9477
       ownhs  -.13305       .13008    -1.0228
     owncoll  -.18433       .19743    -.93368
      tixbot  -.48247       .38721     -1.246
    workthen  -.00199       .04998     -.0398
       yearw  -.33553        .1666     -2.014
      yearm1   .07426       1.6071     .04621
      yearm2  -.09833       1.4601    -.06734
Treatment Interval No 3 - [.1711813360452652, 1]
              Mean        Standard
              Difference  Deviation   t-value
        male  -.01669       .03175    -.52566
       ownhs   .19524       .17768     1.0988
     owncoll   .18711       .27456     .68148
      tixbot   .47912       .50744     .94421
    workthen  -.05865       .07293    -.80421
       yearw   .23415       .22407      1.045
      yearm1  -.70637        1.966    -.35929
      yearm2  -1.1814       1.7682    -.66816

According to a standard two-sided t test:
Decisive evidence against the balancing property
The balancing property is satisfied at a level lower than 0.01
. #delimit cr
delimiter now cr
```

7.2 Flogit glmdose output

The `glmgpscore` command is replaced by `glmdose` and additional options are added. Specifically, the matrix `tp1` contains the value of the treatment we are interested in. See figure 1.

```
. use lotterydataset.dta, clear
. egen max_p=max(prize)
. generate fraction= prize/max_p
. quietly generate cut1 = 23/max_p  if fraction<=23/max_p
. quietly replace cut1 = 80/max_p  if fraction>23/max_p & fraction<=80/max_p
. quietly replace cut1 = 485/max_p if fraction >80/max_p
. mat def tp1 = (0.10\0.20\0.30\0.40\0.50\0.60\0.70\0.80)
. #delimit ;
delimiter now ;
. glmdose male ownhs owncoll tixbot workthen yearw yearm1 yearm2,
> t(fraction) gpscore(gps_flog) predict(y_hat_fl) sigma(sd_fl)
> cutpoints(cut1) index(mean) nq_gps(5) family(binomial) link(logit)
> outcome(year6) dose_response(doseresp_fl) tpoints(tp1) delta(0.1)
> reg_type_t(quadratic) reg_type_gps(quadratic) interaction(1)
> bootstrap(yes) boot_reps(10) analysis(yes) detail
> filename("output_flog") graph("graphflog.eps")
> ;
*******************************************
ESTIMATE OF THE GENERALIZED PROPENSITY SCORE
*******************************************
Generalized Propensity Score
*****************************************************
Algorithm to estimate the generalized propensity score
*****************************************************
  (output omitted)
The outcome variable ´´year6´´ is a continuous variable
The regression model is: Y = T + T^2 + GPS +  GPS^2 + T*GPS
```

Source	SS	df	MS		
Model	4.2029e+09	5	840589784	Number of obs =	202
				F(5, 196) =	4.44
Residual	3.7122e+10	196	189397662	Prob > F =	0.0007
				R-squared =	0.1017
				Adj R-squared =	0.0788
Total	4.1325e+10	201	205596471	Root MSE =	13762

year6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fraction	-63135.37	30152.68	-2.09	0.038	-122600.7	-3670.024
fraction_sq	9555.672	40829.3	0.23	0.815	-70965.47	90076.82
gps_flog	297627.5	137193.5	2.17	0.031	27062.67	568192.4
gps_flog_sq	-931930.1	571320.7	-1.63	0.104	-2058655	194795
fraction_gps_flog	201989.2	290293.7	0.70	0.487	-370510.9	774489.3
_cons	-4979.084	7733.942	-0.64	0.520	-20231.51	10273.34

```
Bootstrapping of the standard errors
..........

The program is drawing graphs of the output
This operation may take a while

(note: file graphflog.eps not found)
(file graphflog.eps saved)

End of the Algorithm

. #delimit cr
delimiter now cr
```

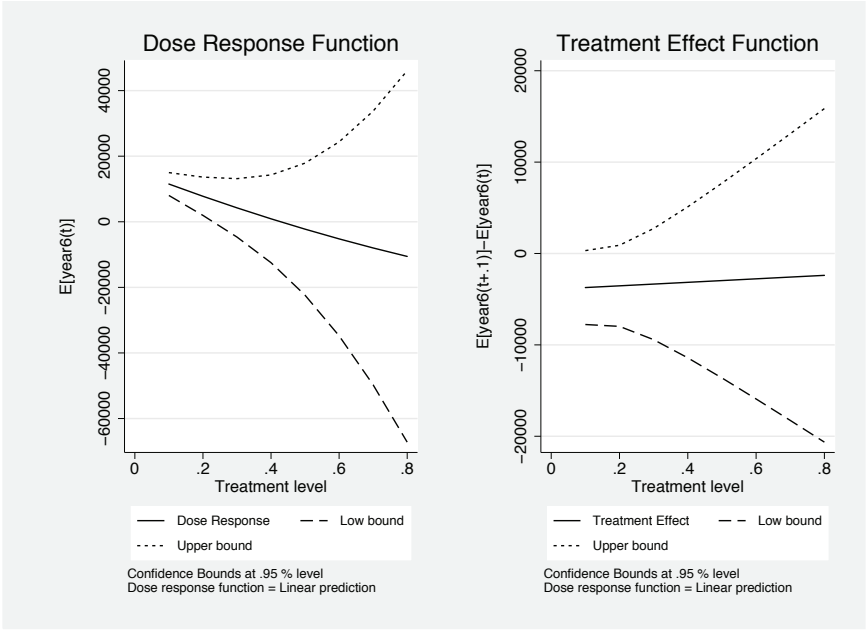

Figure 1. Estimated dose–response function, estimated derivative, and 95% confidence bands for binomial distributed data

8 Conclusions

In recent years, there has been growing interest in the evaluation of policy interventions and, more generally, in the estimation of causal effects. To accomplish these tasks, researchers need ad hoc software and programs. In this article, we provided two Stata programs implementing the GPS in a general setup. The programs are very versatile thanks to the introduction of the GLM estimator in the first step of the estimation of the GPS.

9 Acknowledgments

The article benefited from the useful comments and suggestions of many people. In particular, the authors wish to gratefully acknowledge E. Battistin, K. Hirano, and J. Wooldridge. The opinions expressed by the authors are theirs only and do not necessarily reflect the position of the Institute.

10 References

Becker, S. O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2: 358–377.

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

de Jong, P., and G. Z. Heller. 2008. *Generalized Linear Models for Insurance Data*. New York: Cambridge University Press.

Fryges, H. 2009. The export–growth relationship: Estimating a dose-response function. *Applied Economics Letters* 16: 1855–1859.

Fryges, H., and J. Wagner. 2008. Exports and productivity growth: First evidence from a continuous treatment approach. *Review of World Economics* 144: 695–722.

Hausman, J. A., and G. K. Leonard. 1997. Superstars in the national basketball association: Economic value and policy. *Journal of Labor Economics* 15: 586–624.

Hirano, K., and G. W. Imbens. 2004. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A. Gelman and X.-L. Meng, 73–84. Chichester, UK: Wiley.

Imbens, G. W. 2000. The role of the propensity score in estimating dose–response functions. *Biometrika* 87: 706–710.

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote. 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review* 91: 778–794.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, ed. M. Lechner and F. Pfeiffer, 43–58. Heidelberg: Physica-Verlag.

Leuven, E., and B. Sianesi. 2003. psmatch2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s432001.html.

Liu, J.-L., J.-T. Liu, J. K. Hammitt, and S.-Y. Chou. 1999. The price elasticity of opium in Taiwan, 1914–1942. *Journal of Health Economics* 18: 795–810.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall/CRC.

Papke, L. E., and J. M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* 11: 619–632.

Rabe-Hesketh, S., and B. Everitt. 2007. *A Handbook of Statistical Analyses Using Stata*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

———. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–524.

Wagner, J. 2001. A note on the firm size—Export relationship. *Small Business Economics* 17: 229–237.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the authors

Barbara Guardabascio received a PhD in econometrics and empirical economics from the University of Rome Tor Vergata. Currently, she is a researcher at the Italian National Institute of Statistics. Her primary research interests include forecasting, business cycle analysis, and policy evaluation. She is the coauthor, with G. Cubadda, of "On the use of PLS regression for forecasting large sets of cointegrated time series" and "A medium-N approach to macroeconomic forecasting". She is also the coauthor, with G. Cubadda and A. Hecq, of "A general to specific approach for selecting the best business cycle indicators" and "Building a synchronous common-cycle index for the European Union".

Marco Ventura received a PhD in economics from the Faculty of Statistics at the University of Rome La Sapienza. He is a researcher at the Italian National Institute of Statistics, and his primary research interests are the estimation of causal effects, the evaluation of public policies, and innovation. He has coauthored articles published in international journals.