The Stata Journal

Editors
H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, Texas
editors@stata-journal.com

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
editors@stata-journal.com

Associate Editors
Christopher F. Baum, Boston College
Nathaniel Beck, New York University
Rino Bellocco, Karolinska Institutet, Sweden, and University of Milano-Bicocca, Italy
Maarten L. Buis, WZB, Germany
A. Colin Cameron, University of California–Davis
Mario A. Cleves, University of Arkansas for Medical Sciences
William D. Dupont, Vanderbilt University
Philip Ender, University of California–Los Angeles
David Epstein, Columbia University
Allan Gregory, Queen's University
James Hardin, University of South Carolina
Ben Jann, University of Bern, Switzerland
Stephen Jenkins, London School of Economics and Political Science
Ulrich Kohler, University of Potsdam, Germany
Frauke Kreuter, Univ. of Maryland–College Park
Peter A. Lachenbruch, Oregon State University
Jens Lauritsen, Odense University Hospital
Stanley Lemeshow, Ohio State University
J. Scott Long, Indiana University
Roger Newson, Imperial College, London
Austin Nichols, Urban Institute, Washington DC
Marcello Pagano, Harvard School of Public Health
Sophia Rabe-Hesketh, Univ. of California–Berkeley
J. Patrick Royston, MRC Clinical Trials Unit, London
Philip Ryan, University of Adelaide
Mark E. Schaffer, Heriot-Watt Univ., Edinburgh
Jeroen Weesie, Utrecht University
Ian White, MRC Biostatistics Unit, Cambridge
Nicholas J. G. Winter, University of Virginia
Jeffrey Wooldridge, Michigan State University

Stata Press Editorial Manager
Lisa Gilmore

Stata Press Copy Editors
David Culwell and Deirdre Skaggs

The Stata Journal publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go “beyond the Stata manual” in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The Stata Journal is indexed and abstracted by CompuMath Citation Index, Current Contents/Social and Behavioral Sciences, RePEc: Research Papers in Economics, Science Citation Index Expanded (also known as SciSearch), Scopus, and Social Sciences Citation Index.

For more information on the Stata Journal, including information for authors, see the webpage

http://www.stata-journal.com
**Subscriptions** are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at


**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

<table>
<thead>
<tr>
<th></th>
<th>U.S. and Canada</th>
<th>Elsewhere</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Printed &amp; electronic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-year subscription</td>
<td>$98</td>
<td>$138</td>
</tr>
<tr>
<td>2-year subscription</td>
<td>$165</td>
<td>$245</td>
</tr>
<tr>
<td>3-year subscription</td>
<td>$225</td>
<td>$345</td>
</tr>
<tr>
<td>1-year student subscription</td>
<td>$75</td>
<td>$99</td>
</tr>
<tr>
<td>1-year institutional subscription</td>
<td>$245</td>
<td>$285</td>
</tr>
<tr>
<td>2-year institutional subscription</td>
<td>$445</td>
<td>$525</td>
</tr>
<tr>
<td>3-year institutional subscription</td>
<td>$645</td>
<td>$765</td>
</tr>
<tr>
<td><strong>Electronic only</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-year subscription</td>
<td>$75</td>
<td>$75</td>
</tr>
<tr>
<td>2-year subscription</td>
<td>$125</td>
<td>$125</td>
</tr>
<tr>
<td>3-year subscription</td>
<td>$165</td>
<td>$165</td>
</tr>
<tr>
<td>1-year student subscription</td>
<td>$45</td>
<td>$45</td>
</tr>
</tbody>
</table>

Back issues of the *Stata Journal* may be ordered online at


Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.


The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

[Copyright © 2014 by StataCorp LP](http://www.stata.com/copyright.html)

**Copyright Statement:** The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, *Stata*, Stata Press, Mata, *mata*, and NetCourse are registered trademarks of StataCorp LP.
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 2008) remark that the treatment variable may not be
Estimating the dose–response function

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 re-
duced, 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-prob-
ability 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 esti-
mation 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 \( t \in T \), where \( T \) 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 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

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

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

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

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

Practical implementation of the GPS is accomplished in three steps:

1. 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, $\psi(t) = E[\gamma\{t, r(t, X)\}]$, $t \in T$, is estimated by averaging the estimated conditional expectation, $\hat{\gamma}\{t, r(t, X)\}$, 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 $T$; 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 $\mu$ instead of $\psi$ and $\beta$ instead of $\gamma$. We have slightly changed notation to avoid confusion in the following sections.
Estimating the dose–response function

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 \( \beta \) 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. 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:

\[
\begin{align*}
  f(T) &= c(T, \phi) \exp \left\{ \frac{T\theta - a(\theta)}{\phi} \right\} \quad (3) \\
  g\{E(T)\} &= \beta'X \quad (4)
\end{align*}
\]

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,

\[
E(T) = \dot{a}(\theta), \quad \text{Var}(T) = \phi\ddot{a}(\theta)
\]

where the dots represent the first and the second derivative with respect to \( \theta \).

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, \texttt{gpscore} estimates the GPS assuming \( T|X \) or some transformations of \( T \), \( g(T)|X \), normally distributed. The estimation of \( \beta \) 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).
\( \theta \) and \( \phi \) represent the canonical parameter and the dispersion parameter, respectively. In this context, given \( X \), \( \mu \) is determined through \( g(\mu) \). Given \( \mu \), \( \theta \) is determined through \( \dot{a}(\theta) = \mu \). Finally, given \( \theta \), \( T_i \) 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 \texttt{family()} and \texttt{link()} options: \( a(\theta) \) and \( g(\mu) \).

Indeed, by simply changing \texttt{family()} and \texttt{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 \texttt{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 [...]”. The \texttt{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 \texttt{GLM} allows one to estimate \( \beta \) by maximizing the following quasi-maximum log likelihood (QML) for \( T_i \) independently distributed:

\[
\begin{align*}
\ell(\beta) & \equiv \sum_{i=1}^{N} l_i(\beta) = \sum_{i=1}^{N} \log f(T_i; \beta) = \sum_{i=1}^{N} \left\{ \log c(T_i, \phi) + \frac{T_i \theta_i - a(\theta_i)}{\phi} \right\} \\
\end{align*}
\]

(5)

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

\[
R = r(T, X) = f(\hat{\beta})
\]

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

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

\[
\begin{align*}
\ell^B(\beta) & \equiv \sum_{i=1}^{N} l_i^B(\beta) = \sum_{i=1}^{N} T_i \log\{F(T_i; \beta)\} + (1 - T_i) \log\{1 - F(T_i; \beta)\} \\
\end{align*}
\]

(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 \texttt{p-score}—is \( F(T = 1) \), which is the cumulative logit (or probit) evaluated at \( \hat{\beta} X \) 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 \( \beta \), 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.

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

\[
l_i(\beta) \equiv T_i \log \{ F(\beta'X_i) \} + (1 - T_i) \log \{ 1 - F(\beta'X_i) \}
\]

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 \( \{ 1 - F(\beta'X_i) \} \) 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 \( \text{gp score}() \). The estimated GPS based on the Bernoulli log-likelihood function in (7) is

\[
R_i = F(\hat{\beta}'X_i) \quad \forall \ i
\]

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.

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 $L(\hat{\beta})$ whenever the QML is not Bernoulli.
- take $F(\hat{\beta} X_i)$ 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.

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, $\hat{\gamma}\{t, r(t, X)\}$, 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 $\theta$ and $\phi$ 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

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

where $\hat{\theta}$ and $\hat{\phi}$ 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 $L^B(\hat{\beta})$. 
Estimating the dose–response function

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 `gmpgsc.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 \( t = \text{median of the group} \).

   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 Ichin (2002) and to `psmatch2.ado` by Leuven and Sianesi (2003).

---

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

$$
\varphi\{E(Y_i|T_i, R_i)\} = \lambda(T_i, R_i; \alpha) = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 T_i^3 + \alpha_4 R_i + \alpha_5 R_i^2 + \alpha_6 R_i^3 + \alpha_7 T_i R_i
$$

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,

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

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.

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.

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 $\alpha$ 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.

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

```
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(#) opt_nb(string) 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

\texttt{gpscore(\textit{newvar})} specifies the variable name for the estimated GPS via GLM.

\texttt{sigma(\textit{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.

\texttt{family(\textit{string})} specifies the distribution family name of the treated variable.

\texttt{link(\textit{string})} specifies the link function for the treated variable. The default is the canonical link for the \texttt{family()} specified.

6.2 Optional

\texttt{flag.b(\#)} skips either the balancing or the normal test or both and takes as arguments 0, 1, or 2. If \texttt{flag.b()} is not specified, the program estimates the GPS performing both the balancing and the normal tests. \texttt{flag.b(0)} skips both the balancing and the normal tests; \texttt{flag.b(1)} skips only the balancing test; \texttt{flag.b(2)} skips only the normal test.

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

\texttt{opt_b(\textit{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, \texttt{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 \texttt{prize}) that allows us to use the binary distribution function. In particular, the flogit case has been implemented by using the treatment variable \texttt{fraction}, which is obtained by normalizing the variable \texttt{prize} with respect to its highest value in the sample. Accordingly, the results of \texttt{glgpscore} and \texttt{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] \texttt{gsem family-and-link options}. 
Estimating the dose–response function

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

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>.0103137</td>
</tr>
<tr>
<td>5%</td>
<td>.0202446</td>
</tr>
<tr>
<td>10%</td>
<td>.0231977</td>
</tr>
<tr>
<td>25%</td>
<td>.0351369</td>
</tr>
<tr>
<td>50%</td>
<td>.0654881</td>
</tr>
<tr>
<td>75%</td>
<td>.1299367</td>
</tr>
<tr>
<td>90%</td>
<td>.2702822</td>
</tr>
<tr>
<td>95%</td>
<td>.3482539</td>
</tr>
<tr>
<td>99%</td>
<td>.629324</td>
</tr>
</tbody>
</table>

note: T has noninteger values

Generalized linear models
Optimization : ML
Residual df = 228
Scale parameter = 1

Deviance = 25.91237504  (1/df) Deviance = .1136508
Pearson = 29.27318861  (1/df) Pearson = .128391

Variance function: V(u) = u*(1-u/1)  [Binomial]
Link function : g(u) = ln(u/(1-u))  [Logit]

Log pseudolikelihood = -62.53528122  AIC = .6038733
                      BIC = -1220.805
B. Guardabascio and M. Ventura

### Robust Coefficients and Standard Errors

| Variable | Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------|-------|-----------|-------|-----|----------------------|
| male     | 0.6402121 | 0.1694826   | 3.78  | 0.000 | [0.3080323, 0.9723918] |
| owhns    | -0.1515907 | 0.1086591   | -1.40 | 0.163 | [-0.3645586, 0.0613773] |
| owocoll  | 0.0401978  | 0.0431132   | 0.93  | 0.351 | [-0.0445026, 0.1246982] |
| txtbot   | 0.0202427  | 0.0249659   | 0.81  | 0.417 | [-0.0286895, 0.0691749] |
| workthen | 0.1558366  | 0.2139876   | 0.73  | 0.466 | [-0.2635714, 0.5752446] |
| yearw    | -0.0169543 | 0.0603052   | -0.28 | 0.779 | [-0.1351503, 0.1012416] |
| yearn1   | -0.0055257 | 0.0131275   | -0.42 | 0.674 | [-0.0312552, 0.0202037] |
| yearn2   | 0.0089422  | 0.0134262   | 0.67  | 0.505 | [-0.0173726, 0.038257] |
| _cons    | -2.146518  | 0.5413156   | -3.97 | 0.000 | [-3.207477, -1.085559] |

Robust standard errors reported

Estimated generalized propensity score

Percentiles Smallest

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Value</th>
<th>Largest</th>
<th>Std. Dev.</th>
<th>Sum of Wgt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.0556678</td>
<td>0.0556678</td>
<td>0.0537445</td>
<td>237</td>
</tr>
<tr>
<td>5%</td>
<td>0.0600080</td>
<td>0.0600080</td>
<td>0.0547833</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.0659206</td>
<td>0.0659206</td>
<td>0.0556678</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>0.0749973</td>
<td>0.0749973</td>
<td>0.0563906</td>
<td>237</td>
</tr>
<tr>
<td>50%</td>
<td>0.1254999</td>
<td>0.1254999</td>
<td>Mean</td>
<td>0.1138546</td>
</tr>
<tr>
<td>75%</td>
<td>0.1413647</td>
<td>0.1413647</td>
<td>0.217338</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>0.1541515</td>
<td>0.1541515</td>
<td>Variance</td>
<td>0.0015032</td>
</tr>
<tr>
<td>95%</td>
<td>0.167948</td>
<td>0.167948</td>
<td>Skewness</td>
<td>0.2804054</td>
</tr>
<tr>
<td>99%</td>
<td>0.2175611</td>
<td>0.2175611</td>
<td>Kurtosis</td>
<td>2.511468</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>gps_1</td>
<td>237</td>
<td>0.1138546</td>
<td>0.0387714</td>
<td>0.0537445</td>
<td>0.2256652</td>
</tr>
<tr>
<td>Variable</td>
<td>Obs</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-------</td>
<td>-----------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>gps_2</td>
<td>237</td>
<td>0.1138546</td>
<td>0.0387714</td>
<td>0.0537445</td>
<td>0.2256652</td>
</tr>
<tr>
<td>Variable</td>
<td>Obs</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-------</td>
<td>-----------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>gps_3</td>
<td>237</td>
<td>0.1138546</td>
<td>0.0387714</td>
<td>0.0537445</td>
<td>0.2256652</td>
</tr>
</tbody>
</table>

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
### Estimating the dose–response function

**Treatment Interval No 1** - \([0.0023494709748775, 0.0474060922861099]\)

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Standard Deviation</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>0.07032</td>
<td>0.03214</td>
</tr>
<tr>
<td>ownhs</td>
<td>0.27061</td>
<td>0.13368</td>
</tr>
<tr>
<td>ovnco1</td>
<td>0.14939</td>
<td>0.21863</td>
</tr>
<tr>
<td>tixbot</td>
<td>0.09136</td>
<td>0.43645</td>
</tr>
<tr>
<td>workthen</td>
<td>-0.01029</td>
<td>0.05015</td>
</tr>
<tr>
<td>yearw</td>
<td>0.15477</td>
<td>0.18022</td>
</tr>
<tr>
<td>yearm1</td>
<td>1.4991</td>
<td>1.7217</td>
</tr>
<tr>
<td>yearm2</td>
<td>1.823</td>
<td>1.5597</td>
</tr>
</tbody>
</table>

**Treatment Interval No 2** - \([0.0476247407495975, 0.1631902456283569]\)

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Standard Deviation</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>-0.06435</td>
<td>0.02183</td>
</tr>
<tr>
<td>ownhs</td>
<td>-0.13305</td>
<td>0.13008</td>
</tr>
<tr>
<td>ovnco1</td>
<td>-0.18433</td>
<td>0.19743</td>
</tr>
<tr>
<td>tixbot</td>
<td>4.8247</td>
<td>0.38721</td>
</tr>
<tr>
<td>workthen</td>
<td>-0.00199</td>
<td>0.04998</td>
</tr>
<tr>
<td>yearw</td>
<td>0.33553</td>
<td>0.1666</td>
</tr>
<tr>
<td>yearm1</td>
<td>0.07426</td>
<td>1.6071</td>
</tr>
<tr>
<td>yearm2</td>
<td>-0.10633</td>
<td>1.4601</td>
</tr>
</tbody>
</table>

**Treatment Interval No 3** - \([0.1711813360452652, 1]\)

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Standard Deviation</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>-0.01669</td>
<td>0.03175</td>
</tr>
<tr>
<td>ownhs</td>
<td>0.19524</td>
<td>0.17768</td>
</tr>
<tr>
<td>ovnco1</td>
<td>0.18711</td>
<td>0.27456</td>
</tr>
<tr>
<td>tixbot</td>
<td>0.47912</td>
<td>0.50744</td>
</tr>
<tr>
<td>workthen</td>
<td>-0.05865</td>
<td>0.07293</td>
</tr>
<tr>
<td>yearw</td>
<td>0.23415</td>
<td>0.22407</td>
</tr>
<tr>
<td>yearm1</td>
<td>-0.70637</td>
<td>1.966</td>
</tr>
<tr>
<td>yearm2</td>
<td>-1.1814</td>
<td>1.7682</td>
</tr>
</tbody>
</table>

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 [I]

```
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)
> 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: \( Y = T + T^2 + GPS + GPS^2 + T\times GPS \)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 202</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4.2029e+09</td>
<td>5</td>
<td>840589784</td>
<td>F( 5, 196) = 4.44</td>
</tr>
<tr>
<td>Residual</td>
<td>3.7122e+10</td>
<td>196</td>
<td>189397662</td>
<td>R-squared = 0.1017</td>
</tr>
<tr>
<td>Total</td>
<td>4.1325e+10</td>
<td>201</td>
<td>205596471</td>
<td>Root MSE = 13762</td>
</tr>
</tbody>
</table>

| 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 194785 |
| 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 |
Estimating the dose–response function

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


Estimating the dose–response function


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.