

# Distribution-Free Estimation of Heteroskedastic Binary Response Models in Stata

Jason R. Blevins  
Duke University  
jrb11@duke.edu

Shakeeb Khan  
Duke University  
shakeebk@econ.duke.edu

**Abstract.** This paper considers two recently developed semiparametric estimators for distribution-free binary response models exhibiting multiplicative heteroskedasticity under a conditional median restriction. It shows that these estimators can be implemented in Stata through simple modifications to the nonlinear least squares Probit criterion function. Both estimators are implemented in a new Stata command, **dfbr**, and several examples of its usage are provided.

**Keywords:** st0001, dfbr, binary response, heteroskedasticity, Probit, semiparametric estimation, sieve estimation

## 1 Introduction

This paper develops Stata implementations for two recently developed semiparametric, distribution-free estimators for binary response models of the form

$$y_i = 1 \{x_i' \beta_0 + \varepsilon_i > 0\} \quad (1)$$

where  $y_i \in \{0, 1\}$  is an observed response variable,  $x_i$  is a vector of observed covariates, and  $\varepsilon_i$  is an unobserved disturbance term.  $\beta_0$  is an unknown vector of parameters of interest. Following Manski (1975, 1985) and Horowitz (1992), we impose a conditional median restriction to ensure identification of  $\beta_0$ :

$$\text{med}(\varepsilon_i | x_i) = 0.$$

We wish to estimate  $\beta_0$  given a random sample  $\{y_i, x_i\}_{i=1}^n$ .

This paper proceeds as follows. Section 2 reviews Stata's nonlinear least squares (NLLS) estimation framework and introduces the NLLS Probit model, the model above with  $\varepsilon_i \sim N(0, 1)$ , as a motivating example. Sections 3 and 4 describe, respectively, the sieve nonlinear least squares (SNLLS) and local nonlinear least squares (LNLLS) estimators of Khan (2006). We show that both of these estimators can be easily implemented in Stata by making simple modifications to the standard NLLS Probit regression function. Finally, Section 5 describes **dfbr**, a new Stata module which implements both of these estimators, and provides examples of its usage.

## 2 Nonlinear Least Squares Estimation in Stata

Stata's **nl** command is an interface for fitting an arbitrary nonlinear regression function  $f(x, \theta) = E[y | x, \theta]$  using least squares. There are three ways to provide the regression

function to **nl**: interactively using a substitutable expression, via a substitutable expression program, or using a function evaluator program. We focus on the first approach since it is straightforward to implement for most simple models, including the ones we discuss in the following sections. Furthermore, this is the method used internally by the Stata program **dfbr** introduced in Section 5. See [R] **nl** for further details regarding Stata's NLLS capabilities.

Consider the standard Probit regression model

$$E[y_i | x_i] = \Phi(x_i' \beta_0) \quad (2)$$

where  $\beta_0$  is a vector of parameters of interest. This is precisely the model in (1) when  $\varepsilon_i \sim N(0, 1)$ . Given a sample of size  $n$ ,  $\{y_i, x_i\}_{i=1}^n$ , the nonlinear least squares estimator  $\hat{\beta}_n$  of  $\beta_0$  is defined as

$$\hat{\beta}_n = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n [y_i - \Phi(x_i' \beta)]^2. \quad (3)$$

To be concrete, suppose that we have a binary dependent variable **y** and two independent variables **x1** and **x2**. To estimate the model using the **nl** command, we can express the regression function (2) as a substitutable expression:

```
. nl ( y = normal({b0} + {b1}*x1 + {b2}*x2) )
```

Here, **normal()** is the Normal cumulative distribution function. In this particular example, the intercept **b0** and the two coefficients **b1** and **b2** will be estimated according to (3).

### 3 Local Nonlinear Least Squares Estimator

The local nonlinear least squares estimator introduced by Khan (2006) is defined as

$$\hat{\beta} = \arg \min_{\beta \in \Theta \times 1} \frac{1}{n} \sum_{i=1}^n \left[ y_i - \Phi \left( \frac{x_i' \beta}{h_n} \right) \right]^2.$$

where  $h_n$  is a sequence of positive numbers such that  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ . We adopt the standard semiparametric scale normalization (cf. Horowitz (1992)) and normalize the  $k$ -th element of  $\beta$  to 1 so that  $\hat{\beta} = (\hat{\theta}', 1)'$ .

Other than these two slight modifications, this objective function is identical to that of the NLLS Probit estimator in (3). Thus, in order to implement the estimator in Stata we simply need to normalize one component of  $\beta$  to 1 and allow for  $h_n$  to be specified in a way that depends on the sample size.

Using the two-regressor example from before, we can estimate the model using the bandwidth  $h_n = n^{-1/3}$  as follows:

```
. local h = _N^(-1/3)
. nl ( y = normal(({b1}*x1 + x2)/'h') )
```

where  $N$  is the sample size. We normalized the coefficient on  $x_2$  to 1 by simply omitting this parameter leaving  $b_1$ , the coefficient on  $x_1$ , as the only parameter.

## 4 Sieve Nonlinear Least Squares Estimator

Although the objective function for the sieve nonlinear least squares estimator introduced by Khan (2006) is somewhat more complicated, it is still ultimately a slight variation on the NLLS Probit objective function in (3) and is easy to implement using `nl`. Specifically, the criterion function is

$$Q_n(\theta, \ell) = -\frac{1}{n} \sum_{i=1}^n [y_i - \Phi(x_i' \beta \cdot \exp(\ell(x_i)))]^2$$

where  $\ell$  is an infinite-dimensional scaling parameter and  $\beta = (\theta', 1)'$  is a finite vector of parameters.

In order to use NLLS we introduce a finite-dimensional approximation of  $\ell$  using a linear-in-parameters sieve estimator. Let  $b_{0j}(x_i)$  denote a sequence of known basis functions and let  $b^{\kappa_n}(x_i) = (b_{01}(x_i), \dots, b_{0\kappa_n}(x_i))'$  for some integer  $\kappa_n$ . The function  $g(x_i) \equiv \exp(\ell(x_i))$  in the above objective function can be approximated by  $g_n(x_i) = \exp(b^{\kappa_n}(x_i)' \gamma_n)$  where  $\gamma_n$  is a vector of constants. Let  $\alpha_n \equiv (\Theta, \gamma_n) \in \mathcal{A}_n$  where  $\mathcal{A}_n$  is the sieve space. The estimator can be formally defined by

$$\hat{\alpha}_n = \arg \min_{\alpha \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n [y_i - \Phi(x_i' \beta \cdot g_n(x_i))]^2$$

where  $\beta = (\theta', 1)'$  as before.

To illustrate Stata implementation of this estimator, consider the simple model with two regressors  $x_1$  and  $x_2$ . We use a scaling function consisting of polynomial terms of  $x_i$ :

$$g_n(x_i) = \exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1^2 + \gamma_4 x_2^2 + \gamma_5 x_1 x_2).$$

To estimate the model using `nl`, we construct the corresponding substitutable expression:

```
. nl ( y = normal((b1)*x1 + x2) * exp({g0} + {g1}*x1
+ {g2}*x2 + {g3}*x1*x1 + {g4}*x2*x2 + {g5}*x1*x2) )
```

Again we have normalized the coefficient on  $x_2$  by omitting the corresponding parameter.

## 5 The `dfbr` Command

The Stata program `dfbr` implements both of the estimators described above: the sieve nonlinear least squares estimator and the local nonlinear least squares estimator. It works by automatically generating the required substitutable expressions for an arbitrary list of dependent variables and passing this expression to Stata's `nl` command to

perform the estimation. For sieve estimation the user may supply a set of basis variables such as polynomial terms of the independent variables. If no basis elements are provided the dependent variables are used. For the local nonlinear least squares estimator the user may supply a bandwidth or allow **dfbr** to select it automatically. In both cases the coefficient on the last dependent variable is normalized to 1.

## 5.1 Syntax

Sieve nonlinear least squares estimation:

```
dfbr depvar indepvars [if][in][, sieve basis(basis_vars)]
```

Local nonlinear least squares estimation:

```
dfbr depvar indepvars [if][in], local [bandwidth(#)]
```

## 5.2 Options

Sieve nonlinear least squares options:

- **sieve** specifies the sieve nonlinear least squares estimator (default).
- **basis(basis\_vars)** provides a list of basis variables to use in the linear in parameters sieve approximation of the scaling function. If this option is omitted, a linear combination of the regressors is used.

Local nonlinear least squares options:

- **local** specifies the local nonlinear least squares estimator.
- **bandwidth(#)** specifies the bandwidth. If this option is omitted the default bandwidth  $n^{-1/3}$  is used (where  $n$  is the sample size).

## 5.3 Examples

First, suppose we generate a random sample from a simple binary response model with two normally distributed regressors  $x_1$  and  $x_2$  and a uniformly distributed error term  $u$ :

```
. set seed 100
. set obs 1000
obs was 0, now 1000
. generate x1 = 1 + invnorm(uniform())
. generate x2 = invnorm(uniform())
. generate u = sqrt(12)*uniform() - sqrt(12)/2
. generate y = 0.5 * x1 + x2 - u > 0
```

In particular,  $x_1 \sim N(1, 1)$ ,  $x_2 \sim N(0, 1)$ , and  $u$  is normalized to have mean 0 and variance 1. The coefficient on  $x_2$  is normalized to 1.

To estimate the model using sieve nonlinear least squares using second-degree polynomial terms as the sieve basis functions:

```
. generate x1x2 = x1 * x2
. generate x1_2 = x1^2
. generate x2_2 = x2^2
. dfbr y x1 x2, sieve basis(x1 x2 x1x2 x1_2 x2_2)
Method: Sieve Nonlinear Least Squares (SNLLS)
Sieve basis: 1 x1 x2 x1x2 x1_2 x2_2
(obs = 1000)
```

Source	SS	df	MS			
Model	484.260335	7	69.1800479	Number of obs =	1000	
Residual	156.739665	993	.157844577	R-squared =	0.7555	
Total	641	1000	.641	Adj R-squared =	0.7538	
				Root MSE =	.3972966	
				Res. dev. =	984.708	

  

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/x1	.5152792	.0480166	10.73	0.000	.4210536	.6095048
/g_const	-.3996111	.1800909	-2.22	0.027	-.7530135	-.0462086
/g_x1	-.0548818	.1547367	-0.35	0.723	-.3585302	.2487667
/g_x2	.0596885	.1289139	0.46	0.643	-.1932864	.3126634
/g_x1x2	.0420631	.1332563	0.32	0.752	-.2194332	.3035594
/g_x1_2	.1056956	.0797799	1.32	0.186	-.0508609	.2622522
/g_x2_2	.1856881	.1363003	1.36	0.173	-.0817816	.4531578

Coefficient on x2 normalized to 1.

To estimate the model using the local nonlinear least squares estimator with the default bandwidth:

```
. dfbr y x1 x2, local
Method: Local Nonlinear Least Squares (LNLLS)
Bandwidth: .1 (default: n^(-1/3))
(obs = 1000)
```

Source	SS	df	MS			
Model	424.762751	1	424.762751	Number of obs =	1000	
Residual	216.237249	999	.216453703	R-squared =	0.6627	
Total	641	1000	.641	Adj R-squared =	0.6623	
				Root MSE =	.4652459	
				Res. dev. =	1306.498	

  

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/x1	.5815042	.0132432	43.91	0.000	.5555165	.6074918

Coefficient on x2 normalized to 1.

A custom bandwidth such as 0.15 can be chosen using the bandwidth option:

```
. dfbr y x1 x2, local bandwidth(0.15)
(output omitted)
```



———. 1998. *Semiparametric Methods in Econometrics*. New York: Springer.

Khan, S. 2006. Distribution Free Estimation of Heteroskedastic Binary Response Models Using Probit Criterion Functions. Working Paper, Duke University.

Manski, C. F. 1975. Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics* 3: 205–228.

———. 1985. Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics* 27: 313–333.

**About the authors**

Jason R. Blevins is a Ph.D. candidate in the Department of Economics at Duke University.

Shakeeb Khan is an Associate Professor of Economics at Duke University.