The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–9

Robinson's square root of N consistent semiparametric regression estimator in Stata

Vincenzo Verardi
University of Namur, FNRS
(Centre for Research in the Economics of Development)
Namur, Belgium
and Université Libre de Bruxelles
(European Center for Advanced Research in Economics and Statistics
and Center for Knowledge Economics)
Brussels, Belgium
vverardi@fundp.ac.be

Nicolas Debarsy
University of Namur
(Center of Research in Regional Economics and Economic Policy)
Namur, Belgium
and Université d'Orléans
(Laboratoire d'Economie D'Orléans)
Orléans, France
ndebarsy@fundp.ac.be

Abstract. This paper describes Robinson's (1988) double residual semiparametric regression estimator and Hardle and Mammen's (1993) specification test implementation in Stata. Some simple simulations illustrate how this newly coded estimator outperforms the already available semiparametric `plreg` command.

Keywords: st0001, semipar, Semiparametric estimation
JEL Classification: C14, C21

1 Introduction

The objective of this paper is to present the implementation in Stata of Robinson's (1988) double residual semiparametric regression estimator. Also, to check if the nonparametric part of the relation may be approximated by a polynomial functional form, we introduce Hardle and Mammen's (1993) specification test as an option in the programmed estimator. We also briefly describe this test.

The structure of the paper is the following: in Section 2, Robinson's (1988) semiparametric regression estimator and Hardle and Mammen's (1993) specification test are described. In Section 3, the implemented Stata command (`semipar`) is presented. Some simple simulations assessing the performance of the estimator and of the test are performed in Section 4. In Section 5, we illustrate the use of the `semipar` command with an empirical application. Section 6 concludes.

2 Estimation method

2.1 Robinson's (1988) semiparametric regression estimator

Consider a general model of the type

$$y_i = \theta_0 + \mathbf{x}_i\theta + f(z_i) + \varepsilon_i, \quad i = 1, ..., N \tag{1}$$

where y_i is the value taken by the dependent variable for individual i, $\mathbf{x}_i$ is the row vector of characteristics of individual i, θ_0 is a constant term and ε_i is the disturbance assumed to have zero mean and constant variance σ_ε^2. Variable $\mathbf{z}$ is an explanatory variable that enters the equation nonlinearly according to a non-binding function f. This model can be estimated using Robinson's (1988) double residual methodology that starts by applying a conditional expectation to both sides of (1). This leads to

$$E\left(y_i|z_i\right) = \theta_0 + E\left(\mathbf{x}_i|z_i\right)\theta + f(z_i) \quad i = 1, ..., N \tag{2}$$

By subtracting (2) from (1), we have

$$y_i - E\left(y_i|z_i\right) = \left(\mathbf{x}_i - E\left(\mathbf{x}_i|z_i\right)\right)\theta + \varepsilon_i \quad i = 1, ..., N \tag{3}$$

If the conditional expectations are known, parameter vector θ can easily be estimated by fitting (3) by ordinary least squares. If they are unknown, they have to be estimated by calling on some consistent estimators $y_i = m_y(z_i) + \varepsilon_{1i}$ and $x_{ki} = m_{x_k}(z_i) + \varepsilon_{2ki}$, where $k = 1, ..., K$ is the index of the explanatory variables entering the model parametrically. Robinson's (1988) double residual estimator is hence the OLS estimation of model

$$y_i - \hat{m}_y(z_i) = \left(\mathbf{x}_i - \hat{m}_{\mathbf{x}}(z_i)\right)\theta + \varepsilon_i \quad i = 1, ..., N \tag{4}$$

where $\mathbf{x}_i - \hat{m}_{\mathbf{x}}(z_i)$ is the row-vector of the differences between each explanatory variable x_{ki} and the fitted conditional expectation of x_{ki} given z_i.

The estimated coefficients vector is therefore

$$\hat{\theta} = \left(\sum_i \left(\mathbf{x}_i - \hat{m}_{\mathbf{x}}(z_i)\right)' \left(\mathbf{x}_i - \hat{m}_{\mathbf{x}}(z_i)\right)\right)^{-1} \sum_i \left(\mathbf{x}_i - \hat{m}_{\mathbf{x}}(z_i)\right)' \left(y_i - \hat{m}_y(z_i)\right) \tag{5}$$

with variance (if errors are i.i.d)

$$Var\left(\hat{\theta}\right) = \sigma_\varepsilon^2 \left(\sum_i \left(\mathbf{x}_i - \hat{m}_x(z_i)\right)' \left(\mathbf{x}_i - \hat{m}_x(z_i)\right)\right)^{-1} \tag{6}$$

where σ_ε^2 is the variance of the error term. If errors are non i.i.d., standard sandwich and cluster variance formulas can be used.

Having estimated parameter vector θ, it is now possible to fit the nonlinear relation between z_i and y_i by simply estimating equation (7) presented below nonparametrically.

$$y_i - \mathbf{x}_i\hat{\theta} = \theta_0 + f(z_i) + \varepsilon_i, \quad i = 1, ..., N \tag{7}$$

2.2 Hardle and Mammen's (1993) test

It is sometimes suggested that nonparametric functions may be approximated by some parametric polynomial alternative. To test for the appropriateness of such an approximation, Hardle and Mammen (1993) develop a statistic which compares the nonparametric and parametric regression fits using squared deviations between them. The test-statistic is:

$$T_n = N\sqrt{h}\sum_{i=1}^{N}\left(\hat{f}(z_i) - \hat{f}(z_i, \theta)\right)^2 \pi(\cdot) \tag{8}$$

where $\hat{f}(z_i)$ is the nonparametric function estimated in (7), $\hat{f}(z_i, \theta)$ is an estimated parametric function, h is the bandwidth used and $\pi(\cdot)$ is a weighting function for the squared deviations between fits. To obtain critical values for the test, Hardle and Mammen (1993) suggest to call on simulated values obtained by wild bootstrap. Obviously, an absence of rejection of the null (i.e. "accepting" the parametric model) means that the polynomial adjustment is at least of the degree that has been tested.

We implemented this estimator and the specification test in Stata under the command `semipar` which is described below.

3 The semipar command

The `semipar` command fits Robinson's double residual estimator in the case of a unique variable entering the model nonparametrically. The default kernel regression used for all stages is a gaussian kernel weighted local polynomial fit.[1] The optimal bandwidth used minimizes the conditional weighted mean integrated squared error.

The general syntax for the command is:

```
semipar varlist [if] [in] [weight] , nonpar(varname) [ generate(string)
   partial(string) kernel(string) degree(#) trim(#) nograph ci level(#)
   title(string) ytitle(string) xtitle(string) robust cluster(varname)
   test(#) nsim(#) weight_test(varname) ]
```

The first option, `nonpar`, is compulsory and necessary to declare which variable

[1] The kernel is of order 2.

enters the model nonparametrically. All other choices are optional. The first one, `generate`, reproduces the "nonparametrically" fitted dependent variable. The user chooses the name of this new variable by defining it in parentheses. Similarly, the `partial` option is needed to generate a new variable that contains the parametric residuals (i.e. the left-hand side of equation (7)). The `kernel` allows to change the kernel function. The `degree` option allows the user to specify the degree of the local polynomial fit used to nonparametrically estimate the regressions. By default, it is set to 1. The `trim` option allows to trim the data relying on a value of the probability distribution function of the `nonpar` variable. The default value is set to 0 (no trimming). The option `nograph` should be used if the user does not want the graph of the nonparametric fit of the variable set in `nonpar` to appear. The `ci` option allows to visualize the confidence interval around the nonparametric fit[2], while the `level` option sets the level of confidence for inference (by default set to 95%). The options `title`, `ytitle` and `xtitle` are used to indicate respectively the title and the labels of the axes of the graph illustrating the nonparametric relation between the dependent variable and the variable defined in the `nonpar` option. The `robust` and `cluster` options call for standard errors of the estimated parameters that are respectively resistant to heteroskedasticity and clustered errors. The `test` option implements Hardle and Mammen's (1993) statistic to test whether the nonparametric fit could be approximated by a polynomial fit, the order of which must be set by the user. For the sake of clarity, we rescaled the statistic in such a way that it can be compared with the quantile of a Normal distribution. Note however that the test is not normally distributed. The `nsim` option defines the number of bootstrap replicates used to get inference. Its default value is set to 100. Finally, the `weight_test` option allows to give different weights to the squared deviation between the nonparametric fit and the polynomial adjustment in the computation of the test (i.e. introducing $\pi(\cdot)$ in equation (8)). By default, this weighting vector is set to ι_N/N with ι_N a unit vector of dimension N. To assess the performance of the programmed estimator, in the next section we present some simple simulations in which we compare this estimator with the already available `plreg` command. The latter implements Yatchew's (1998) difference estimator where the nonparametric part in (1) is partialled out by differencing rather than by removing the conditional expectations. Since the highest efficiency of Yatchew's (1998) estimator is attained by a differencing of order 10, we will use this differencing order as a benchmark.

4 Simulations

The simulation setup is the following. To begin, we generate (for a sample of 300 observations) two explanatory variables x_2 and x_3 from two independent $N(0,1)$. An additional random variable x_1 is generated from a discrete Uniform distribution on $[-10,10]$. This sample design remains unchanged for all simulations. Then, for each replication, we generate an error term e from a standard normal and create variable y according to DGP $y = x_1 + x_1^2 + x_2 + x_3 + e$. We run the `semipar` and `plreg` estimators for each replication and Table 1 reports both the bias and MSE of the coefficients associated

[2] Further information about confidence intervals can be found in the help of `lpoly`.

Table 1: Comparison between `semipar` and `plreg`

	Bias x_2	Bias x_3	MSE x_2	MSE x_3
plreg	-0.4695	-0.1039	0.2208	0.0112
semipar	-0.0435	-0.0183	0.0022	0.0007

with x_2 and x_3. We carry out 1000 simulations. The variable that enters the equation nonparametrically is generated from a discrete Uniform distribution on purpose to illustrate the fragility of `plreg` with respect to this kind of data. Robinson's (1988) estimator, that is based on partialling out the nonparametric part removing conditional expectations rather than by differencing, behaves much better.

In this setup, Robinson's (1988) estimator leads to smaller biases than Yatchew's (1998) differencing estimator. From equation (7), this also implies that the nonparametric fit is better estimated by `semipar` than by `plreg`.

To illustrate the fitting performance of the proposed estimation procedure, we generate four samples according to the following DGPs:

a) $y = x_2 + x_3 + e$

b) $y = x_1 + x_2 + x_3 + e$

c) $y = x_1 + x_1^2 + x_2 + x_3 + e$

d) $y = x_1 - x_1^2 - x_1^3 + x_2 + x_3 + e.$

In Figure 1, we present the scatter plots, the non-parametric fit (thick plain line) and the true DGP (red dashed line) related to the four DGPs described above. As expected, the results are unambiguous.

In the absence of any relation between x_1 and y (panel a), no clear pattern emerges and the non-parametric curve lies close to the horizontal line (the true DGP). In the three other cases (panels b, c and d), the nonparametric estimation of the relation matches the true functional form quite well.

As mentioned in the previous section, the T_n statistic assesses the adequacy of a polynomial adjustment compared to a nonparametric fit. Table 2 presents the performance of the test for the DGPs described above. The rows indicate the order of the generated polynomial while the columns specify the order of the polynomial that has been tested. Thus, the diagonal (and the upper triangle) elements are the simulated sizes of the test while elements below the main diagonal are some measure of power. To construct this table we replicated the DGPs 1000 times. Each time a new error term is randomly drawn and a new dependent variable is generated (the design space remains unchanged). Inference for the test is based on 100 bootstrap replications. We observe that the test has good rejection rates when the order of the polynomial adjustment tested is lower than the generated one. Besides, the size of the test (whose theoretical value is set at 5%), is very close to its nominal value.

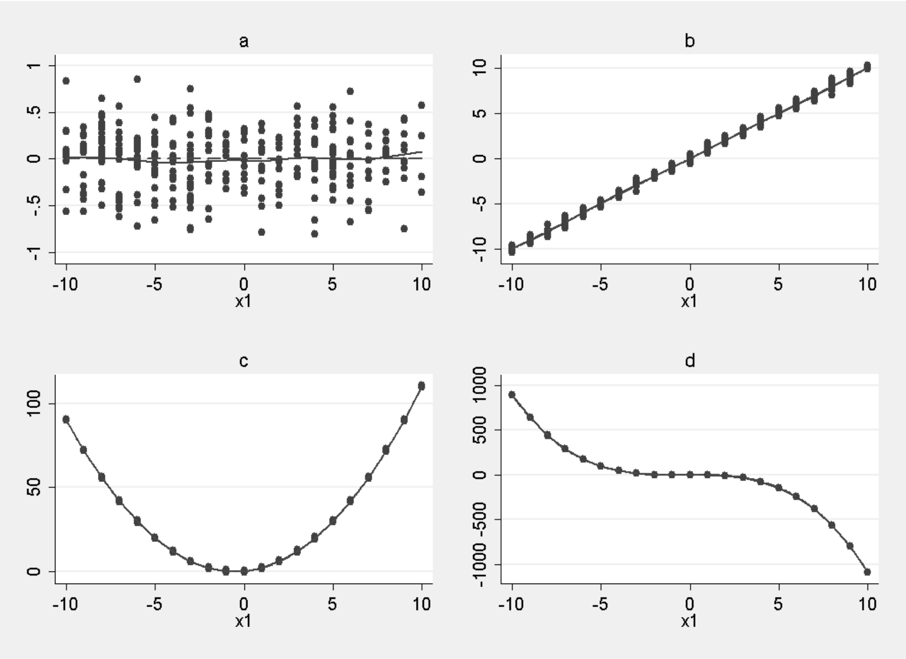

Figure 1: Non-parametric fit of the four DGPs

Table 2: Performance of the comparison Test T_n

		Order tested			
		0	1	2	3
	0	0.053	0.06	0.055	0.039
True	1	1	0.064	0.055	0.021
Order	2	1	1	0.06	0.062
	3	1	1	1	0.066

Figures correspond to rejection rates of the test.

5 Example

To illustrate the usefulness of this semiparametric model in empirical applications, we call on a dataset used by Wooldridge (2002) that studies the effects of an incinerator location on housing prices. The data are for houses that were sold during the year 1981 in North Andover, MA; 1981 was the year construction began on a local garbage incinerator. The dependent variable is the log of price of houses (*lprice*) and the variable of interest is the distance from the house to the incinerator measured in feet and expressed in logs (*ldist*). To control for confounding effects, the author suggests to include the log of interstate distance (*linst*), the log of the square footage of the house (*larea*), the log of the lot size in square feet (*lland*), the number of rooms (*rooms*), the number of bathrooms (*baths*), and the age of the house (*age*) as additional covariate. However, he also asserts that the effect of the log of the interstate distance is not linear and proposes to consider it squared. In this application we carry out this exercise again but do not impose any functional form to the log of interstate distance and estimate the model semiparametrically. We then check if the square approximation is appropriate. More precisely, we run the following command lines:

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/HPRICE3
. semipar lprice ldist larea lland rooms baths age, nonpar(linst) xtitle(linst) ci
                                                     Number of obs =     321
                                                     R-squared     =  0.4437
                                                     Adj R-squared =  0.4331
                                                     Root MSE      =  0.2646
------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldist |   .1083941   .0640184     1.69   0.091    -.0175636    .2343519
       larea |   .4887243   .0668208     7.31   0.000     .3572527    .6201959
       lland |   .0866459    .036037     2.40   0.017     .0157423    .1575495
       rooms |   .0436451   .0221781     1.97   0.050     9.12e-06     .087281
       baths |   .0806555   .0335251     2.41   0.017      .014694     .146617
         age |   -.003481   .0005436    -6.40   0.000    -.0045506   -.0024114
------------------------------------------------------------------------------
```

The results of the parametric part (see Stata output above) show that the distance from the incinerator does not seem to be significant (the t-stat associated with the coefficient is smaller than the critical value of 1.96).

As far as the effect of the log of the interstate distance is concerned, Figure (2) shows that it is clearly nonlinear.

Indeed, when the interstate distance increases, the effect of house prices first increases and then decreases. When we check if the quadratic approximation proposed by Wooldridge (2002) is appropriate, it turns out that this assumption is clearly rejected by Hardle and Mammen's (1993) test (see below). However, when we compare it with a polynomial adjustment of degree 3, the null is no longer rejected which means that instead of a semiparametric model, a pure parametric model with a polynomial fit of degree 3 of *linst* could be used.

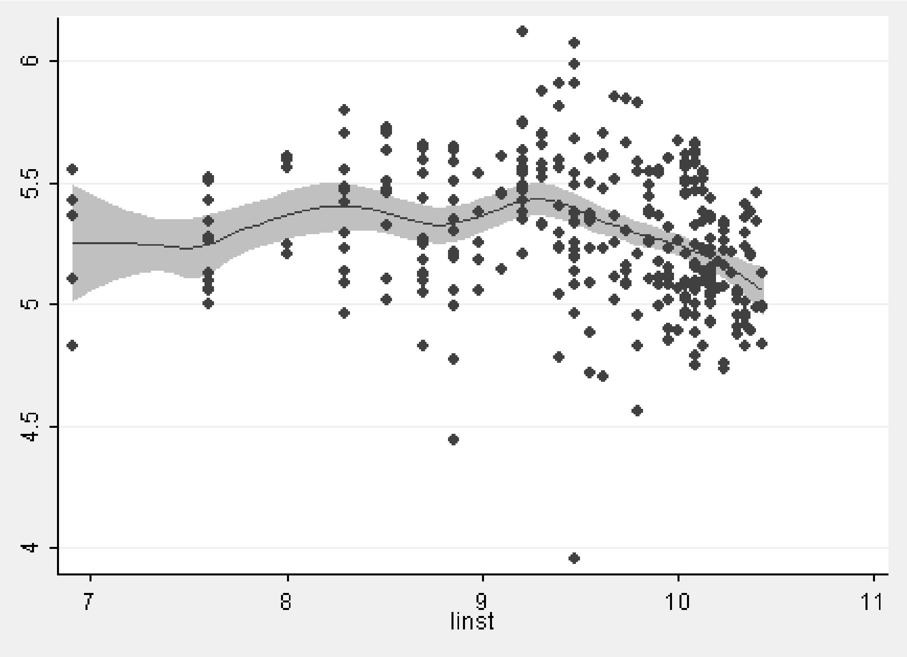

Figure 2: Nonlinear link between the price and interstate distance (in logs)

The two Stata outputs below summarize results of the Hardle and Mammen (1993) test when the polynomial adjustment tested is of order 2 or 3 respectively. These outputs do not present the results concerning the parametric part since they are the same as in the output presented above.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/HPRICE3
. semipar lprice ldist larea lland rooms baths age, nonpar(linst) nograph test(2)
Simulation the distribution of the test statistic
bootstrap replicates (100)
───┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100

H0: Parametric and non-parametric fits are not different
-------------------------------------------------------
Test statistic T: 2.6499135
Critical value (95%): 1.959964
P-value: .02
```

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/HPRICE3
. semipar lprice ldist larea lland rooms baths age, nonpar(linst) nograph test(3)
Simulation the distribution of the test statistic
bootstrap replicates (100)
───┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
H0: Parametric and non-parametric fits are not different
-------------------------------------------------------
Test statistic T: 1.067547
Critical value (95%): 1.959964
P-value: .27
```

6 Conclusion

In econometrics, semiparametric regression estimators have become standard tools for applied researchers. In this paper, we present Robinson's (1988) double residual semiparametric regression estimator and Hardle and Mammen's (1993) specification test. We then present the Stata codes we created to implement them in practice. Some simple simulations and an empirical application to illustrate the usefulness of the procedure are also shown.

7 Acknowledgments

We would like to thank all our colleagues at CRED, ECARES, and CERPE as well as Emmanuel Flachaire, Marjorie Gassner, Whitney Newey, Christopher Parmeter and Sébastien Van Bellegem. The authors gratefully acknowledge the National Science Foundation of Belgium (FNRS) for its financial support.

About the author

Vincenzo Verardi holds a PhD in economics from the Université Libre de Bruxelles. He is an associate researcher of the National Science Foundation of Belgium and is a professor of economics and econometrics at the University of Namur and the Université Libre de Bruxelles. His research fields are applied econometrics, robust methods, political economy, and public economics.
Nicolas Debarsy holds a PhD in economics from the University of Namur (Belgium). He is currently post-doc researcher at the University of Namur and University of Orléans (France). His current research fields are applied econometrics and spatial econometrics.

8 References

Hardle W., E. Mammen. 1993. Comparing nonparametric versus parametric regression fits, *Annals of Statistics* 21: 1926-1947.

Robinson P.M. 1988. Root-N-consistent semiparametric regression. *Econometrica* 56: 931-954.

Wooldridge J.M. 2002. *Econometric analysis of cross section and panel data*, MIT Press, London.

Yatchew A. 1998. Nonparametric regression techniques in economics, *Journal of Economic Literature* 57: 135-143.