The Stata Journal

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 or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers 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

http://www.stata.com/bookstore/sj.html

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

U.S. and Canada		Elsewhere	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

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

http://www.stata.com/bookstore/sjj.html

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

http://www.stata-journal.com/archives.html

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.

The Stata Journal (2012)
12, Number 4, pp. 726–735

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. In this article, we describe Robinson's (1988, *Econometrica* 56: 931–954) double residual semiparametric regression estimator and Härdle and Mammen's (1993, *Annals of Statistics* 21: 1926–1947) specification test implementation in Stata. We use some simple simulations to illustrate how this newly coded estimator outperforms the already available semiparametric `plreg` command (Lokshin, 2006, *Stata Journal* 6: 377–383).

Keywords: st0278, semipar, semiparametric estimation, double residual estimator

1 Introduction

In this article, we aim 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 Härdle and Mammen's (1993) specification test as an option in the programmed estimator. We also briefly describe this test.

The structure of the article is as follows: In section 2, we describe Robinson's (1988) semiparametric regression estimator and Härdle and Mammen's (1993) specification test. In section 3, we present the implemented Stata command (`semipar`). In section 4, we perform some simple simulations assessing the performance of the estimator and of the test. In section 5, we illustrate the use of the `semipar` command with an empirical application. Section 6 concludes the article.

2 Estimation method

2.1 Robinson's (1988) semiparametric regression estimator

Consider a general model of the type

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

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 nonbinding 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

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

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

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

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 $k = 1, \ldots, K$ is the index of the explanatory variables entering the model parametrically. Robinson's (1988) double residual estimator is hence the ordinary least squares estimation of model

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

where $\mathbf{x}_i - \widehat{m}_{\mathbf{x}}(z_i)$ 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

$$\widehat{\theta} = \left[\sum_i \left\{\mathbf{x}_i - \widehat{m}_{\mathbf{x}}(z_i)\right\}' \left\{\mathbf{x}_i - \widehat{m}_{\mathbf{x}}(z_i)\right\}\right]^{-1} \sum_i \left\{\mathbf{x}_i - \widehat{m}_{\mathbf{x}}(z_i)\right\}' \left\{y_i - \widehat{m}_y(z_i)\right\}$$

with variance (if errors are independent and identically distributed)

$$\mathrm{Var}\left(\widehat{\theta}\right) = \sigma_\varepsilon^2 \left[\sum_i \left\{\mathbf{x}_i - \widehat{m}_x(z_i)\right\}' \left\{\mathbf{x}_i - \widehat{m}_x(z_i)\right\}\right]^{-1}$$

where σ_ε^2 is the variance of the error term. If errors are nonindependent and nonidentically distributed, standard sandwich and cluster variance formulas can be used.

Having estimated parameter vector θ, we can now fit the nonlinear relation between z_i and y_i by simply estimating (4) nonparametrically:

$$y_i - \mathbf{x}_i\widehat{\theta} = \theta_0 + f(z_i) + \varepsilon_i \quad i = 1, \ldots, N \tag{4}$$

2.2 Härdle 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, Härdle and Mammen (1993) developed a statistic that compares the nonparametric and parametric regression fits by using squared deviations between them. The test statistic is

$$T_n = N\sqrt{h}\sum_{i=1}^{N}\left\{\widehat{f}(z_i) - \widehat{f}(z_i, \theta)\right\}^2 \pi(\cdot) \tag{5}$$

where $\widehat{f}(z_i)$ is the nonparametric function estimated in (4), $\widehat{f}(z_i, \theta)$ 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, Härdle and Mammen (1993) suggest calling on simulated values obtained by wild bootstrap. Obviously, an absence of rejection of the null (that is, "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.

3.1 Syntax

The general syntax for the command is

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

1. The kernel is of order 2.

The nonpar() option is required to declare which variable enters the model nonparametrically. All the other options are optional. The generate() option reproduces the "nonparametrically" fit dependent variable; the user chooses the name of this new variable by defining it in parentheses. Similarly, the partial() option generates a new variable that contains the parametric residuals [that is, the left-hand side of (4)].

The degree() option allows the user to specify the degree of the local polynomial fit used to nonparametrically estimate the regressions; the default is degree(1). The trim() option allows the user to trim the data by relying on a value of the probability distribution function of the nonpar() variable; the default is trim(0) (no trimming). The kernel() option allows the user to change the kernel function.

The option nograph should be used if the user does not want to see the graph of the nonparametric fit of the variable set in nonpar(). The ci option allows the user 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() indicate, respectively, the overall title, the title of the y axis, and the title of the x axis 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 uses Härdle 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; the default is nsim(100).

Finally, the weight_test() option allows the user to give different weights to the squared deviation between the nonparametric fit and the polynomial adjustment in the computation of the test [that is, introducing $\pi(\cdot)$ in (5)]. By default, this weighting vector is set to ι_N/N with ι_N being 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 user-written command plreg (Lokshin 2006). plreg implements Yatchew's (1998) difference estimator, where the nonparametric part in (1) is partialled out by differencing rather than by removing the conditional expectations. Because the highest efficiency of Yatchew's estimator is attained by a differencing of order 10, we will use this differencing order as a benchmark.

2. Further information about confidence intervals can be found in [R] **lpoly**.

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 variables $N(0,1)$. 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 the data-generating process (DGP) $y = x_1 + x_1^2 + x_2 + x_3 + e$. We run the `semipar` and `plreg` estimators for each replication.

Table 1 reports both the bias and the mean squared error (MSE) of the coefficients associated with x_2 and x_3. We carry out 1,000 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 these kinds of data. Robinson's (1988) estimator, which is based on partialling out the nonparametric part by removing conditional expectations rather than by differencing, behaves much better.

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

In this setup, Robinson's estimator leads to smaller biases than Yatchew's differencing estimator. From (4), this also implies that the nonparametric fit is better estimated by `semipar` than it is 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 scatterplots, the nonparametric fit (thick plain line), and the true DGP (dashed line) related to the four DGPs described above. As expected, the results are unambiguous.

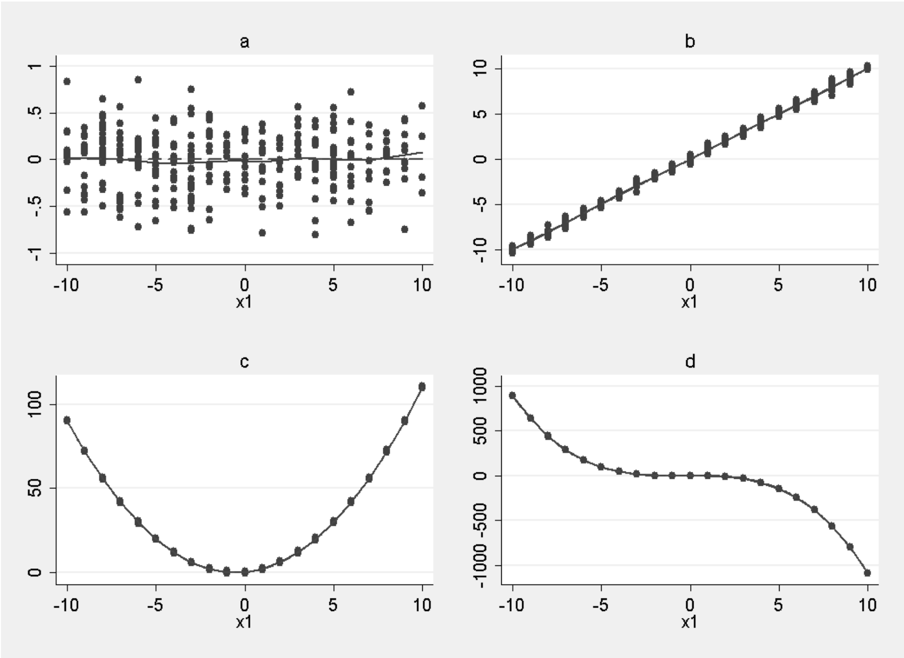

Figure 1. Nonparametric fit of the four DGPs

In the absence of any relation between x_1 and y (panel a), no clear pattern emerges, and the nonparametric 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 previously mentioned, the T_n statistic assesses the adequacy of a polynomial adjustment compared with 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 and 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 1,000 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.

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 use a dataset by Wooldridge (2002) that studies the effects of an incinerator's location on housing prices. The data are for houses that were sold in North Andover, Massachusetts, during 1981, the year construction began on a local garbage incinerator. The dependent variable is the log of the 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 the 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 covariates. 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 fit the model semiparametrically. We then check whether 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 |    .108394   .0640184     1.69   0.091    -.0175637    .2343517
       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 the Stata output above) show that the distance from the incinerator does not seem to be significant (the t statistic associated with the coefficient is smaller than the critical value of 1.96).

Figure 2 shows that the log of the interstate distance is clearly nonlinear. Indeed, when the interstate distance increases, the effect of house prices first increases and then decreases. When we check whether the quadratic approximation proposed by Wooldridge (2002) is appropriate, it turns out that this assumption is clearly rejected by Härdle 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.

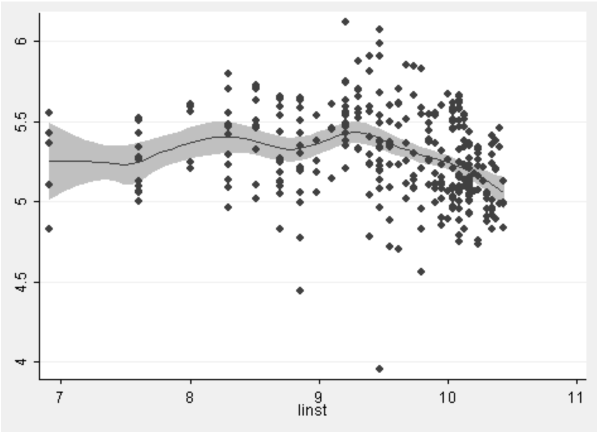

Figure 2. Nonlinear link between the price and interstate distance (in logs)

The two Stata outputs below summarize results of the Härdle 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 because those results 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)
  (output omitted)
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
--------------------------------------------------------
Standardized Test statistic T: 2.7793574
Critical value (95%): 1.959964
Approximate P-value: 0
. use http://fmwww.bc.edu/ec-p/data/wooldridge/hprice3
. semipar lprice ldist larea lland rooms baths age, nonpar(linst) nograph
> test(3)
  (output omitted)
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
--------------------------------------------------------
Standardized Test statistic T: .96211213
Critical value (95%): 1.959964
Approximate P-value: .38
```

6 Conclusion

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

7 Acknowledgments

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

8 References

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

Lokshin, M. 2006. Difference-based semiparametric estimation of partial linear regression models. *Stata Journal* 6: 377–383.

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*. Cambridge, MA: MIT Press.

Yatchew, A. 1998. Nonparametric regression techniques in economics. *Journal of Economic Literature* 36: 669–721.

About the authors

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 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 a postdoctoral researcher at the University of Namur and the University of Orléans (France). His current research fields are applied econometrics and spatial econometrics.