



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

# THE STATA JOURNAL

## Editor

H. Joseph Newton  
Department of Statistics  
Texas A & M University  
College Station, Texas 77843  
979-845-3142; FAX 979-845-3144  
jnewton@stata-journal.com

## Associate Editors

Christopher Baum  
Boston College  
Rino Bellocco  
Karolinska Institutet, Sweden and  
Univ. degli Studi di Milano-Bicocca, Italy  
A. Colin Cameron  
University of California–Davis  
David Clayton  
Cambridge Inst. for Medical Research  
Mario A. Cleves  
Univ. of Arkansas for Medical Sciences  
William D. Dupont  
Vanderbilt University  
Charles Franklin  
University of Wisconsin–Madison  
Joanne M. Garrett  
University of North Carolina  
Allan Gregory  
Queen's University  
James Hardin  
University of South Carolina  
Ben Jann  
ETH Zurich, Switzerland  
Stephen Jenkins  
University of Essex  
Ulrich Kohler  
WZB, Berlin

**Stata Press Production Manager**

**Stata Press Copy Editor**

## Editor

Nicholas J. Cox  
Geography Department  
Durham University  
South Road  
Durham City DH1 3LE UK  
n.j.cox@stata-journal.com

Jens Lauritsen  
Odense University Hospital  
Stanley Lemeshow  
Ohio State University

J. Scott Long  
Indiana University  
Thomas Lumley  
University of Washington–Seattle  
Roger Newson  
Imperial College, London  
Marcello Pagano  
Harvard School of Public Health  
Sophia Rabe-Hesketh  
University of California–Berkeley  
J. Patrick Royston  
MRC Clinical Trials Unit, London  
Philip Ryan  
University of Adelaide  
Mark E. Schaffer  
Heriot-Watt University, Edinburgh  
Jeroen Weesie  
Utrecht University  
Nicholas J. G. Winter  
University of Virginia  
Jeffrey Wooldridge  
Michigan State University

Lisa Gilmore

Gabe Waggoner

**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 web sites, 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*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

# Difference-based semiparametric estimation of partial linear regression models

Michael Lokshin  
The World Bank  
Washington, DC  
mlokshin@worldbank.org

**Abstract.** This article describes the `plreg` command, which implements the difference-based algorithm for fitting partial linear regression models.

**Keywords:** `st0109`, `plreg`, nonparametric regression, difference-based estimator, partial linear regression

## 1 Introduction

Only in rare cases, economic theory implies a particular functional form for an empirical model specification. Incorrect parameterization of the regression equation might result in inconsistent estimates. Sometimes, the researcher might feel more confident about the functional form of some parts of the regression equation but be less confident about the form of the other parts. Combining the parametric and nonparametric techniques to yield the semiparametric regression model could then help obtain consistent estimates of the parameters of interest.

In this article, I describe implementing the difference-based algorithm to fit the partial linear regression model. The econometric problem of fitting a partial linear model arises in a variety of settings. For example:

- Yatchew (1997) estimates the relationship between variable costs of distributing electricity per customer as a nonlinear function of the scale of operation as measured by the number of customers. The other control variables in the model include measures of customer density, remaining life of distribution assets, and a proxy for local wage rates.
- Yatchew (1998) applies the partial linear regression technique to estimating the hedonic price of housing attributes. Parametric variables include lot size, area of living space, and presence of various amenities. The location effect, which has no natural parametric specification, is incorporated nonparametrically.
- Mesnard and Ravallion (2001) estimate the effect of wealth on business startups among migrants returning to their home country, Tunisia. The paper tests for nonlinear wealth effects on the transition to self-employment, consistent with the argument that the extent of aggregate business activity in the economy depends on the distribution of wealth.

## 2 Methods

Consider a semiparametric regression,

$$y_i = f(z_i) + x_i\beta + \epsilon_i \quad (1)$$

where  $z$  is a random variable,  $x$  is a  $p$ -dimensional random variable,  $E[y|x, z] = f(z) + x\beta$ , and  $\epsilon_i$  is an independently and identically distributed mean-zero error term, such that  $\text{Var}[y|x, z] = \sigma_\epsilon^2$ . The function  $f$  is a smooth, single-valued function with a bounded first derivative. In this model, the parametric ( $x\beta$ ) and nonparametric [ $f(z)$ ] parts are additively separable.

Following the approach suggested by Yatchew (1997), to fit the partial linear model (1), we first rearrange (sort) the data such that  $z_1 < z_2 < \dots < z_T$ , where  $T$  is the number of observations in the sample. Then the first difference of (1) results in the following:

$$\begin{aligned} \{y_{i(n)} - y_{i(n-1)}\} &= [f\{z_{i(n)}\} - f\{z_{i(n-1)}\}] \beta \{x_{i(n)} - x_{i(n-1)}\} + \\ &\quad \epsilon_{i(n)} - \epsilon_{i(n-1)} \quad n = 2, \dots, T \end{aligned} \quad (2)$$

When the sample size increases,  $f\{z_{i(n)}\} - f\{z_{i(n-1)}\} \rightarrow 0$  because the derivative of  $f$  is bounded. Under standard assumptions, (2) could be estimated by ordinary least squares (OLS). The vector of estimated parameters  $\hat{\beta}_{\text{diff}}$  has the approximate sampling distribution

$$\hat{\beta}_{\text{diff}} \rightarrow N\left(\beta, \frac{1}{T} \frac{1.5\sigma_\epsilon^2}{\sigma_u^2}\right)$$

where  $\sigma_u^2$  is the conditional variance of  $x$  given  $z$ . The error term in (2) has a first-order moving average structure, thus reducing efficiency of the OLS estimator. The efficiency could be improved by using higher-order differences (Yatchew 1997). The generalization of (2) for the  $m$ th-order differencing can be expressed as

$$\sum_{j=1}^m d_j y_{i-j} = \beta \left( \sum_{j=1}^m d_j x_{i-j} \right) + \sum_{j=1}^m d_j f(z_{i-j}) + \sum_{j=1}^m d_j v_{i-j} \quad (3)$$

where  $d_0, \dots, d_m$  are differencing weights satisfying the conditions

$$\sum_{j=1}^m d_j = 0 \quad \text{and} \quad \sum_{j=1}^m d_j^2 = 1 \quad (4)$$

The first condition in (4) ensures that the differencing removes the nonparametric component in (3) as the sample size increases. The second normalization condition implies that the residuals in (3) have variance of  $\sigma_u^2$ . With the optimal choice of weights, (3) could be estimated by OLS. If  $m$  is large enough, the estimator approaches asymptotic efficiency.<sup>1</sup>

---

1. The Monte Carlo simulations using the `plreg` command demonstrate that the noticeable gains in efficiency from the higher-order differencing occur on samples with 30,000 or more observations. On the smaller samples (1,000–3,000 observations), using the higher-order differencing might result in biased estimates.

Define  $\Delta \mathbf{y}$  to be the  $(T - m) \times 1$  vector with elements  $(\Delta \mathbf{y})_i = \sum_{j=1}^m d_j y_{i-j}$  and  $\Delta \mathbf{x}$  to be the  $(T - m) \times p$  matrix with elements  $(\Delta \mathbf{x})_i = \sum_{j=1}^m d_j x_{i-j}$ . Then,

$$\hat{\beta}_{\text{diff}} = (\Delta \mathbf{x}' \Delta \mathbf{x})^{-1} \Delta \mathbf{x}' \Delta \mathbf{y} \rightarrow N \left\{ \beta, \frac{1}{T} \left( 1 + \frac{1}{2m} \right) \sigma_\epsilon^2 \Sigma_{x|z}^{-1} \right\} \quad (5)$$

$$s_{\text{diff}}^2 = \frac{1}{T} \left( \Delta \mathbf{y} - \Delta \mathbf{x} \hat{\beta}_{\text{diff}} \right)' \left( \Delta \mathbf{y} - \Delta \mathbf{x} \hat{\beta}_{\text{diff}} \right) \rightarrow \sigma_\epsilon^2 \quad (6)$$

$$\hat{\Sigma}_{x|z} = \frac{1}{T} (\Delta \mathbf{x})' \Delta \mathbf{x} \rightarrow \Sigma_{x|z} \quad (7)$$

This method allows performing inferences on  $\beta$  as if there were no nonparametric component  $f$  in the model. Once  $\hat{\beta}_{\text{diff}}$  is estimated, a variety of nonparametric techniques could be applied to estimate  $f$  as if  $\beta$  were known. Formally, subtracting the estimated parametric part from both sides of (1), we get

$$y_i - x_i \hat{\beta}_{\text{diff}} = x_i (\beta - \hat{\beta}_{\text{diff}}) + f(z_i) + \epsilon_i \cong f(x_i) + \epsilon_i$$

Because  $\hat{\beta}_{\text{diff}}$  converges quickly enough to true  $\beta$ , the consistency, optimal rate of convergence, and construction of confidence intervals for  $f$  remain valid and could be estimated by standard smoothing methods.

Using estimates (5), one can perform the differencing test for the parametric specification of  $f$ . Suppose that  $g(z, \pi)$  is the known function of  $z$  and some unknown parameter  $\pi$ . We want to test the null hypothesis that  $y_i = g(z_i, \pi) + x_i \beta_p$  against the alternative hypothesis that  $y_i = f(z_i) + x_i \beta$ . Parameters  $\pi$  and  $\beta_p$  and mean square residual  $s_{\text{res}}^2$  could be obtained by estimating the parametric regression of  $y$  on  $x$  and  $z$ . Then,

$$\mathbf{V} = \sqrt{mT} (s_{\text{res}}^2 - s_{\text{diff}}^2) / s_{\text{diff}}^2 \rightarrow N(0, 1) \quad (8)$$

### 3 The plreg command

The `plreg` command uses two alternative sets of differencing weights,  $d_1, \dots, d_m$ . Optimal weights do not have analytical expressions but have been tabulated (up to  $m = 10$ ) by Hall, Kay, and Titterton (1990) and by Yatchew (1998). In contrast to the former's optimal weight sequences, Yatchew's decline in absolute values toward zero. The nonlinear function  $f$  is estimated by Stata's `lowess` procedure. `plreg` also outputs the result of a significance test on  $z$ , which is a special case of (8) where  $g(z, \pi)$  is a constant function, so that the restricted model is a linear regression function,  $y_i = \pi + x_i b$ .

### 3.1 Syntax

```
plreg depvar indepvars [if] [in], nlf(varname) [generate(newvar)
      order(#) wh level(#) collinear lowess_options]
```

where *depvar* is a dependent variable in (1) and *indepvars* is a vector of variables in a linear (parametric) portion of regression (1).

### 3.2 Options

`nlf(varname)` is required and specifies the argument of an unknown function,  $f$ .

`generate(newvar)` creates a new variable *newvar* containing the smoothed values of the argument of  $f$ . These values are estimated by the locally weighted regression using **lowess**. A corresponding graph of the estimated function  $f$  could also be output; see [R] **lowess**.

`order(#)` specifies the differencing order. Tenth-order differencing is the maximum allowed. If `order()` is not specified, the model is fitted by first-order differencing.

`wh` specifies a form of the vector of differencing weights  $d_1, \dots, d_m$ , as in (3). By default, Yatchew (1998) weights are used. If `wh` is specified, Hall, Kay, and Titterton (1990) weights are used for differencing.

`level(#)` sets the confidence level; the default is `level(95)`.

`collinear` specifies that collinear variables be kept.

*lowess\_options* control the way **lowess** generates the smoothed values for the argument of the nonlinear function.

### 3.3 Saved results

In addition to the standard results saved by **regress**, **plreg** saves the following in `e()`:

#### Scalars

<code>e(s2diff)</code>	residual variance (6)
<code>e(s2lin)</code>	variance of the residual in specification that assumes that $f$ is a constant function
<code>e(order)</code>	order of differencing
<code>e(stest)</code>	value of the test on the significance of the variable that enters (1) nonlinearly

#### Matrices

<code>e(b)</code>	matrix of coefficients of differencing (5)
<code>e(V)</code>	variance-covariance matrix of differencing (5)

### 3.4 Postestimation commands

Most postestimation commands available after **regress** are also available after **plreg**. The postestimation commands are based on the estimation of the difference (3).

## 4 Example

I illustrate using the `plreg` command by replicating the example from Yatchew (2003). Data for that example come from the survey of 81 municipal electricity distributors in Ontario, Canada, during 1932.<sup>2</sup>

The cost of distributing electricity is modeled in a simple Cobb–Douglas framework,

$$\text{tc}_i = f(\text{cust}_i) + \beta_1 \text{wage}_i + \beta_2 \text{pcap}_i + \beta_3 \text{puc}_i + \beta_4 \text{kwh}_i + \beta_5 \text{life}_i + \beta_6 \text{lf}_i + \beta_7 \text{kmwire}_i + \epsilon_i$$

where `tc` is the log of total cost per customer, `cust` is the log of number of customers, `wage` is the log of wage rate, `pcap` is the log price of capital, `puc` is the dummy variable for the public utility commissions that deliver additional services and may benefit from economy of scope, `kwh` is the log of kilowatt hours per customer, `life` is the remaining life of distribution assets, `lf` is the load factor, and `kmwire` is the log of kilometers of distribution wire per customer. The objective of the analysis is to assess scale economies in electricity distribution.

The parametric effect,  $\beta$ , is estimated by first-order differencing, as in (2). We also estimate by OLS a pure parametric specification, where the scale effect,  $f$ , is modeled with a quadratic polynomial. Stata output of the estimations using `plreg_example.dta` is shown below.

```
. use plreg_example
. plreg tc wage pcap puc kwh life lf kmwire, nlf(cust) generate(func) bwidth(1)
> nodraw
```

Partial linear regression model with Yatchew's weighting matrix

Source	SS	df	MS	Number of obs = 80		
Model	1.765078594	7	.252154085	F(7, 73)	=	12.663
Residual	1.453568962	73	.019911904	Prob > f	=	0.0000
				R-squared	=	0.5484
				Adj R-squared	=	0.5051
Total	3.219	80	.040233094	Root MSE	=	0.1411

tc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wage	.4484555	.3695674	1.213	0.229	-.2880912	1.185002
pcap	.458975	.0760358	6.036	0.000	.3074359	.6105141
puc	-.0856378	.042962	-1.993	0.050	-.1712609	-.0000148
kwh	-.0105118	.0879159	-.1196	0.905	-.185728	.1647045
life	-.506133	.1318116	-3.84	0.000	-.7688332	-.2434328
lf	1.25216	.4595468	2.725	0.008	.3362849	2.168036
kmwire	.3516307	.0943774	3.726	0.000	.1635368	.5397245

Significance test on cust: V = 5.757 P>|V| = 0.000

2. I use the data for this example with the permission of Dr. Yatchew. You can download the data from <http://www.chass.utoronto.ca/~yatchew/>.

The syntax of this command does not specify an option for the order of differencing using the default first-order differencing; `bwidth(1)` is an option of the Stata `lowess` command, which determines the bandwidth for nonparametric smoothing.

The significance test of the variable (`cust`) that enters the specification nonlinearly (7) indicates that the log of number of customers is highly significant ( $p$ -value of 0.000). The estimation of the fully parametric model with a quadratic polynomial of the log of number of customers shows that although the effect of exogenous variables is qualitatively similar between these two specifications, the magnitudes of some coefficients are different. For example, the effect of log wage on log of total cost per customer declines from 0.83 in the fully parametric model to 0.45 in the partial linear model estimation.

```
. regress tc cust custsq wage pcap puc kwh life lf kmwire
```

Source	SS	df	MS	Number of obs = 81		
Model	2.76114864	9	.306794293	F( 9, 71) = 12.76		
Residual	1.70734029	71	.024047046	Prob > F = 0.0000		
				R-squared = 0.6179		
				Adj R-squared = 0.5695		
Total	4.46848893	80	.055856112	Root MSE = .15507		

tc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cust	-.832789	.1749502	-4.76	0.000	-1.18163	-.4839481
custsq	.0402137	.0091974	4.37	0.000	.0218746	.0585529
wage	.8325415	.32466	2.56	0.012	.1851878	1.479895
pcap	.5620181	.0741125	7.58	0.000	.414242	.7097941
puc	-.0705723	.0388506	-1.82	0.074	-.1480383	.0068937
kwh	-.0174608	.0889375	-0.20	0.845	-.1947972	.1598756
life	-.602922	.1192685	-5.06	0.000	-.8407367	-.3651073
lf	1.243992	.4343841	2.86	0.005	.3778543	2.110129
kmwire	.4452568	.085974	5.18	0.000	.2738295	.616684
_cons	2.750979	2.138662	1.29	0.203	-1.513392	7.01535

Using these two estimations, we can conduct a test of quadratic versus nonparametric scale effect. Substituting corresponding values into (8), we get  $V = \sqrt{81}(0.240 - 0.199)/0.199 = 1.854$ , which corresponds to a  $p$ -value of 0.032.

`plreg` uses the Stata `lowess` routine to generate the nonparametric smoothing of nonlinear function  $f$ . Figure 1 illustrates the nonparametric and fully parametric estimates of the return to scale in electricity distribution plotted against the log of the total number of customers. Quadratic specification fits the data closely to the nonparametric specification.



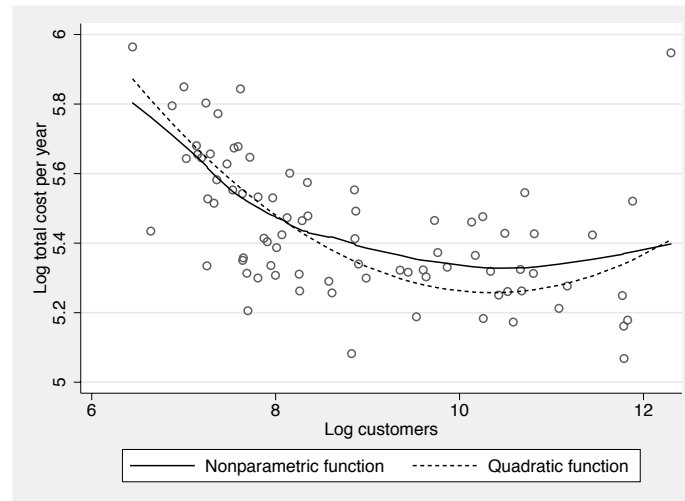


Figure 1: Electricity distribution plotted against the log of the total number of customers

## 5 References

- Hall, P., J. Kay, and D. Titterton. 1990. Asymptotically optimal difference-based estimation of variance in non-parametric regression. *Biometrika* 77: 521–528.
- Mesnard, A., and M. Ravallion. 2001. Is inequality bad for business? A nonlinear microeconomic model of wealth effect on self-employment. The World Bank, Washington, DC, Policy Research Working Paper No. WPS2527.  
[http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2001/02/10/000094946\\_01012705513588/Rendered/PDF/multi\\_page.pdf](http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2001/02/10/000094946_01012705513588/Rendered/PDF/multi_page.pdf).
- Yatchew, A. 1997. An elementary estimator of the partial linear model. *Economic Letters* 57: 135–143.
- . 1998. Nonparametric regression techniques in economics. *Journal of Economic Literature* 36: 669–721.
- . 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge: Cambridge University Press.

### About the author

Michael Lokshin is a senior economist at the Research Department of the World Bank.