

Semi-parametric difference-based estimation of partial linear regression models

Michael Lokshin
The World Bank
mlokshin@worldbank.org

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

Keywords: Non-parametric 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. The incorrect parameterization of the regression equation might result in inconsistent estimates. In some cases, 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. Then combining the parametric and non-parametric techniques to yield the semi-parametric regression model could help obtain the consistent estimates of the parameters of interest.

In this article we describe the implementation of the difference-based algorithm to estimate the partial linear regression model. The econometric problem of estimating 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 non-linear 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 estimation of 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 start-ups among the migrants returning to their home country, Tunisia. The paper tests for non-linear 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 semi-parametric regression:

$$y_i = f(z_i) + x_i\beta + \varepsilon_i \quad (2.1)$$

where z is a random variable, x is a p -dimensional random variable, $E[y | x, z] = f(z) + x\beta$, and ε_i is i.i.d. mean-zero error term, such that $Var[y | x, z] = \sigma_\varepsilon^2$. The function f is a smooth, single valued function with a bounded first derivative. In this model the parametric ($x\beta$) and non-parametric ($f(z)$) parts are additively separable.

Following the methodology suggested by Yatchew (1997), to estimate the partial linear model (2.1) we first rearrange (sort) the data in such a way 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:

$$(y_{i(n)} - y_{i(n-1)}) = (f(z_{i(n)}) - f(z_{i(n-1)})) + \beta(x_{i(n)} - x_{i(n-1)}) + \varepsilon_{i(n)} - \varepsilon_{i(n-1)} \quad n = 2, \dots, T \quad (2.2)$$

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, equation (2.2) could be estimated by the ordinary least squares. The vector of estimated parameters $\hat{\beta}_{diff}$ has the approximated sampling distribution:

$$\hat{\beta}_{diff} \rightarrow N\left(\beta, \frac{1}{T} \frac{1.5\sigma_\varepsilon^2}{\sigma_u^2}\right) \quad (2.3)$$

where σ_u^2 is conditional variance of x given z . The error term in (2.2) has an MA(1) structure, thus reducing efficiency of the OLS estimator. The efficiency could be improved by using higher order differences (Yatchew 1997). The generalization of (2.2) for the m th-order differencing can be expressed as:

$$\sum_{j=1}^m d_j y_{i-j} = \beta \left(\sum_{j=1}^m d_j x_{i-j} \right) + \sum_{j=1}^m d_j f(z_{i-j}) + \sum_{j=1}^m d_j v_{i-j} \quad (2.4)$$

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

$$\sum_{j=1}^m d_j = 0 \quad \text{and} \quad \sum_{j=1}^m d_j^2 = 1 \quad (2.5)$$

The first condition in (2.5) ensures that the differencing removes the non-parametric component in (2.4) as the sample size increases. The second normalization condition implies that the residuals in (2.4) have variance of σ_u^2 . With the optimal choice of weights equation (2.4) could be estimated by OLS. By selecting m sufficiently large, the estimator approaches asymptotic efficiency.

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:

$$\hat{\beta}_{diff} = (\Delta \mathbf{x}' \Delta \mathbf{x})^{-1} \Delta \mathbf{x}' \Delta \mathbf{y} \rightarrow N\left(\beta, \frac{1}{T} \left(1 + \frac{1}{2m}\right) \sigma_\varepsilon^2 \sum_{x|z}^{-1}\right) \quad (2.6)$$

$$s_{diff}^2 = \frac{1}{T} (\Delta \mathbf{y} - \Delta \mathbf{x} \hat{\beta}_{diff})' (\Delta \mathbf{y} - \Delta \mathbf{x} \hat{\beta}_{diff}) \rightarrow \sigma_{\varepsilon}^2 \quad (2.7)$$

$$\hat{\Sigma}_{x|z} = \frac{1}{T} (\Delta \mathbf{x})' \Delta \mathbf{x} \rightarrow \Sigma_{x|z} \quad (2.8)$$

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

$$y_i - x_i \hat{\beta}_{diff} = x_i (\beta - \hat{\beta}_{diff}) + f(z_i) + \varepsilon_i \cong f(x_i) + \varepsilon_i \quad (2.9)$$

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

Using estimates (2.6) it is possible to perform the differencing test for the parametric specification of f . Suppose $g(z, \pi)$ is the known function of z and some unknown parameter π . 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 π and β_p and mean square residual s_{res}^2 could be obtained by estimating the parametric regression of y on x and z . Then:

$$V = \sqrt{mT} (s_{res}^2 - s_{diff}^2) / s_{diff}^2 \rightarrow N(0,1) \quad (2.10)$$

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 et. al., (1990) and by Yatchew (1998). In contrast to Hall's et. al., Yatchew's optimal weight sequences declines in absolute values toward zero. The non-linear 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 (2.10) 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` is implemented as a **Stata** ado file. The generic syntax for the command is:

```
plreg depvar varlist, [if exp] [in range] , nlf(varname)
      [generate(newvar) order(#) WH level(#) lowess_options]
```

where `depvar` is a dependent variable in equation (1). `varlist` is a vector of variables in a linear (parametric) portion of regression (1).

3.2 General options

`nlf(varname)` specifies a non-optional argument of an unknown function f .

`generate(newvar)` in an optional argument that creates a new variable `newvar` containing the smoothed values of the argument of f . These values are estimated by the locally weighted regression using **Stata** procedure `lowess`. Corresponding graph of the estimated function f could also be outputted, see help for `lowess` procedure in **Stata**.

`order(#)` is an optional argument to specify the differencing order. Maximum 10th-order differencing is allowed. If `order(#)` is not specified, the model is estimated by the first order differencing.

`WH` is an optional argument that specifies a form of the vector of differencing weights d_1, \dots, d_m , as in (2.4). By default, Yatchew (1998) weights are used. If `WH` is specified Hall et. al. (1990) weights are used for differencing.

`level` in an optional argument to set a confidence level; default is `level(95)`.

`lowess_options` options to control the way `lowess` procedure generates the smoothed values for the argument of non-linear function.

3.3 Saved results

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

Scalars

- `e(s2diff)` residual variance (2.7)
- `e(s2lin)` variance of the residual in specification that assumes that f is a constant function
- `e(order)` order of differencing
- `e(stest)` value of the test on the significance of the variable that enters (2.1) nonlinearly

Matrices

- `e(b)` matrix of coefficients of differencing equation (2.6)
- `e(V)` variance-covariance matrix of differencing equation (2.6)

3.4 Post-estimation commands

Most of post-estimation commands available for `regress` are also available for `plreg`. The post-estimation commands are based on the estimation of the difference equation (2.4).

4 Example

We illustrate the use of the `plreg` command by replicating the example from Yatchew (2003). Data for that example comes from the survey of 81 municipal electricity distributors in Ontario, Canada during 1993¹.

The cost of distributing electricity is modeled in a simple Cobb-Douglas framework:

$$tc_i = f(cust_i) + \beta_1 wage_i + \beta_2 pcap_i + \beta_3 PUC_i + \beta_4 kwh_i + \beta_5 life_i + \beta_6 lf_i + \beta_7 kmwire_i + \varepsilon_i \quad (4.1)$$

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 f is estimated by the first-order differencing, as in (2.2). We also estimate by OLS a pure parametric specification where the scale effect f is modeled with a quadratic polynomial. **Stata** output of these estimations using the dataset `plreg_example.dta` is shown below.

```
. use plreg_example, clear;
. plreg tc wage pcap puc kwh life lf kmwire, nlf(cust) gen(func) bw(1);
```

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: $V = 5.757$ $P>|V| = 0.000$

The significance test of the variable ($cust$) that enters the specification non-linearly (2.8) indicates that the log of number of customers is highly significant (P-value of 0.000). The

¹ The data for that example is used with a permission of Dr. Yatchew and could be downloaded from <http://www.chass.utoronto.ca/~yatchew/>

estimation of the fully parametric model with a quadratic polynomial of the log of number of customers shows that while qualitatively the effect of exogenous variables is 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.

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

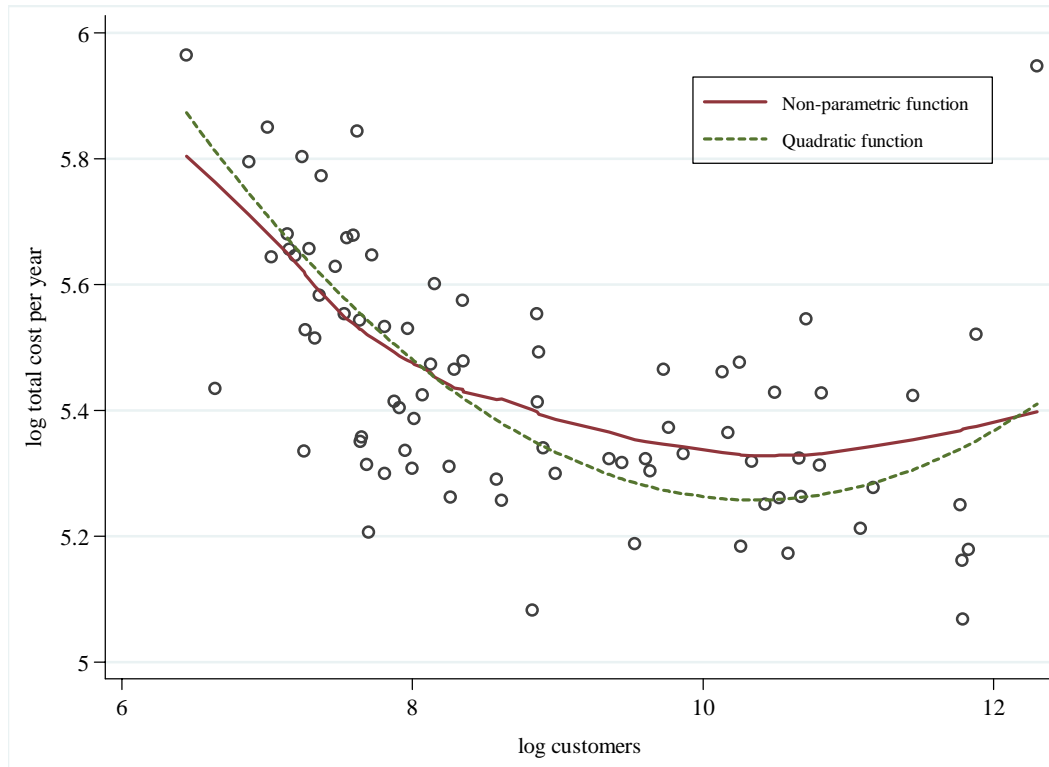
Source	SS	df	MS	Number of obs = 81		
Model	2.76114864	9	.306794293	F(9, 71)	=	12.76
Residual	1.70734029	71	.024047046	Prob > F	=	0.0000
-----				R-squared	=	0.6179
Total	4.46848893	80	.055856112	Adj R-squared	=	0.5695
-----				Root MSE	=	.15507

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 equation (2.8) we get:

$$V = \sqrt{81}(0.240 - 0.199)/0.199 = 1.854, \text{ that corresponds to P-value of } 0.032.$$

plreg utilizes the Stata routine lowess to generate the non-parametric smoothing of nonlinear function f . Figure 1 illustrates the non-parametric 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 non-parametric specification.



5 References

- Hall, P., Kay, J., and D. Titterton (1990) "Asymptotically Optimal Difference-based Estimation of Variance in Non-parametric regression," *Biometrika*, Vol. 77(3): 521-528
- Mesnard A., and M. Ravallion (2001) "Is inequality bad for business? A Nonlinear Microeconomic Model of Wealth Effect on Self-Employment," Policy Research Working Paper #2527, The World Bank, Washington DC
- Robinson, P., (1988) "Root-N-Consistent Semi-Parametric Regression," *Econometrica*, Vol. 56: 931-54
- Yatchew, A., (1997) "An Elementary Estimator of the Partial Linear Model," *Economic Letters*, Vol. 57: 135-43
- Yatchew, A., (1998) "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, Vol. 36(2): 669-721
- Yatchew, A., (2003) *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press, Cambridge, UK