Varying Coefficient Models in Stata

by
Fernando Rios-Avila
Levy Economics Institute

Introduction

Non-parametric regressions are a powerful statistical tool that can be used to model relationships between dependent and independent variables with minimal assumptions on the underlying functional forms. However, these types of models have two weaknesses:

1. Added flexibility creates curse of dimensionality
2. Procedures for model selection, in particular cross-validation, are computationally intensive in large samples.

An alternative is to use semiparametric regression modeling combining the flexibility of non-parametric with the structure of standard models.

A set of commands are introduced that aim to calibrate, estimate and visualize semiparametric model known as varying coefficient models, with a single smoothing variable.

Background

Standard semiparametric and non-parametric methods are in principle easy to implement. However, model selection (choice of Bandwidth) is computationally intensive and few commands are available for their implementation [making this methods less common to use].

Up to Stata 14:

Limited capabilities [`lpoly, lowess, fp[fracpoly], mfp`],

More flexibility from community contributed commands (`[xt]semipar, plreg, sml, pspline, bspline, mvrs`)

Stata 15 - game changer: **`npregress`** [full nonparametric models].

$$y = h(x_1, x_2, x_3, z)$$

`npregress` estimates local conditional weighted averages for every observed point in the sample. This implies a rapid decline in the effective number of observations used for each local average: more flexibility with more constraints.

Middle Ground: Varying Coefficient Model

A set of commands are proposed to estimate a particular Semiparametric model, that combines the structure of linear regression models, with the flexibility of non-parametric models: a varying coefficient model (Hastie and Tabshiran, 1993).

$$y = \beta_0(z) + \beta_1(z)x_1 + \beta_2(z)x_2 + \beta_3(z)x_3 + e$$

This model assumes all coefficients in the model are smooth functions of a single smoothing variable, z. They can be estimated using local linear kernel weighted regressions (Li and Racine, 2007), given a choice of kernel function K and bandwidth h:

$$\boldsymbol{\hat{\beta}(z), \hat{\delta}(z) = argmin} \sum (\boldsymbol{y_i} - \boldsymbol{X'}\beta(z) - (\boldsymbol{Z} - \boldsymbol{z})\boldsymbol{X'}\delta(z))^2 * \boldsymbol{K}\left(\frac{\boldsymbol{Z-z}}{\boldsymbol{h}}\right)$$

Method: Model Selection (h): vc_bw and and vc_bwalt

Non/semi-parametric model are sensitive to choice of bandwidth "h" (less so to the choice of Kernel function). The process implies a trade-off between variance and bias.

The commands `vc_bw` (*NewtonRampson*) and `vc_bwalt` (*NelderMead*) can be used to select the appropriate bandwidth by minimizing the following function:

$$CV_{loo}(h) = \sum_{i \in B} \omega(z_i)\left(\boldsymbol{y_i} - \boldsymbol{X'}\widehat{\beta_{-i}}(z_i, h) - (\boldsymbol{Z} - \boldsymbol{z_i})\boldsymbol{X'}\widehat{\delta_{-i}}(z_i, h)\right)^2$$

Syntax:

```
vc_bw[alt] [varlist], vcoeff(z) [knots(#)  bwi(#)
trisample(varname) kernel(kfunc)]
```

Note:

- By default, $\hat{\beta}(z,h)$ is estimated at every step for all possible values of z. To increase speed process, use **knots(#),** and request fewer number of regressions to be estimated.
- The CV procedure is performed over equal width blocks of Z and block average Z.
- The LOO errors use `cv_regress`, using the leverage statistic for the local OLS.

Method: Model Estimation vc_reg and vc_bsreg

The estimation of the model can be obtain focusing on specific points of reference over z.

Syntax:

```
vc_reg varlist, vcoeff(z) [bw(#) kernel(kfunc) k(#)
klist(numlist) cluster(vname) robust hc2 hc3] [BS options]
```

Note:

- If not specified `bw(#)  kernel(kfunc)` are taken from `vc_bw/vc_bwalt`
- `k(#)` indicates a fixed number of equidistance points for the estimation of the models
- `klist(numlist)` indicates specific points for the estimation of the models
- Weights estimation uses `_gkweights`. (kwgt=K(.)/K(0)).
- `vc_reg` reports SE from OLS (iweights if robust not used) (Li & Racine,2010).
- `vc_bsreg` reports Bootstrapped SE, similar to `npregress`.

Method: Model prediction and Evaluation

Syntax:

```
vc_predict varlist, vcoeff(z) [bw(#) kernel(kfunc) xb(vname)
looerr(vname) lev(vname) stest]
```

This command can get predictions , predictions of LOO errors (For CV), and predictions of Leverage (Degrees of Freedom). Also reports E(Kobs) and R2

$$CV = log\left(\frac{\sum \hat{e}_{-i}^2}{N}\right);\ DF = \sum lev_i;\ E(Kobs) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{N}\frac{K\left(\frac{x_j - x_i}{h}\right)}{K(0)};R^2 = 1 - \frac{SSE}{SST}$$

Specification test compare the Semi parametric model to :

$$\boldsymbol{y = \beta_1 X + \gamma_z z + \beta_2 X * Z + \beta_3 X * Z^2 + \beta_4 X * Z^3 + e}$$

Using F test. $F = \frac{(SSR_p - SSR_{vc})/(DF_{vc} - DF_p)}{SSR_{vc}/(N - DF_{vc})} \sim F^2$ (Hastie and Tibshirani 1990)

Method: Model Visualization

When more than 1 point of reference is used with `vc_reg, vc_graph` can visualize main coefficients $\beta(z)$, or their changes $\delta(z)$

Syntax:

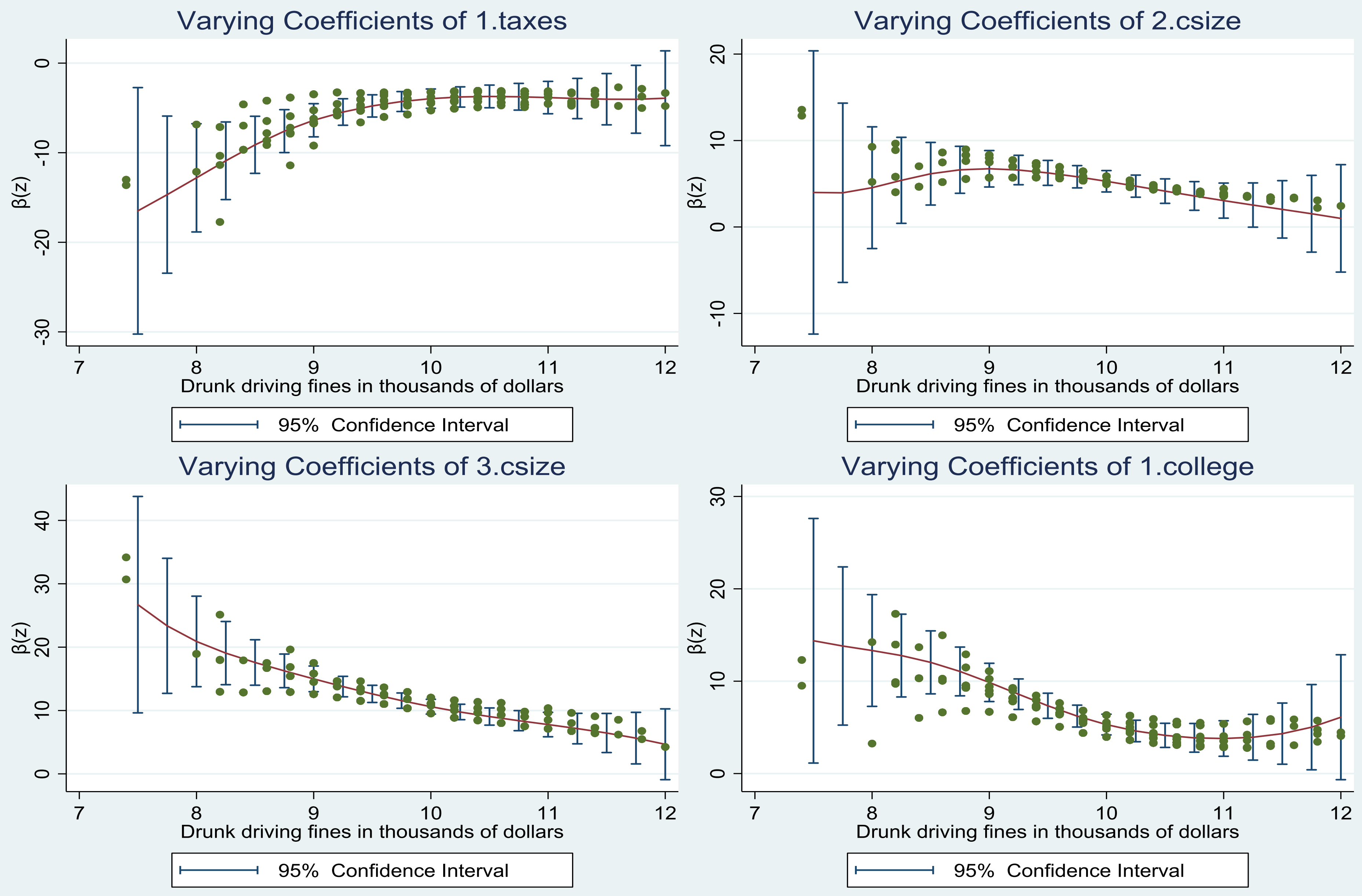
```
vc_graph [Selected varlist],  [ci(#) constant delta
xvar(H(z)) graph(name)]
```

`vc_graph` plots the coefficients of the selected variables. Each graph is stored in memory with the name "graph#".

Note:

- `xvar(.)` can change the scale of the running variable for the Graph. Say $H(z) = \log(z)$. H(z) can display the figure in log scale. Uses `vt_xtoy`.
- Useful if CV and model estimation is done for a transformation of z.
- $hz = \hat{F}(z)$ may help where Z is sparse.

Illustration: Citations and fines (DUI)



Varying Coefficients of 1.taxes

Varying Coefficients of 2.csize

Varying Coefficients of 3.csize

Varying Coefficients of 1.college

Conclusions and Discussion

The commands proposed aim to facilitate the estimation of varying coefficient models. These models can be used for analysis and visualization tool for heterogeneous effects across a selected variable. Further work required:

- Increase Speed of Cross validation (using C or Mata).
- Additional specification tests.
- Allow for discreet running groups.
- Explore theoretical properties of using transformations for local linear estimation.
- Explore theoretical properties of estimator for the Endogenous selection.

Contact

Fernando Rios-Avila
Levy Economics Institute
Email: friosavi@levy.org
Website:www.levy.org
Phone:845-758-7719

References

Hastie, T., and R. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T., and R. Tibshirani. 1993. Varying-Coefficient Models. *Journal of the Royal Statistical Society* Series B (Methodological), 55(4): 757–796.

Li Q. and J. S. Racine. 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press: United Kingdom.

Li Q. and J. S. Racine. 2010. Smooth Varying-Coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory* 26(6):1607-1637.