Nonparametric Econometrics in R

Philip Shaw

Fordham University

November 17, 2011

The NP Package

- R seems to be the only software package that has a list of comprehensive nonparametric routines
- A variety of nonparametric econometrics can be run under the NP package
- Much of the code was written and is maintained by Jeffrey Racine, McMaster University

Cereal Data

- Scanner data on price (dpfwgtavgppoz) and quantity (lndfshr) of cereal across five brands
- n=10450
- Includes cereal characteristics such as fiber, calories, sugar, etc.

Density Estimation

Suppose you are interested in estimating the unconditional density function:

$$\hat{f}(x) = \frac{1}{hn} k(\frac{x_i - x}{h}) \tag{1}$$

where:

$$k(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2} \tag{2}$$

h is referred to as the bandwidth

Density Estimation

For various types of data:

$$K_\gamma(x_i, x) = W_h(x_i^c, x^c) L(x_i^d, x^d, \lambda) J(x_i^s, x^s, \lambda) \tag{3}$$

$$W_h(x_i^c, x^c) = \prod_j^{r_1} \frac{1}{h_j} w\left(\frac{x_j^c - x_{ji}^c}{h_j}\right) \tag{4}$$

$$L(x_i^d, x^d, \lambda) = \prod_j^{r_2} \lambda_j^{I(x_{ji}^d \neq x_j^d)} \tag{5}$$

$$J(x_i^s, x^s, \lambda) = \prod_j^{r_3} \lambda_j^{|x_{ji}^s - x_j^s|} \tag{6}$$

where x^c are continous, x^d are discrete, and x^s are discrete ordered variables.

Bandwidth Selection

For density estimation:

$$CV(h) = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{k}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i, j=1}^{n} k\left(\frac{x_i - x_j}{h}\right) \quad (7)$$

For kernel regression:

$$\gamma^{opt} = \underset{\gamma \in R^l}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_t - \hat{m}_{-i}(x_i))^2 M(x_i) \quad (8)$$

where $0 \leq M(x_i) \leq 1$.

$$AIC = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{m}(x_i)\}^2 + \frac{1 + tr(H)/n}{1 - \{tr(H) + 2\}/n} \quad (9)$$

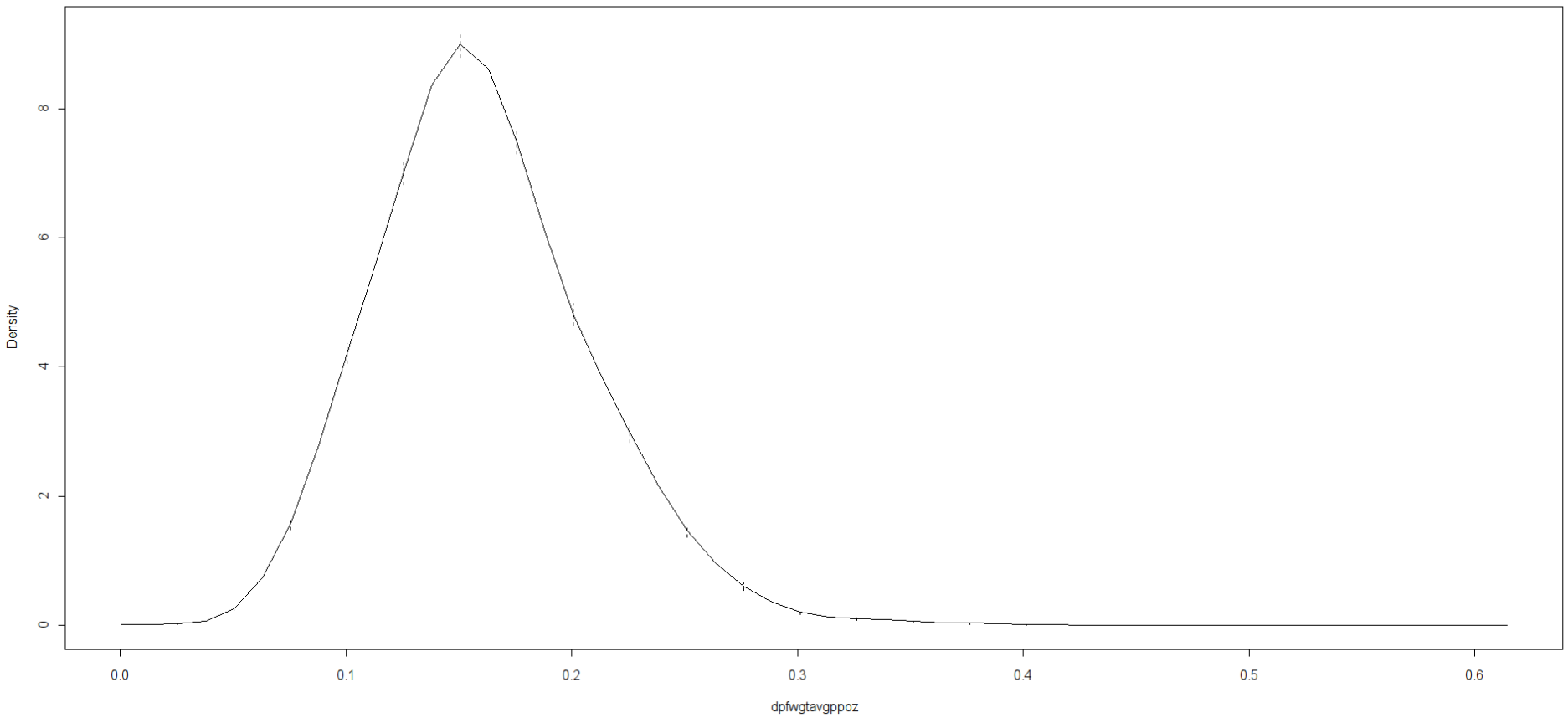
where $H_{ij} = K_{\gamma,ij} / \sum_{l=1}^{n} K_{\gamma,il}$.

The npudensbw command

```
> bw <- npudensbw(~dpfwgtavgppoz)
> bw
Data (10450 observations, 1 variable(s)):
dpfwgtavgppoz
Bandwidth(s): 0.01340203
Bandwidth Selection Method: Maximum Likelihood Cross-Validation
Formula:  dpfwgtavgppoz
Bandwidth Type: Fixed
Objective Function Value: -1.66763 (achieved on multistart 1)
Continuous Kernel Type: Second-Order Gaussian
No. Continuous Vars.: 1
```

The npudens command

```
> fhat <- npudens(bw)
> plot(fhat, plot.errors.method="bootstrap")
```



Kernel Regression

Now suppose we want to estimate the demand curve for cereal
Typically papers have assumed that demand is a linear function of price such that:

$$q = \beta_0 + \beta_1 p + u \tag{10}$$

Instead we would like to estimate the demand curve without imposing a linear assumption on the model:

$$q = f(p) + u \tag{11}$$

Under the assumption that $E(u|p) = 0$ we can consistently estimate the function $f(p)$ as:

$$E[\hat{q}|x] = \hat{f}(p) = \frac{\sum_{i=1}^{n} q_i K_\gamma(p_i, p)}{\sum_{i=1}^{n} K_\gamma(p_i, p)} \tag{12}$$

The npregbw command

For our example we can select the optimal bandwidth for the kernel regression using the following R code:

$>$ bw1 $<-$ npregbw(formula=lndfshr$\tilde{}$ dpfwgtavgppoz)

$>$ bw1

Regression Data (10450 observations, 1 variable(s)):

dpfwgtavgppoz

Bandwidth(s): 0.008413665

Regression Type: Local-Constant

Bandwidth Selection Method: Least Squares Cross-Validation

Formula: lndfshr dpfwgtavgppoz

Bandwidth Type: Fixed

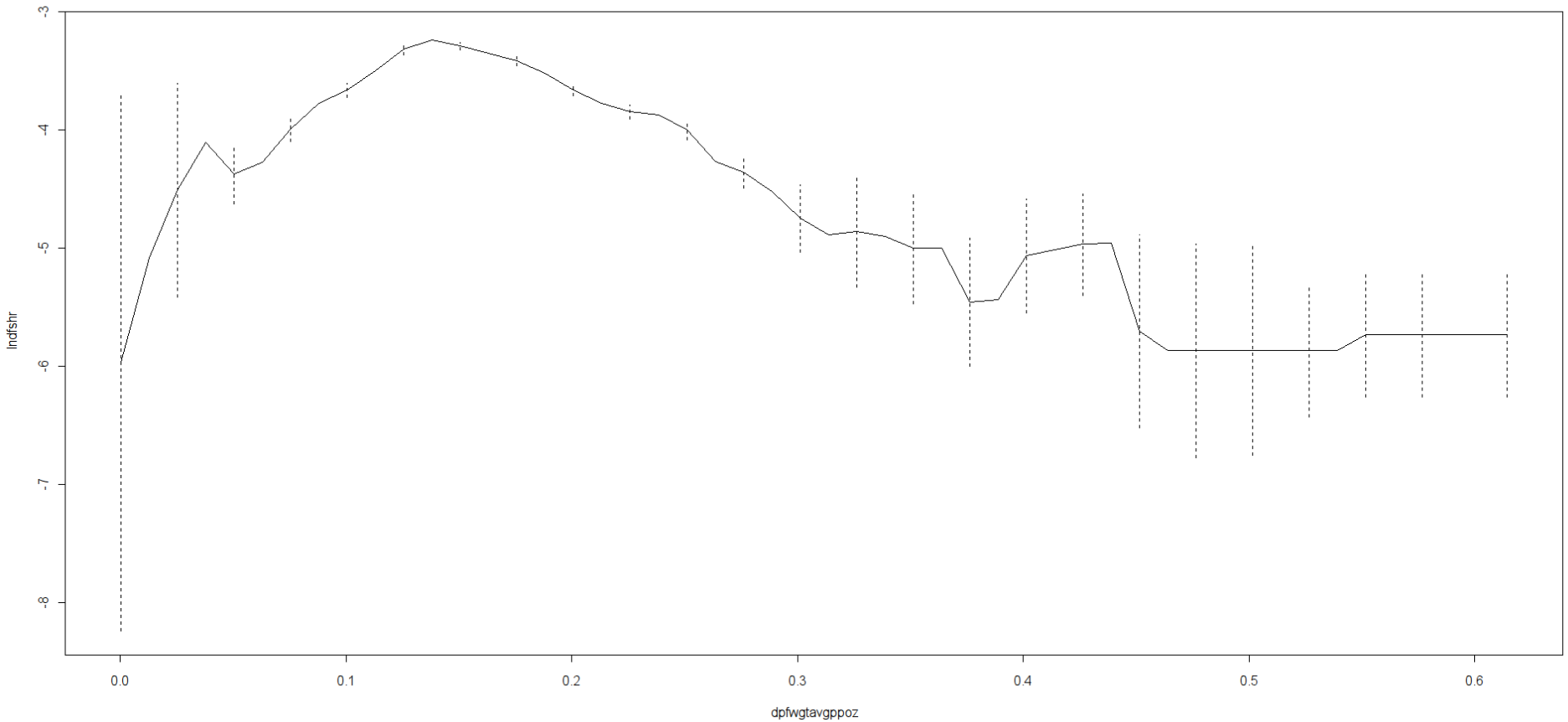
Objective Function Value: 0.963636 (achieved on multistart 1)

Continuous Kernel Type: Second-Order Gaussian

No. Continuous Explanatory Vars.: 1

The npreg command

```
> fhat1 < − npreg(bw1)
> plot(fhat1, plot.errors.method= "bootstrap" )
```



Significance Testing

Suppose we would like to test the following condition:

$$H_0: \quad E(y|x) = E(y) \tag{13}$$

versus

$$H_1: \quad E(y|x) \neq E(y) \tag{14}$$

In the linear model this test is equivalent to using a t-test on the estimated coefficient under the zero null
The problem is that this test is in general not a consistent test
Racine proposes a consistent test in the nonparametric setting

The npsigtest command

> npsigtest(bw1,boot.num=100,boot.method = "wild")
Kernel Regression Significance Test
Type I Test with Wild Bootstrap (100 replications)
Explanatory variables tested for significance:
dpfwgtavgppoz (1)
dpfwgtavgppoz
Bandwidth(s): 0.008413665
Significance Tests
P Value: dpfwgtavgppoz $< 2.22e-16^{***}$

Functional Form Testing

Suppose we would like to test the following condition:

$$H_0: \quad E(y|x) = m(x, \gamma_0), \text{ for almost all } x \text{ and for some } \gamma_0 \in B \subset R^p \tag{15}$$

versus

$$H_1: \quad E(y|x) \neq m(x, \gamma_0) \tag{16}$$

> model <- lm(lndfshr ~ dpfwgtavgppoz, x=TRUE, y=TRUE)
> model
Call:
lm(formula = lndfshr ~ dpfwgtavgppoz, x = TRUE, y = TRUE)
Coefficients:
(Intercept) dpfwgtavgppoz
-3.068 -2.841

The npcmstest command

$>$ npcmstest(model = model, xdat = dpfwgtavgppoz, ydat = lndfshr)
Consistent Model Specification Test
Parametric null model: lm(formula = lndfshr dpfwgtavgppoz, x = TRUE, y = TRUE)
Number of regressors: 1
Wild Bootstrap (100 replications)
Test Statistic Jn: 119.5579 P Value: $< 2.22e-16^{***}$

Other commands in the NP package

- Conditional density estimation
- Local linear regression
- Statistical test of distribution equality
- Instrumental variable estimation
- Semiparametric single index models