NONPARAMETRIC AND SEMIPARAMETRIC METHODS

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Overview
2. Nonparametric Methods
3. Semiparametric Methods

1. Overview

- What does “nonparametric” mean? Traditionally, it means without specifying a parametric distribution for a random variable (or variables). For example, suppose we want to estimate $\mu = E(y)$ for a random variable y. One can think of the sample average (from, say, a random sample) as a nonparametric estimator because it is unbiased and consistent for a wide range of distributions: those with $E(|y|) < \infty$. (It is also the MLE for some common distributions: Bernoulli, Poisson, normal, exponential.)
- In current usage, one rarely thinks of the sample average as a “nonparametric” estimator.

- As a practical matter, “nonparametric” now is synonymous with estimating an “infinite dimensional” feature, such as an unconditional density or a conditional mean function. (In other words, without assuming the density or regression function is in a class described by a finite number of parameters.)
- Pure nonparametric methods are still used rarely in economics, partly because the dimension of economics problems is usually “large.” (For example, many explanatory variables in regression analysis.)

- Also not clear how useful a multi-dimensional graph is in describing relationships, or have essentially unrestricted slopes in all directions.
- We usually want summary numbers, such as (average) partial effects. Flexible parametric approaches often do nicely. (We have seen several examples where average partial effects from different nonlinear models are similar.)

- Nonparametric methods are useful as descriptive devices for looking at distributions of a single variable, or a relationship between two variables.
- For example, if y and x are scalars, we can estimate $E(y|x)$ without assuming a specific functional form. But such a description rarely has a causal interpretation.
- Can even plot $E(y|x_1, x_2)$ in three dimensions. But what about, say, $E(y|x_1, x_2, x_3)$? Visualizing is difficult and even summary measures are difficult to decide on.

- Semiparametric methods are more promising and used more often. Part of the problem is parametric – depends on a finite number of parameters, but some critical feature that needs to be estimated is allowed to be "infinite dimensional."
- Most useful semiparametric methods are for estimating a finite set of population parameters but relaxing some key assumptions.

EXAMPLE

- Suppose y is a binary response and we have exogenous covariates $\mathbf{x}$.

Consider three situations.

$P(y = 1|\mathbf{x}) = \Phi(\alpha + \mathbf{x}\boldsymbol{\beta})$ where $\Phi(\bullet)$ is the standard normal cdf: parametric

$P(y = 1|\mathbf{x}) = G(\alpha + \mathbf{x}\boldsymbol{\beta})$ for unknown G:$(\bullet)$: $\mathbb{R}^K \to [0,1]$: semiparametric

$P(y = 1|\mathbf{x}) = H(\mathbf{x})$ where H: $\mathbb{R}^K \to [0,1]$: nonparametric

- The term “seminonparametric” has been used, too. It often looks a lot like nonparametric or semiparametric analysis in that there is either an infinite-dimensional function to estimate or a finite dimensional parameter vector and also something that is infinite dimensional.
- If there is no finite-dimensional parameter, seminonparametric is really nonparametric. If there is a finite-dimensional parameter, one needs to decide whether the finite-dimensional parameter is if interest – and has estimators with the usual “nice” properties – or whether one is just trying to achieve flexibility.

2. Nonparametric Methods

- Some nonparametric estimators are “automatic,” the leading example being an estimator of a cumulative distribution function.
- But density and regression estimators use various “smoothing” methods. Broadly speaking, these can be put into two categories: **global smoothing** and **local smoothing**.
- Global smoothing is usually implemented via **series estimation** (sometimes called **sieve estimation**), which is simply a flexible parametric model where the approximation gets better as the sample size increases.

- Local smoothing is implemented as a local averaging, or at least averaging where observations far away receive little weight. Kernel estimation of densities and regression functions falls into this category.
- Both global smoothing and local smoothing require the choice of either the number of terms in the series or the amount of local averaging. These must be chosen by the researcher (sometimes using established rules), or a data driven method.

Estimating the Cumulative Distribution Function

- Let x denote a random variable with cdf $F(\bullet)$, so that $F(a) = P(x \leq a)$. Notice that

$$P(x \leq a) = E\{1[x \leq a]\}$$

so that an unbiased and consistent estimator (with random sampling) is

$$\hat{F}(a) = N^{-1} \sum_{i=1}^{N} 1[x_i \leq a],$$

which is simply the fraction of the N observations that are less than or equal to a.

- $\hat{F}(a)$ is usually called the **empirical cdf**. Notice that $\hat{F}(\bullet)$ is nondecreasing but discontinuous at points represented by the data. It is a step function, continuous from the right.
- Write

$$\sqrt{N}\,[\hat{F}(a) - F(a)] \;\; = N^{-1/2}\sum_{i=1}^{N}\{1[x_i \leq a] - F(a)\}.$$

- By the CLT,

$$\sqrt{N}\,[\hat{F}(a) - F(a)] \;\; \overset{d}{\to} Normal(0, Var\{1[x_i \leq a]\})$$

- But $w_i \equiv 1[x_i \leq a]$ is just a binary variable, so

$Var(w_i) = F(a)[1 - F(a)]$.

$$\sqrt{N}\,[\hat{F}(a) - F(a)] \overset{d}{\rightarrow} Normal(0, F(a)[1 - F(a)])$$

- An asymptotic 95% CI for $F(a)$ is

$$\hat{F}(a) \pm 1.96\sqrt{\hat{F}(a)[1 - \hat{F}(a)]/N}$$

Density Estimation

- Suppose the underlying density is continuous. Can nevertheless use a **histogram estimator**. Given a set of bins – usually gotten by specifying the number of bins, obtaining the range of the data, and then dividing up the line into bins of equal size – use the fraction of observations falling into each bin as the estimated density. Can superimpose a parametric density to see how simple models fit.
- To smooth out the estimated density, generally use a **kernel estimator**.

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} k\left(\frac{x_i - x}{h} \right)$$

where $h > 0$ is called the **bandwidth** and $k(\bullet)$ is the **kernel function**.

- Typically, $k(\bullet) \;\; \geq 0$ and

$$\int_{\mathbb{R}} k(v)dv = 1$$
$$k(v) = k(-v)$$
$$\int_{\mathbb{R}} v^2 k(v)dv > 0.$$

In other words, $k(\bullet)$ is a symmetric density about zero with nonzero variance.

- Some examples are

$$k(v) = \frac{1}{2}1[-1 < v < 1] \text{ (rectangular or uniform)}$$

$$k(v) = 1 - |v|, \ -1 < v < 1 \text{ (triangular)}$$

$$k(v) = \frac{3}{4}(1 - v^2), \ -1 < v < 1 \text{ (Epanechnikov)}$$

- If $h \to 0$ and $Nh \to \infty$, can show the kernel density estimator is consistent (pointwise), that is,

$$\hat{f}(x) \xrightarrow{p} f(x)$$

for all x in the interior of the support of the distribution. See Li and Racine (2007, *Nonparametric Econometrics*).

- How to choose the bandwidth? (1) Guess and experiment; (2) Use rules based on experience or optimality for common distributions (say, suppose f is normal and minimize the mean squared error); (3) Use various data-driven methods, such as **cross validation**.

- For a broad class of densities, the “optimal” bandwidth has the form

$$h = c_0 N^{-1/5}$$

for $c_0 > 0$.

- Optimality is based on integrated mean squared error:

$$\int E[\hat{f}(x) - f(x)]^2 dx.$$

For each x, $E[\hat{f}(x) - f(x)]^2 = [Bias(\hat{f}(x))]^2 + Var(\hat{f}(x))$, so the optimal h trades off bias and variance over the range of x.

- The **normal reference rule-of-thumb** is

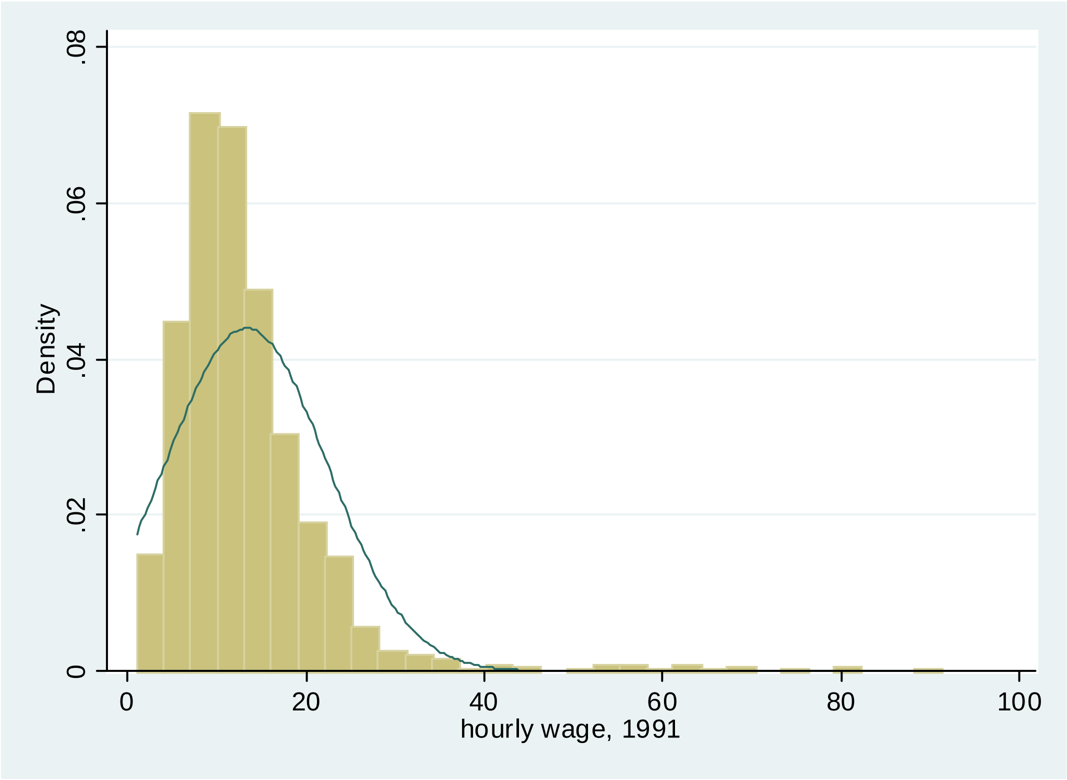
$$h = 1.06\sigma N^{-1/5},$$

which is optimal for the normal density. Have to replace $\sigma = sd(x_i)$ with its usual estimate. This is the default in Stata.

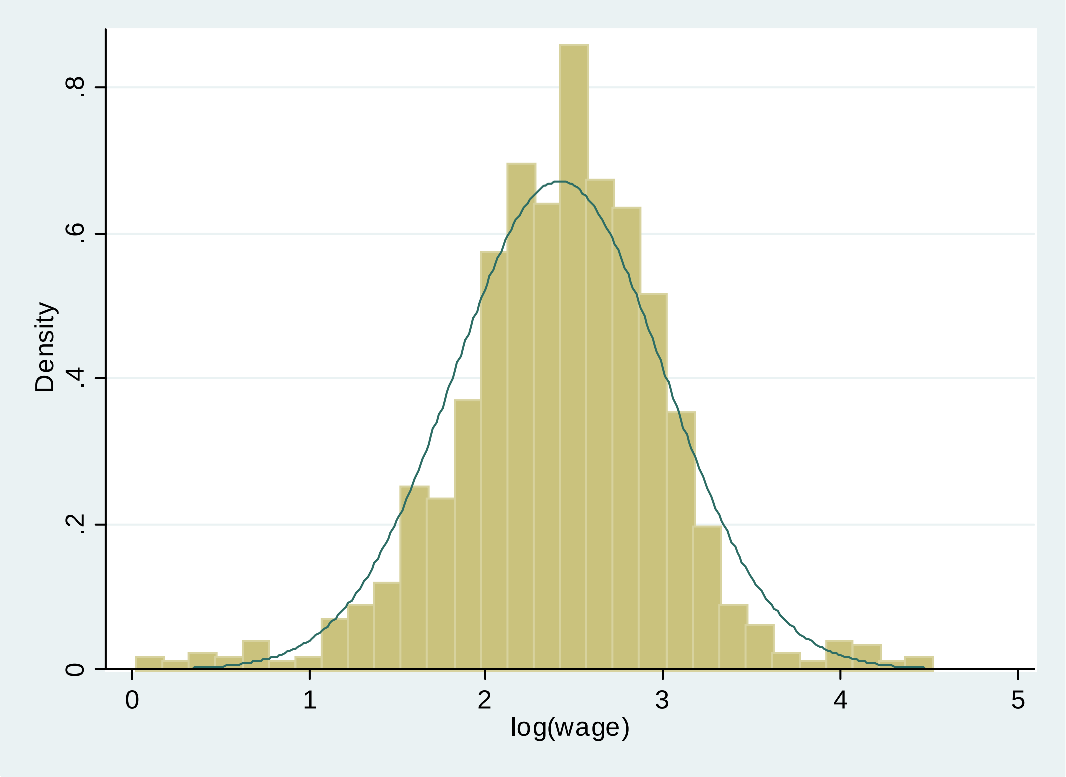
- If population density is highly skewed or multimodal, normal reference ROT can oversmooth. Look at a histogram first.

- However, using $h = 1.06\sigma N^{-1/5}$ when the distribution is nonnormal does not cause inconsistency because it satisfies the rule for consistency. We use it because it is optimal for the leading case, but the rule does not cause inconsistency across a wide class of densities.

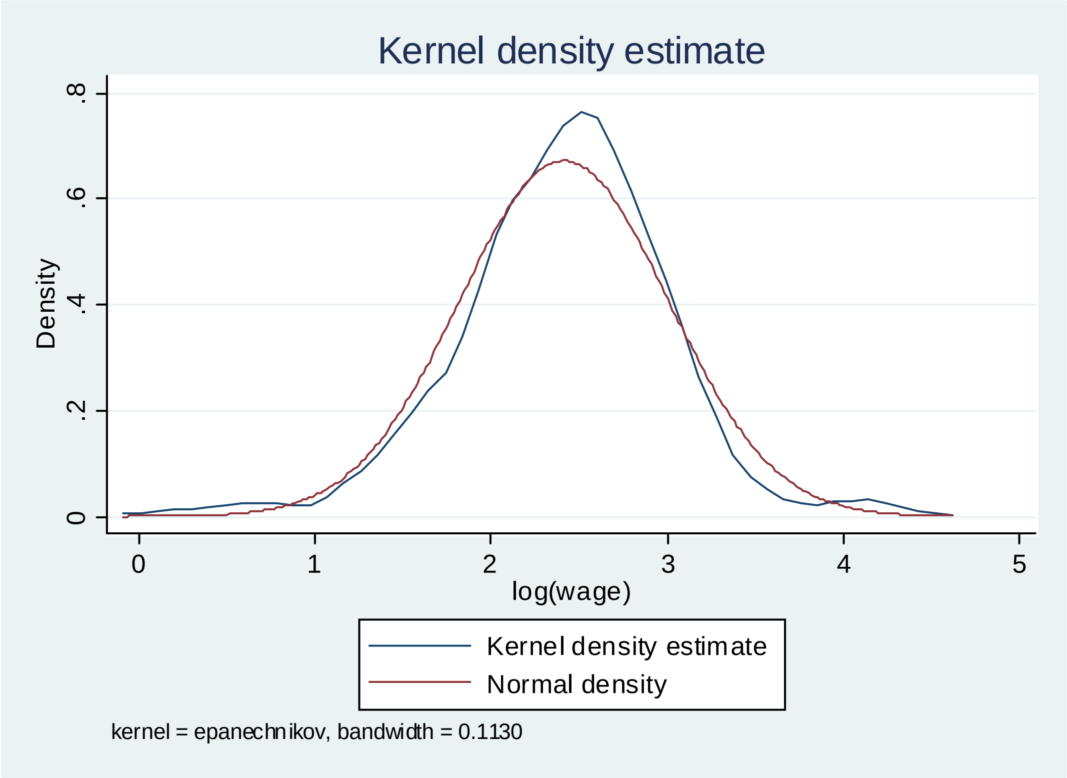
```
. use htv

. hist wage, normal
(bin=30, start=1.0235294, width=3.009523)
```



```
. hist lwage, normal
(bin=30, start=.02325686, width=.14969983)
```



```
. kdensity lwage, kernel(epan) normal
```



Regression Estimation

- Consider a simple regression model where (x_i, y_i) are random draws, and we hope to estimate $m(x) = E(y_i | x_i = x)$.

- Kernel estimators are weighted averages of the y_i. For a given value x, observations with x_i closer to x receive greater weight.

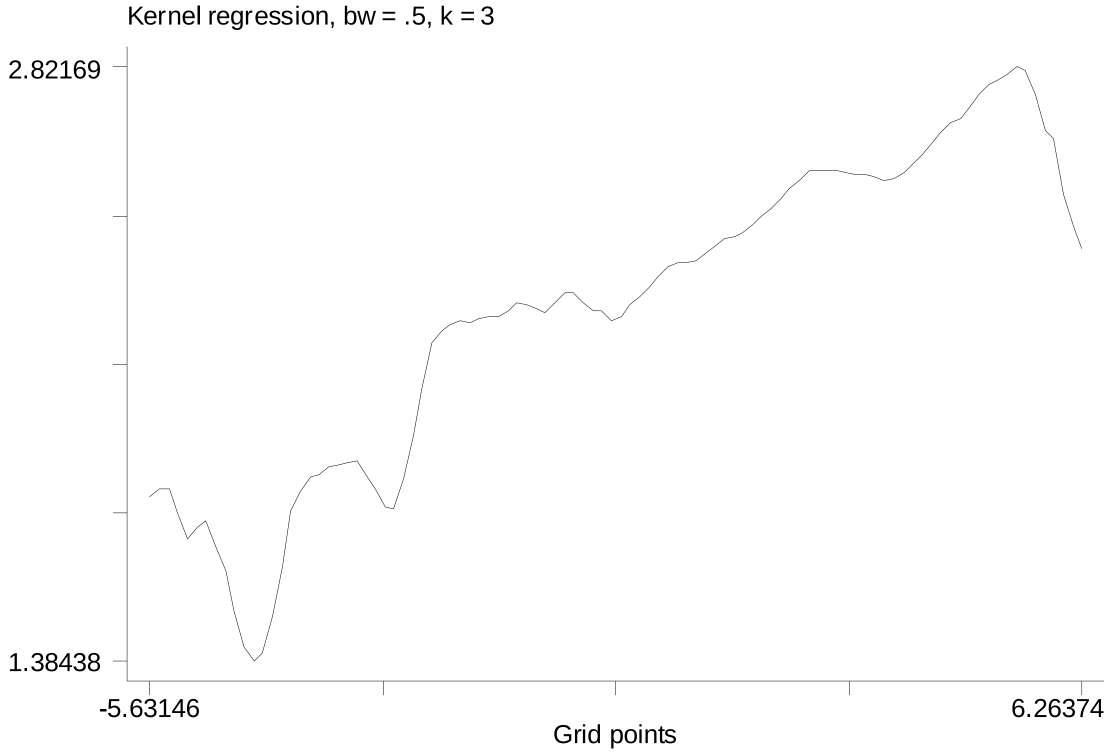
$$\hat{m}(x) = \frac{\sum_{i=1}^{N} k(\frac{x_i - x}{h}) y_i}{\sum_{i=1}^{N} k(\frac{x_i - x}{h})} \equiv \sum_{i=1}^{N} w_{N,i}(x) y_i$$

where

$$w_{N,i}(x) = \frac{k(\frac{x_i - x}{h})}{\sum_{r=1}^{N} k(\frac{x_r - x}{h})}$$

- These weights are nonnegative and sum to unity. Typically, $k(\bullet)$ is a unimodal, symmetric density about zero, so the largest weight is at for i with $x_i = x$ (if there are any such i).
- Can choose $k(\bullet)$ so that observations far enough away receive no weight. Rectangular and triangular densities have this feature.
- Kernel regression is called “local smoothing” because most of the weight is on nearby observations; what is happening far from x receives little or no weight.

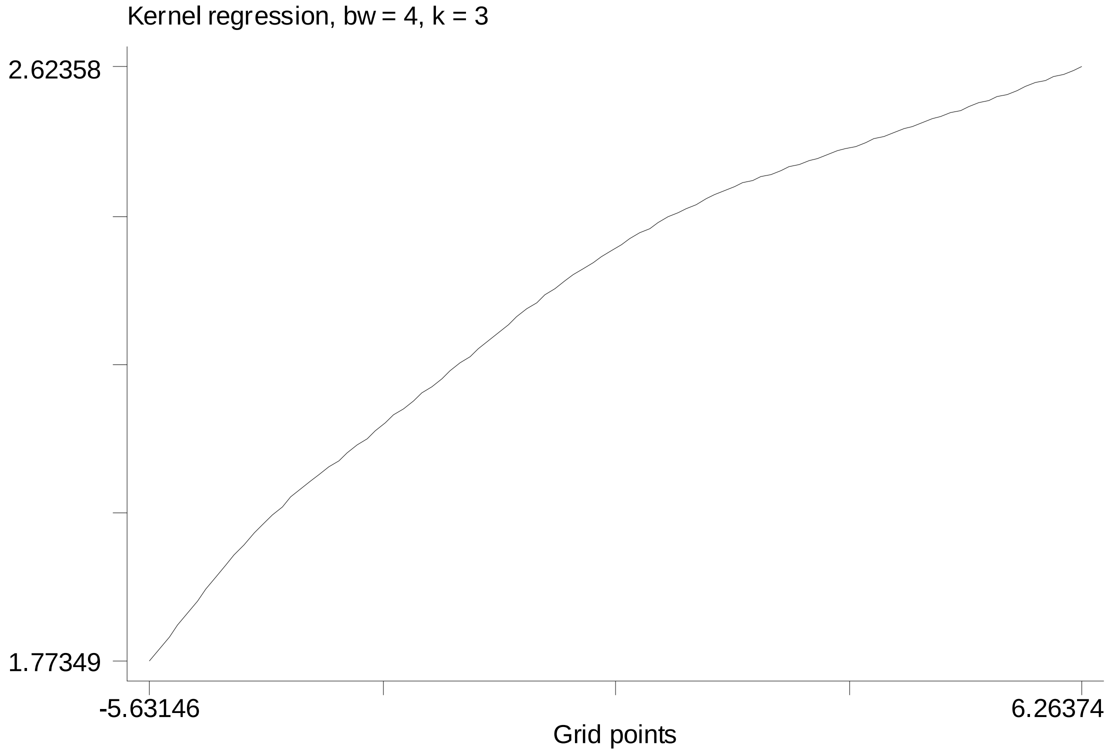
```
. kernreg lwage abil, b(.5) k(3) np(100) gen(lwageh_p5 abilg_p5)
```



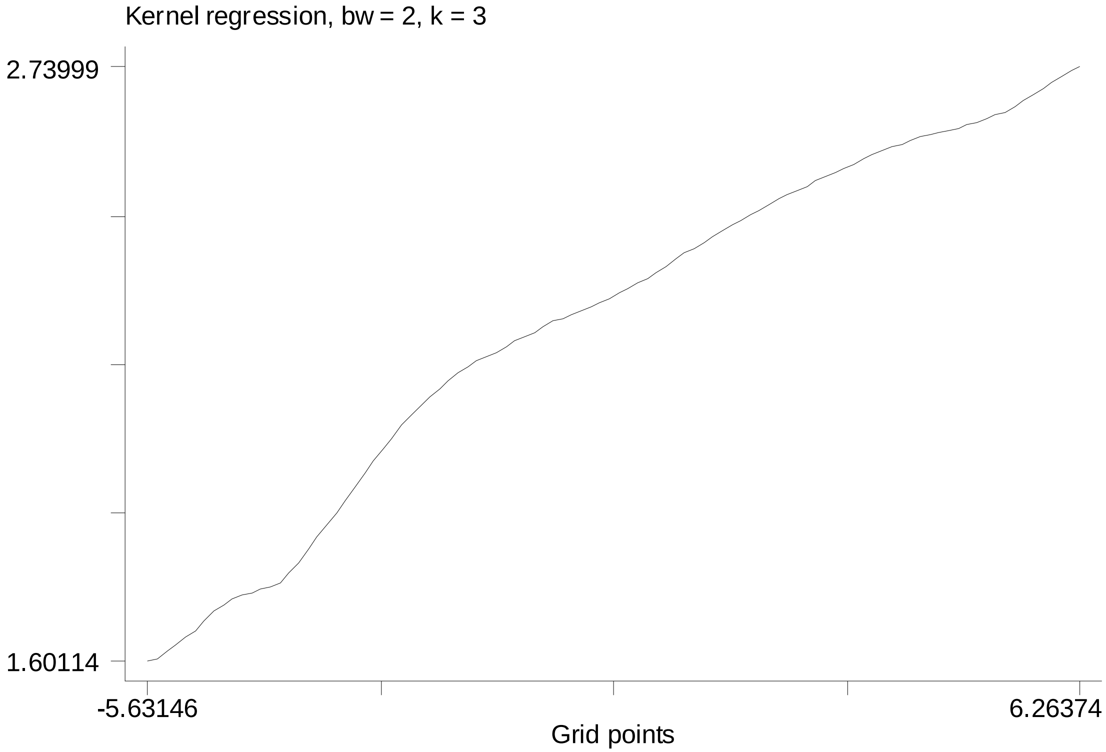
```
. list lwage abil  lwageh_p5 abilg_p5 in 1/10

     +-------------------------------------------+
     |    lwage        abil   lwageh~5    abilg_p5 |
     |-------------------------------------------|
  1. | 1.857899   -5.631463   1.779906   -5.631463 |
  2. | 1.301366   -5.468668   1.801753    -5.51131 |
  3. |  2.30092   -5.344852   1.799685   -5.391156 |
  4. | 1.277095    -4.91838   1.740911   -5.271002 |
  5. | 1.311393   -4.729329    1.67767   -5.150849 |
     |-------------------------------------------|
  6. | 2.358675   -4.686911    1.71085   -5.030695 |
  7. | 1.756499   -4.671546   1.724063   -4.910542 |
  8. | 1.965497   -4.572411   1.656987   -4.790388 |
  9. | .3856625   -4.354986   1.599717   -4.670234 |
 10. | 1.585721   -4.337047   1.504929   -4.550081 |
     +-------------------------------------------+

. * Note: Data have been sorted by ability.

. corr lwage lwageh_p5
(obs=100)

             |    lwage lwageh~5
-------------+------------------
       lwage |   1.0000
   lwageh_p5 |   0.3062   1.0000

. di .3062^2
.09375844
```

```
. reg lwage abil

      Source |       SS       df       MS              Number of obs =    1230
-------------+------------------------------           F(  1,  1228) =  190.21
       Model |  58.1036671     1  58.1036671           Prob > F      =  0.0000
    Residual |  375.115595  1228  .305468726           R-squared     =  0.1341
-------------+------------------------------           Adj R-squared =  0.1334
       Total |  433.219262  1229  .352497365           Root MSE      =  .55269

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        abil |   .0995388   .0072173    13.79   0.000     .0853792    .1136984
       _cons |   2.234976   .0204078   109.52   0.000     2.194938    2.275014
------------------------------------------------------------------------------

. gen lwageh_lin = _b[_cons] + _b[abil]*abilg_p5
(1130 missing values generated)
```

```
. corr lwage lwageh_lin
(obs=100)

             |    lwage lwageh~n
-------------+------------------
       lwage |   1.0000
  lwageh_lin |   0.3181   1.0000


. di .3181^2
.10118761

. * This R-squared looks at the fit of the linear model at the grid points chosen
. * for the kernel estimation, so it is a fair comparison. The OLS estimates
. * are not chosen to minimize the SSR at the grid points.
. kernreg lwage abil, b(4) k(3) np(100) gen(lwageh_4 abilg_4)
```



```
. kernreg lwage abil, b(2) k(3) np(100) gen(lwageh_2 abilg_2)
```



```
. corr lwage lwageh_2
(obs=100)

             |    lwage lwageh_2
-------------+------------------
       lwage |   1.0000
    lwageh_2 |   0.3030   1.0000


. di .3030^2
.091809
```

Global Smoothing of Regression

- Can use flexible linear models if the range of *y* is essentially unrestricted – such as log(*wage*).

```
. sum abil

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        abil |      1230    1.796596    2.184406  -5.631463   6.263742

. gen sabil = (abil - 1.8)/2.2

. gen sabilsq = sabil^2

. gen sabilcu = sabil^3

. gen sabilqu = sabil^4
. reg lwage sabil sabilsq sabilcu

      Source |       SS       df       MS              Number of obs =    1230
-------------+------------------------------           F(  3,  1226) =   64.37
       Model |  58.9534459     3  19.6511486           Prob > F      =  0.0000
    Residual |  374.265816  1226  .305273912           R-squared     =  0.1361
-------------+------------------------------           Adj R-squared =  0.1340
       Total |  433.219262  1229  .352497365           Root MSE      =  .55252

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sabil |   .1889656   .0266104     7.10   0.000     .1367586    .2411726
     sabilsq |  -.0051833   .0190512    -0.27   0.786    -.0425598    .0321933
     sabilcu |   .0085218   .0096005     0.89   0.375    -.0103134     .027357
       _cons |   2.425275   .0213327   113.69   0.000     2.383422    2.467127
------------------------------------------------------------------------------

. test sabilsq sabilcu

 ( 1)  sabilsq = 0
 ( 2)  sabilcu = 0

       F(  2,  1226) =    1.39
            Prob > F =    0.2490
```

```
. reg lwage sabil

      Source |       SS       df       MS              Number of obs =    1230
-------------+------------------------------           F(  1,  1228) =  190.21
       Model |  58.103667     1   58.103667           Prob > F      =  0.0000
    Residual | 375.115595  1228  .305468726           R-squared     =  0.1341
-------------+------------------------------           Adj R-squared =  0.1334
       Total | 433.219262  1229  .352497365           Root MSE      =  .55269

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sabil |   .2189854    .015878    13.79   0.000     .1878343    .2501364
       _cons |   2.414146   .0157591   153.19   0.000     2.383228    2.445063
------------------------------------------------------------------------------

. reg lwage sabil sabilsq, robust

Linear regression                                      Number of obs =    1230
                                                       F(  2,  1227) =   89.72
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1355
                                                       Root MSE      =  .55247

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sabil |   .2061115   .0183971    11.20   0.000     .1700183    .2422047
     sabilsq |  -.0178432   .0127278    -1.40   0.161    -.0428138    .0071274
       _cons |   2.431703    .019943   121.93   0.000     2.392576    2.470829
------------------------------------------------------------------------------
```

3. Semiparametric Methods

- Some estimators that do not involve estimating an infinite dimensional object (along with a finite-dimensional parameter) have been dubbed “semiparametric.” Powell’s (1984) censored LAD estimator is an example.

- Recall the setup:

$$Med(w_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

but the data are, say, top coded. So we observe

$$y_i = \min(y_i, c)$$

- Now

$$Med(y_i|\mathbf{x}_i) = \min(\mathbf{x}_i\boldsymbol{\beta}, c)$$

and so LAD can be applied to this median function.

- Notice that the only parameter to estimate is the $K \times 1$ vector $\boldsymbol{\beta}$. This was labeled "semiparametric" because its main competitors are to specify a full distribution, $D(w_i|\mathbf{x}_i)$ – usually normal – and then apply MLE.

- Powell’s approach might be thought of as “clever” parametric estimation: he found an estimating equation that can be used in a standard procedure, LAD. (However, the nonsmoothness in $\min(\mathbf{x}_i\boldsymbol{\beta}, c)$ and the LAD function make the asymptotics nonstandard.) Remember that Powell’s approach does not generally identify $E(w_i|\mathbf{x}_i)$.
- Certain approaches to estimating coefficients in corner-solution panel data models – in particular, Honoré (1992, *Econometrica*) are similar: they are based on clever objective functions or moment conditions that depend on the finite-dimensional parameter of interest.

Partial Linear Model

- Consider a different kind of semiparametric problem, often known as a **partial linear model** (**PLM**):

$$E(y|\mathbf{x},\mathbf{z}) = \mathbf{x}\boldsymbol{\beta} + g(\mathbf{z})$$

where $\mathbf{x}$ is $1 \times K$ and $\mathbf{z}$ is $1 \times M$. Here, $g(\bullet)$ is an unkown function, and $\mathbf{x}$ does not include a constant.

- Is $\boldsymbol{\beta}$ of interest, or $g : \mathbb{R}^M \to \mathbb{R}$? They both might be.
- This setup excludes the possibility of interactions between $\mathbf{x}$ and $\mathbf{z}$. (The vector $\mathbf{x}$ may include nonlinear functions of variables not in $\mathbf{z}$.)

- What if **z** is discrete taking on a finite number of values, say $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_R\}$. Then, with enough data, the problem is easy: just saturate the model with dummy variables for the different outcomes on **z** – probably an intercept and R – 1 dummies.
- Robinson (1988, *Econometrica*) showed how to estimate $\boldsymbol{\beta}$ assuming **z** is continuous. Usually, **z** is actually a scalar, but the theory is not much easier.
- Identification is the most interesting aspect. Write

$$\begin{aligned} y &= \mathbf{x}\boldsymbol{\beta} + g(\mathbf{z}) + u \\ E(u|\mathbf{x},\mathbf{z}) &= 0 \end{aligned}$$

- Of course, $E(u|\mathbf{z}) = 0$, and so

$$E(y|\mathbf{z}) = E(\mathbf{x}|\mathbf{z})\boldsymbol{\beta} + g(\mathbf{z}).$$

Subtract this from $y = \mathbf{x}\boldsymbol{\beta} + g(\mathbf{z}) + u$ to get

$$y - E(y|\mathbf{z}) = [\mathbf{x} - E(\mathbf{x}|\mathbf{z})]\boldsymbol{\beta} + u$$

- This is the population version of the well-known “partialling out” result from linear regression. Here, we partial out the general mean functions, $E(y|\mathbf{z})$ and $E(\mathbf{x}|\mathbf{z})$.

- The technical issue is now estimating $E(y|\mathbf{z})$ and $E(x_j|\mathbf{z})$, $j = 1,\ldots,K$. Robinson and others consider kernel estimation. Attractive especially when we have just one variable in $\mathbf{z}$. Could use series estimation, too.
- Given estimates, define nonparametric residuals,

$$\begin{aligned} \ddot{y}_i &= y_i - \hat{E}(y_i|\mathbf{z}_i) \\ \ddot{\mathbf{x}}_i &= \mathbf{x}_i - \hat{E}(\mathbf{x}_i|\mathbf{z}_i) \end{aligned}$$

- Then $\hat{\boldsymbol{\beta}}$ is just the OLS estimator (without an intercept) of $\ddot{y}_i$ on $\ddot{\mathbf{x}}_i$, $i = 1,\ldots,N$:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \ddot{\mathbf{x}}_i'\ddot{\mathbf{x}}_i\right)^{-1}\left(\sum_{i=1}^{N} \ddot{\mathbf{x}}_i'\ddot{y}_i\right)$$

- Robinson shows $\hat{\boldsymbol{\beta}}$ is consistent and $\sqrt{N}$-asymptotically normal. As in the case where $g(\cdot)$ is parametric, the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is as if $\ddot{y}_i$ and $\ddot{\mathbf{x}}_i$ are obtained using the (unknown) $E(y_i|\mathbf{z}_i)$ and $E(\mathbf{x}_i|\mathbf{z}_i)$. In other words, we can just act as if our random sample is $\{(\ddot{\mathbf{x}}_i, \ddot{y}_i) : i = 1, \ldots, N\}$ and then use usual OLS inference (probably robust to heteroskedasticity).
- If we use the same series estimation for all conditional means – for example, polynomials of degree P, collected in $\mathbf{r}(\mathbf{z})$, which includes an intercept – then we just do the usual inference in the regression

$$y_i \text{ on } \mathbf{x}_i,\ \mathbf{r}(\mathbf{z}_i),\ i = 1, \ldots, N.$$

- In other words, the naive approach of just approximating $g(\cdot)$ by some flexible functional form, and then using the usual inference, leads to the right place.
- Define

$$v_i = y_i - \mathbf{x}_i\boldsymbol{\beta}.$$

Then

$$E(v_i|\mathbf{z}_i) = g(\mathbf{z}_i),$$

and so we can apply nonparametric regression. But we replace v_i with $\hat{v}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$. See, for example, Li and Racine (2007, *Nonparametric Econometrics: Theory and Practice*).

- In some cases, $\mathbf{z}_i$ actually depends on parameters that have to be estimated. A good application of the PLM approach is to sample selection corrections without full distributional assumptions. Recall we start with

$$\begin{aligned} y_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \\ y_2 &= 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0] \end{aligned}$$

- Maintain independence of (u_1, v_2) and $\mathbf{x}$, so that $E(u_1|v_2,\mathbf{x}) = E(u_1|v_2) \equiv g_1(v_2)$.

- Then

$$E(y_1|\mathbf{x}, v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + g_1(v_2)$$

and

$$\begin{aligned} E[g_1(v_2)|\mathbf{x}, y_2 = 1] &= E[g_1(v_2)|\mathbf{x}, v_2 > -\mathbf{x}\boldsymbol{\delta}_2] \\ &= \int_{-\mathbf{x}\boldsymbol{\delta}_2}^{\mathbb{R}} g_1(\mathfrak{v}_2)f_2(\mathfrak{v}_2)d\mathfrak{v}_2 \equiv h_1(\mathbf{x}\boldsymbol{\delta}_2) \end{aligned}$$

where we use the fact that the density of v_2 given **x** does not depend on **x** (and this density is $f_2(\bullet)$).

- So, we have shown that

$$E(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + h_1(\mathbf{x}\boldsymbol{\delta}_2)$$

where $h_1(\bullet)$ is unknown. Now, we can apply the PLM methods on the selected sample.

- Assume, for the moment, that we can estimate $\boldsymbol{\delta}_2$. Then the semiparametric problem arises because we do not want to assume $h_1(\bullet)$ is known (as we would under normality of v_2, and then $h_1(\bullet)$ would be the proportional to the IMR).

- Notice that our goal here is to estimate $\boldsymbol{\beta}_1$, a population parameter. We are looking beyond the standard Heckman approach to achieve robustness of estimation. (We can and did discuss this in a parametric setting, for example, assume $E(u_1|v_2)$ is a quadratic. But v_2 was still assumed standard normal.)
- If we have a suitable estimator $\hat{\boldsymbol{\delta}}_2$ of $\boldsymbol{\delta}_2$, we can define $\hat{z}_i \equiv \mathbf{x}_i\hat{\boldsymbol{\delta}}_2$ and then apply PLM methods. Notice how we must have something in $\mathbf{x}$, with a nonzero coefficient in $\boldsymbol{\delta}_2$, that is not in $\mathbf{x}_1$. This is why achieving identification off of the nonlinearity of the IMR is very questionable.

- If we use a P^{th} order polynomial, we would run the regression,

$$y_{i1} \text{ on } \mathbf{x}_{i1}, 1, (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2), (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)^2, \ldots, (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)^P \text{ with } y_{i2} = 1,$$

where $\mathbf{x}_{i1}$ no longer includes an intercept (and we cannot identify the intercept).

- Newey (1988, *Journal of Applied Econometrics*) considers more exotic possibilities. For example, first use a monotonic transformation, such as the logistic, $\Lambda(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$, before including the powers in the regression. (Less sensitive to outliers.)

- How can we estimate $\boldsymbol{\delta}_2$?
- If v_2 is independent of $\mathbf{x}$, we have

$$E(y_2|\mathbf{x}) = P(y_2 = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\delta}_2)$$

where $G(\bullet)$ is an unknown function.

- Ichimura (1993) shows how to estimate the slopes in $\boldsymbol{\delta}_2$ up to scale without assuming a parametric model for $G(\bullet)$. Uses kernel smoothing and a least squares approach.
- Klein and Spady (1993) use a log likelihood instead, but also use local smoothing.

Semiparametric Estimation of Index Models

● Ichimura's approach is actually more general and does not require y to be a binary response. In fact, start with a general index model for $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = g(\mathbf{x}\boldsymbol{\beta}_o)$$

where it is now important to distinguish the true value from a generic value. The function $g(\bullet)$ is unknown but assumed smooth.

- Because $E(y|\mathbf{x})$ depends only on $\mathbf{x}\boldsymbol{\beta}_o$, we know $E(y|\mathbf{x}\boldsymbol{\beta}_o) = E(y|\mathbf{x})$.

Recall a property of the conditional mean: for any (vector) function $\mathbf{q}(\mathbf{x})$ of $\mathbf{x}$,

$$E\{[y - E(y|\mathbf{x})]^2\} \leq E\{[y - E(y|\mathbf{q}(\mathbf{x}))]^2\}.$$

- Just says that one cannot do better for predicting y by throwing away information.

- But for any value $\boldsymbol{\beta}$, $\mathbf{x}\boldsymbol{\beta}$ is a function of $\mathbf{x}$. Further, the index assumption implies $E(y|\mathbf{x}) = E(y|\mathbf{x}\boldsymbol{\beta}_o)$, and so

$$E\{[y - E(y|\mathbf{x}\boldsymbol{\beta}_o)]^2\} \leq E\{[y - E(y|\mathbf{x}\boldsymbol{\beta})]^2\}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^K$.

- Therefore, if we can find $E(y|\mathbf{x}\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$, estimate $\boldsymbol{\beta}_o$ by solving

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} [y_i - E(y_i|\mathbf{x}_i\boldsymbol{\beta})]^2.$$

- Issues: (1) Identification; (2) Estimation.
- For identification, $\mathbf{x}_i$ cannot include an intercept, and a normalization is needed on $\boldsymbol{\beta}$, such as $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. (Note this rules out the possibility $E(y|\mathbf{x}) = E(y)$.)
- Ichimura shows in fact that $\mathbf{x}$ needs at least one *continuous* element, say x_1, with nonzero β_1. Then a different normalization sets $\beta_1 \equiv 1$.
- Ichimura uses kernel estimation for $E(y_i|\mathbf{x}_i\boldsymbol{\beta})$. $\hat{\boldsymbol{\beta}}$ solves a problem such as

$$\min_{\substack{\boldsymbol{\beta} \\ \boldsymbol{\beta}'\boldsymbol{\beta}=1}} \sum_{i=1}^{N}[y_i - \hat{E}(y_i|\mathbf{x}_i\boldsymbol{\beta})]^2.$$

- Showing $\sqrt{N}$-consistency of $\hat{\boldsymbol{\beta}}$, and deriving asymptotic properties of the partial effects, is challenging but has been done.
- Another approach is to use a series-type estimator for $\hat{E}(y_i|\mathbf{x}_i\boldsymbol{\beta})$ followed by the least squares problem.

- Or, just make standard models more flexible in a series framework.

For example, if y is binary, use

$$P(y = 1|\mathbf{x}) \approx \Phi[\mathbf{x}\boldsymbol{\beta} + \eta_2(\mathbf{x}\boldsymbol{\beta})^2 \ldots + \eta_P(\mathbf{x}\boldsymbol{\beta})^P]$$

where **x** contains a constant in this formulation. Then estimate $\eta_2, \ldots, \eta_P$ along with $\boldsymbol{\beta}$.

- Ideally, one can study the large-sample properties of partial effects (in addition to those of $\hat{\boldsymbol{\beta}}$) as P increases with N, but this is a hard problem.
- Recall that it is easy to test that the η_h are all zero using the score approach.

- Klein and Spady exploit the binary nature of y_i, and solve

$$\max_{\substack{\boldsymbol{\beta} \\ \boldsymbol{\beta}'\boldsymbol{\beta}=1}} \sum_{i=1}^{N} \{(1-y_i)\log[1-\hat{E}(y_i|\mathbf{x}_i\boldsymbol{\beta})] \ + y_i \log[\hat{E}(y_i|\mathbf{x}_i\boldsymbol{\beta})]\}.$$

where $\hat{E}(y_i|\mathbf{x}_i\boldsymbol{\beta})$ is again obtained using kernel estimation.

- More efficient than using sum of squared residuals. Consistency follows by the Kullback-Leibler inequality because $D(y|\mathbf{x}) = D(y|\mathbf{x}\boldsymbol{\beta}_o)$.

- Need to worry about "trimming" in the kernel estimation (density estimator in the denominator).

- Can extend the KS idea to quasi-log likelihoods.
- Do not fall into the "inconsistent parameter estimates" trap. It is common to hear of the perils of using an LPM, or probit, or logit, for estimating $\boldsymbol{\beta}$ in

$$P(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})$$

when $G(\bullet)$ is unknown. Whether LPM, probit or logit consistently estimate $\boldsymbol{\beta}$ in a general specification is largely irrelevant. They can do a good job of estimating the *ratios* of the coefficients (which is all we can compare in terms of parameters across index specifications, anyway).

- More importantly, how well are the partial effects estimated using simpler models?

- If we think, say, probit is not flexible enough, we can move beyond it (logit, too) by using a series approach as mentioned earlier. Or, we can use other extensions, such as the "heteroskedastic probit" model,

$$P(y = 1|\mathbf{x}) = \Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$$

where **x** includes an intercept but $\mathbf{x}_1$ does not.

- The heteroskedastic probit model is is not even nested in the index model,

$$P(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta}).$$

- Manski (1975) proposed a different approach for binary response.

Write the index model as

$$y = 1[\mathbf{x}\boldsymbol{\beta} + e > 0]$$

under the restriction

$$Med(e|\mathbf{x}) = 0$$

- The indicator function is monotonic (but not strictly!), and so

$$Med(y|\mathbf{x}) = 1[\mathbf{x}\boldsymbol{\beta} > 0]$$

- Manski’s maximum score estimator is a LAD estimator:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - 1[\mathbf{x}_i\boldsymbol{\beta} > 0]|$$

where some normalization is needed on **β** because multiplying $\mathbf{x}_i\boldsymbol{\beta}$ by a positive constant does not change the truth of the event in brackets.

- Like other semiparametric methods, **β** does not include an intercept. And one element of **x** is needed to be continuous (and often its coefficient is set to unity).

- This problem is so “unsmooth” that $\hat{\boldsymbol{\beta}}$ is only consistent at the rate $N^{1/3}$, that is,

$$N^{1/3}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1),$$

and its limiting distribution is nonnormal.

- Horowitz (1992, *Econometrica*) showed how to smooth the indicator function to get asymptotic normality and faster convergence rate.
- By construction, Manski’s approach does not allow estimation of average partial effects. The average structural function is

$$ASF(\mathbf{x}) = E_e\{1[\mathbf{x}\boldsymbol{\beta} + e > 0]\} \quad = 1 - F_e(-\mathbf{x}\boldsymbol{\beta})$$

- If e has a symmetric distribution with continuously differentiable $F_e(\cdot)$, then

$$ASF(\mathbf{x}) = F_e(\mathbf{x}\boldsymbol{\beta})$$

and the APEs can be gotten by differentiating or differencing. For continuous x_j,

$$APE_j(\mathbf{x}) = f_e(\mathbf{x}\boldsymbol{\beta})\beta_j$$

where $f_e(\cdot) > 0$ is the density of e.

- So, being able to estimate the β_j up to the same scale factor allows us to identify the ratios of APEs for continuous variables.

- A puzzling situation. Suppose y given $\mathbf{x}$ follows a heteroskedastic probit model:

$$\begin{aligned} y &= 1[\mathbf{x}\boldsymbol{\beta} + e > 0] \\ e|\mathbf{x} &\sim Normal(0, \exp(2\mathbf{x}_1\boldsymbol{\delta}_1)) \end{aligned}$$

Then $Med(e|\mathbf{x}) = 0$ and so Maximum Score can be used to estimate the slopes up to scale. With MLE, we can estimate $\boldsymbol{\beta}$ (including an intercept). But, again, these give us the relative effects.

- The partial effects obtained from differentiating

$$P(y|\mathbf{x}) = \Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$$

need not even be the same sign as the APEs.

- If we knew the distribution of $\mathbf{x}_1$ we can, in principle, obtain F_e from $e|\mathbf{x} \sim Normal(0, \exp(2\mathbf{x}_1\boldsymbol{\delta}_1))$ by integrating out $\mathbf{x}_1$. F_e will not be normal. Can consistently estimate F_e by averaging out $\mathbf{x}_{i1}$ in the sample. In fact, its density is consistently estimated by

$$\hat{f}_e(e) = N^{-1}\sum_{i=1}^{N}(2\pi)^{-1/2}\exp(-\mathbf{x}_{i1}\hat{\boldsymbol{\gamma}})\exp[-\exp(-2\mathbf{x}_{i1}\hat{\boldsymbol{\gamma}})e^2/2].$$

- What should we report as the partial effects? $\partial P(y|\mathbf{x})/\partial x_j$ or $f_e(\mathbf{x}\boldsymbol{\beta})\beta_j$? The APE approach argues for the latter, consistently estimated from

$$ASF(\mathbf{x}) = N^{-1}\sum_{i=1}^{N}\Phi[\exp(-\mathbf{x}_{i1}\hat{\boldsymbol{\delta}})\mathbf{x}\hat{\boldsymbol{\beta}}].$$

- But, if we started with a random slope model such as

$$P(y_i = 1|\mathbf{x}_i, \mathbf{b}_i) = \Phi(\mathbf{x}_i\mathbf{b}_i)$$
$$\mathbf{b}_i|\mathbf{x}_i \sim Normal(\boldsymbol{\beta}, \boldsymbol{\Psi})$$

then

$$P(y_i = 1|\mathbf{x}_i) = \Phi[\mathbf{x}_i\boldsymbol{\beta}/(\mathbf{x}_i'\boldsymbol{\Psi}\mathbf{x}_i)^{1/2}]$$

(A normalization is needed in $\boldsymbol{\Psi}$, $\Psi_{11} = 1$, so that the variance is unity when the other parameters are zero.)

- The ASF in this case is

$$ASF(\mathbf{x}) = \Phi[\mathbf{x}\boldsymbol{\beta}/(\mathbf{x}'\boldsymbol{\Psi}\mathbf{x})^{1/2}];$$

that is, the ASF and the response probability are the same.

- The problem is that we can get the same $P(y_i = 1|\mathbf{x}_i)$ by starting with a different heteroskedastic probit:

$$y_1 = 1[\mathbf{x}_i\boldsymbol{\beta} + e_i > 0]$$
$$e_i|\mathbf{x}_i \sim Normal(0, \mathbf{x}_i'\boldsymbol{\Psi}\mathbf{x}_i)$$

- Yet the APEs in this latter case are obtained from $F_e(\mathbf{x}\boldsymbol{\beta})$, not $\Phi[\mathbf{x}\boldsymbol{\beta}/(\mathbf{x}'\boldsymbol{\Psi}\mathbf{x})^{1/2}]$.
- There is a fundamental lack of identification, and seems to be no resolution when the focus is on APEs.