Lecture 4 –Nonparametric and semiparametric estimation

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

Nonparametric regression

- Suppose we have a sample, $(X_i, Y_i), i = 1, \ldots, n$ where Y_i is scalar and $X_i \in \mathbb{R}^d$.
- Two perspectives on nonparametric regression:
  - either
    - Econometric model: $Y_i = g(X_i) + u_i$ where $E(u_i \mid X_i) = 0$
    - the function g is an unknown parameter that we want to estimate

Nonparametric regression

- Suppose we have a sample, $(X_i, Y_i), i = 1, \dots, n$ where Y_i is scalar and $X_i \in \mathbb{R}^d$.
- Two perspectives on nonparametric regression:
  - either
    - Econometric model: $Y_i = g(X_i) + u_i$ where $E(u_i \mid X_i) = 0$
    - the function g is an unknown parameter that we want to estimate
  - **or**
    - we want to predict Y_i

Nonparametric regression

- Suppose we have a sample, $(X_i, Y_i), i = 1, \ldots, n$ where Y_i is scalar and $X_i \in \mathbb{R}^d$.
- Two perspectives on nonparametric regression:
  - either
    - Econometric model: $Y_i = g(X_i) + u_i$ where $E(u_i \mid X_i) = 0$
    - the function g is an unknown parameter that we want to estimate
  - **or**
    - we want to predict Y_i
    - the MSE-minimizing predictor of Y_i is the CEF, $E(Y_i \mid X_i)$
    - estimate the CEF to predict Y_i

Two simple versions

- A fully saturated regression model.
  - This is not feasible if some values in the support of X_i have only one observation.
  - More generally, this leads to predicted values with a high variance.

Two simple versions

- Two simple estimators:
    - A “histogram” estimator: $\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(|X_i - x| \leq h)}{\sum_{i=1}^{n} \mathbf{1}(|X_i - x| \leq h)}$

Two simple versions

- Two simple estimators:
  - A “histogram” estimator: $\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(|X_i - x| \leq h)}{\sum_{i=1}^{n} \mathbf{1}(|X_i - x| \leq h)}$
  - The k-nearest neighbor estimator: $\hat{g}(x)$ equals the simple average of the k observations with smallest values of $|X_i - x|$.

Two simple versions

- Two simple estimators:
  - A “histogram” estimator: $\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(|X_i - x| \leq h)}{\sum_{i=1}^{n} \mathbf{1}(|X_i - x| \leq h)}$
  - The k-nearest neighbor estimator: $\hat{g}(x)$ equals the simple average of the k observations with smallest values of $|X_i - x|$.
- These are in a sense equivalent.
- But not as you vary x.
- And choice of h, k differ.

Kernel regression

- Suppose $Y_i = g(X_i) + u_i$ and $E(u_i \mid X_i) = 0$
  - The kernel regression estimator (Nadaraya-Watson) is

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right) Y_i}{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right)}$$

where $K(u)$ is a *kernel* function and h is a *bandwidth* parameter.

Kernel regression

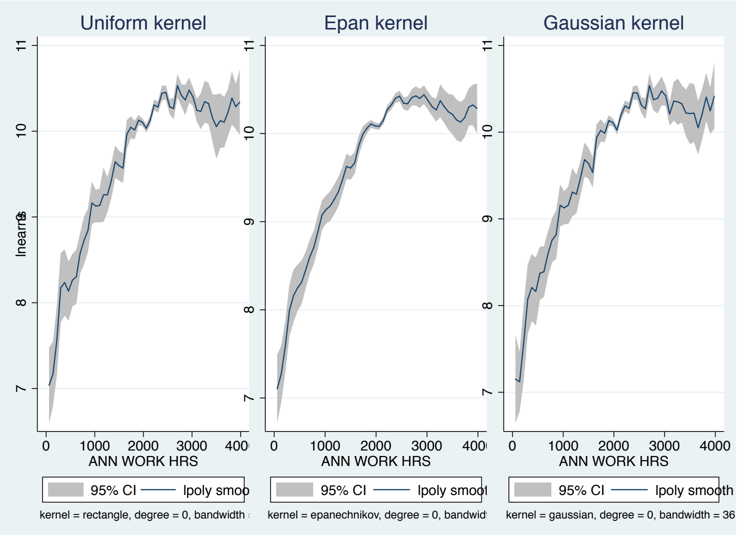
- Suppose $Y_i = g(X_i) + u_i$ and $E(u_i \mid X_i) = 0$
  - The kernel regression estimator (Nadaraya-Watson) is

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right) Y_i}{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right)}$$

    where $K(u)$ is a *kernel* function and h is a *bandwidth* parameter.
  - Example kernel functions:
    - uniform: $K(u) = \mathbf{1}(|u| \leq 1)$
    - triangular: $K(u) = (1 - |u|)\mathbf{1}(|u| \leq 1)$
    - gaussian: $K(u) = \phi(u)$

Different kernel functions

- The choice of kernel function is often not that important.

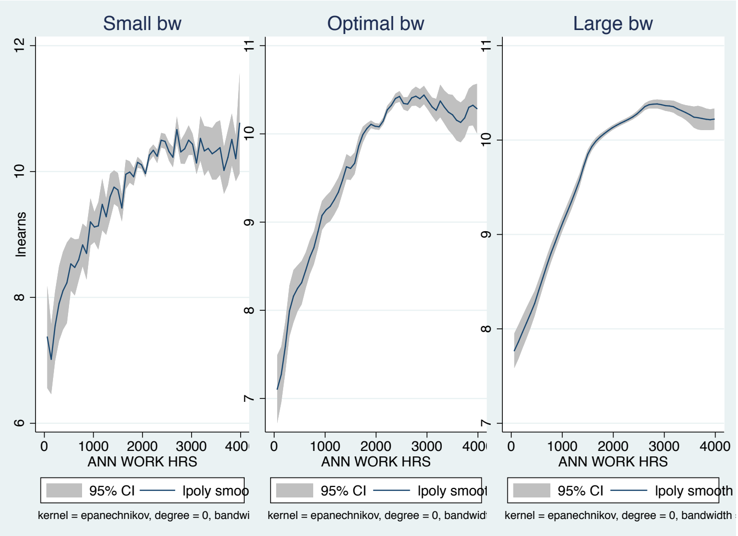


Kernel regression

- The choice of bandwidth is important.
    - too small → high variance
    - too large → high bias

Different bandwidths

- Comparison of optimal bandwidth h^* with $h^*/4$ and $4h^*$



Kernel regression

- Let f denote the density of X_i, which we assume for now is scalar.
- Under some regularity conditions:

$$Bias(\hat{g}(x)) = h^2\left(g'(x)\frac{f'(x)}{f(x)} + g''(x)\right)\int u^2 K(u)du \\ + O(n^{-1}h^{-1}) + o(h^2)$$

Kernel regression

- Let f denote the density of X_i, which we assume for now is scalar.
- Under some regularity conditions:

$$Bias(\hat{g}(x)) = h^2 \left(g'(x) \frac{f'(x)}{f(x)} + g''(x) \right) \int u^2 K(u) du + O(n^{-1} h^{-1}) + o(h^2)$$

and

$$Var(\hat{g}(x)) = \frac{Var(u_i \mid X_i = x)}{nhf(x)} \int (K(u))^2 \, du + o(n^{-1} h^{-1})$$

Kernel regression

- Let f denote the density of X_i, which we assume for now is scalar.
- Under some regularity conditions:

$$Bias(\hat{g}(x)) = h^2 \left(g'(x) \frac{f'(x)}{f(x)} + g''(x) \right) \int u^2 K(u) du + O(n^{-1} h^{-1}) + o(h^2)$$

  and

$$Var(\hat{g}(x)) = \frac{Var(u_i \mid X_i = x)}{nhf(x)} \int (K(u))^2 \, du + o(n^{-1} h^{-1})$$

- The h that balances bias and variance is $O(n^{-1/5})$.

Kernel regression

- Asymptotic normality:

$$\sqrt{nh}(\hat{g}(x) - g(x) - Bias(\hat{g}(x))) \\ \rightarrow_d N\left(0, f(x)^{-1} Var(u_i \mid X_i = x) \int K(u)^2 du\right)$$

Kernel regression

- Asymptotic normality:

$$\sqrt{nh}(\hat{g}(x) - g(x) - Bias(\hat{g}(x))) \\ \to_d N\left(0, f(x)^{-1} Var(u_i \mid X_i = x) \int K(u)^2 du\right)$$

- if $\sqrt{nh}h^2 \to 0$ then $\sqrt{nh}(\hat{g}(x) - g(x))$ has the same asy. dist.
  - some *undersmoothing* is necessary

Kernel regression

- Some important assumptions (see, e.g., Pagan and Ullah):
    - g and f are twice cts diff'ble near x
    - $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2 K(u)du < \infty$
    - $h \to 0$ and $nh \to \infty$
    - x is an interior point, f'' is cts. and bounded near x

Kernel regression

- Bandwidth choice:
  - Optimal bandwidth minimizes the MISE: $\int E((\hat{g}(x) - g(x))^2)f(x)dx$
    - plug-in
    - cross-validation
- Trimming

Kernel regression

- (Pointwise) Confidence intervals:
  - Estimate asy. variance.
  - Estimate, undersmooth, or ignore the bias.
  - Then the construction is typical.
- Uniform confidence intervals (Hardle, 1990)

Kernel regression

- In Stata: `lpoly yvar xvar, degree(0)`
  - what this name/syntax means will be apparent shortly...
  - default kernel is Epanechnikov; specify other with `kernel()`
  - default bandwidth is the plug-in estimator; specify other with `bw)`
  - use `at()` to specify support points
  - `ci` provides confidence intervals; pointwise, ignores the bias
  - trimming can be done manually...

Multivariate kernel regression

- Suppose that $X_i \in \mathbb{R}^d$.
- Under sufficient regularity conditions,
    - $Bias(\hat{g}(x)) \approx h^2 \sum_{k=1}^{d} \left(g_k(x) \frac{f_k(x)}{f(x)} + g_{kk}(x) \right) \int u^2 K(u) du$
    - $Var(\hat{g}(x)) \approx \frac{Var(u_i|X_i=x)}{nh^d f(x)} \left(\int (K(u))^2 du \right)^d$
    - optimal convergence rate depends on d – curse of dimensionality.

Kernel regression

- Recommendations:
  - Standardize regressors first.
  - Take care with observations in the tails, or on a boundary.
  - If inference is important in your application
    - Do not ignore the bias.
    - Choose the bandwidth carefully (try various different methods).
    - Remember the difference between pointwise and uniform CIs.
  - If $d > 2$, try to think of plausible economic restrictions to impose.

Local polynomial regression

- The NW kernel regression estimator can be viewed as a “locally constant” regression.
- A local linear regression minimizes a weighted sum of squares

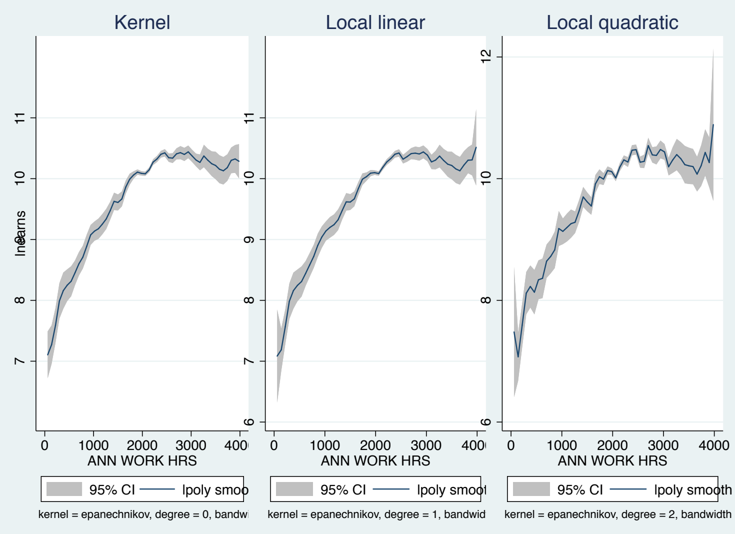
$$\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right) \left(Y_i - a_0 - a_1(X_i - x_0)\right)^2$$

.

- This can be extended by replacing summand with $(Y_i - a_0 - a_1(X_i - x_0) - \ldots - a_k(X_i - x_0)^k)^2$.

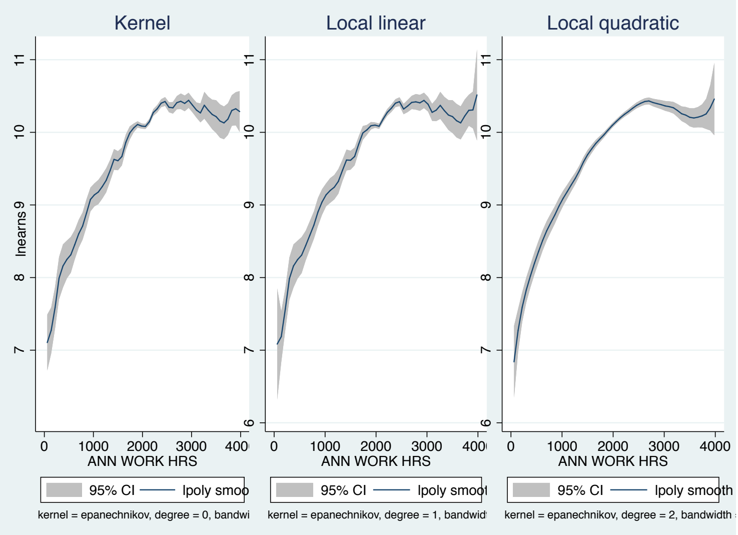
Comparison

- Comparison of different order local polynomials at fixed bandwidth:



Comparison

- Letting Stata choose the optimal bandwidth:



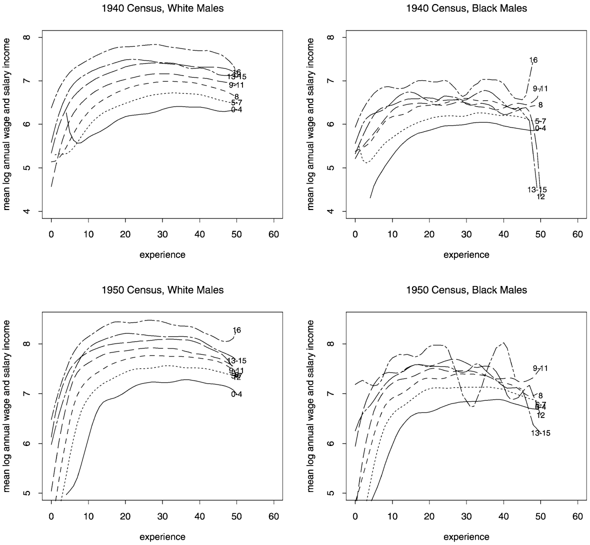
Local polynomial regression

- The asymptotic variance of the local linear regression (LLR) estimator is the same as for the NW estimator.
- The bias is $\frac{1}{2}h^2 g''(x) \int u^2 K(u) du$.

Other nonparametric regression estimators

- A series estimator (or other sieve estimators) approximates $g(x_0)$ globally using polynomials (or other basis functions).

Section 3 of Heckman, Lochner, Todd (2006)

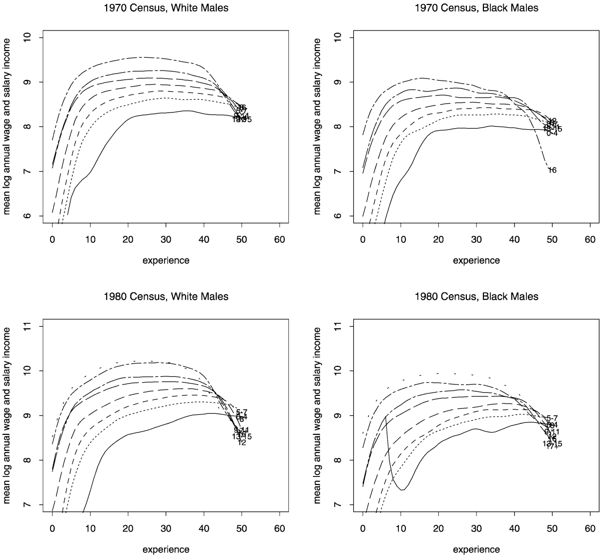

1940 Census, White Males

1940 Census, Black Males

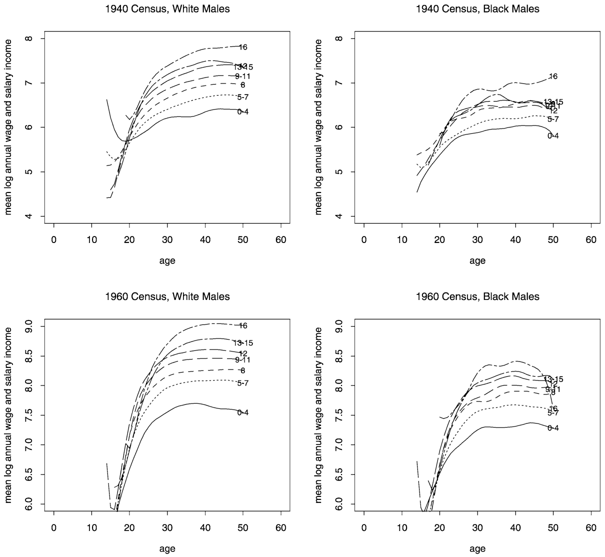
1950 Census, White Males

1950 Census, Black Males

Section 3 of Heckman, Lochner, Todd (2006)



Section 3 of Heckman, Lochner, Todd (2006)



1940 Census, White Males

1940 Census, Black Males

1960 Census, White Males

1960 Census, Black Males

Section 3 of Heckman, Lochner, Todd (2006)

- An implication of the economic model is that $E(y_i \mid x_i, s_i = s_1) - E(y_i \mid x_i, s_i = s_2)$ does not vary with x_i.
- To test this:
  - estimate $\hat{g}(x, s)$ for various values of x, s
  - form a test statistic and use the asymptotic variance derived previously

Section 3 of Heckman, Lochner, Todd (2006)

- An implication of the economic model is that $E(y_i \mid x_i, s_i = s_1) - E(y_i \mid x_i, s_i = s_2)$ does not vary with x_i.
- Two issues that they address:
  - $Cov\left(\hat{g}(x_1, s_1), \hat{g}(x_2, s_2)\right)$?
  - what formula to use for asymptotic variance?
    - they use a method that mimics the OLS standard error formula, instead of the plug-in formula.
  - Kernel? (quartic kernel)
  - bandwidth choice?

Section 3 of Heckman, Lochner, Todd (2006)

- An implication of the economic model is that $E(y_i \mid x_i, s_i = s_1) - E(y_i \mid x_i, s_i = s_2)$ does not vary with x_i.
- Two issues that they address:
    - $Cov\left(\hat{g}(x_1, s_1), \hat{g}(x_2, s_2)\right)$?
    - what formula to use for asymptotic variance?
        - they use a method that mimics the OLS standard error formula, instead of the plug-in formula.
    - Kernel? (quartic kernel)
    - bandwidth choice? (“The bandwidth parameter is equal to 5 years. Estimates are not very sensitive to changes in the bandwidth parameter in the range of 3-10 years.”)

Section 3 of Heckman, Lochner, Todd (2006)

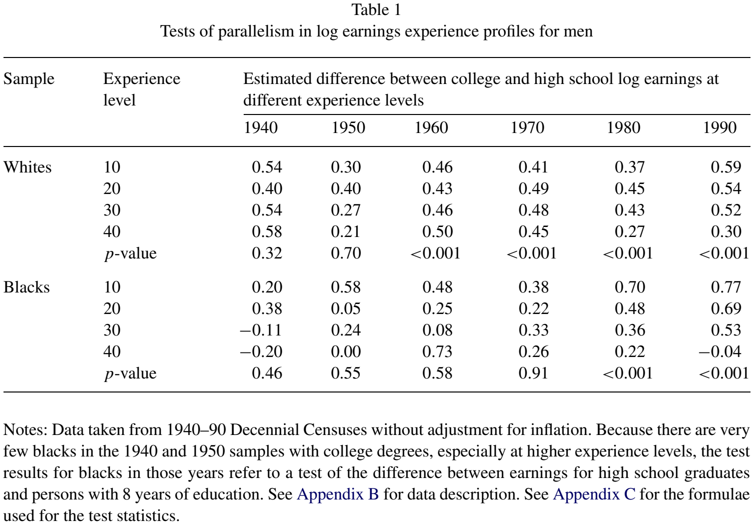


Table 1
Tests of parallelism in log earnings experience profiles for men

Sample	Experience level	Estimated difference between college and high school log earnings at different experience levels					
		1940	1950	1960	1970	1980	1990
Whites	10	0.54	0.30	0.46	0.41	0.37	0.59
	20	0.40	0.40	0.43	0.49	0.45	0.54
	30	0.54	0.27	0.46	0.48	0.43	0.52
	40	0.58	0.21	0.50	0.45	0.27	0.30
	p-value	0.32	0.70	<0.001	<0.001	<0.001	<0.001
Blacks	10	0.20	0.58	0.48	0.38	0.70	0.77
	20	0.38	0.05	0.25	0.22	0.48	0.69
	30	−0.11	0.24	0.08	0.33	0.36	0.53
	40	−0.20	0.00	0.73	0.26	0.22	−0.04
	p-value	0.46	0.55	0.58	0.91	<0.001	<0.001

Notes: Data taken from 1940–90 Decennial Censuses without adjustment for inflation. Because there are very few blacks in the 1940 and 1950 samples with college degrees, especially at higher experience levels, the test results for blacks in those years refer to a test of the difference between earnings for high school graduates and persons with 8 years of education. See Appendix B for data description. See Appendix C for the formulae used for the test statistics.

Semiparametric models

- A semiparametric model has a parametric component (β) and a nonparametric component (g).
- Examples:
  - A partially linear regression model: $Y_i = \beta' X_{1i} + g(X_{2i}) + u_i$.
  - A linear regression model, $Y_i = \beta' X_i + u_i$, with unknown heteroskedasticity, $g(X_i) = \sigma^2(X_i)$.
  - Average derivative: $Y_i = g(X_i) + u_i$, $\beta = E\left(\frac{\partial g(X_i)}{\partial X_i}\right)$
  - Matching estimators
- Newey (1990) is a good source for an introduction to technical conditions in semiparametric estimation.

Semiparametric models

- Advantages of some semiparametric estimators:
  - Often they will converge at rate $\sqrt{n}$ to a normal distribution (just like a parametric estimator).
  - If this happens, it is also often the case that the asymptotic distribution does not depend on what estimator is used for the nonparametric component.
  - In very nice cases, the asymptotic distribution is the same as if g had been known. (*adaptive*)
  - In some cases we can derive a *semiparametric efficiency bound*.

Semiparametric models

- Issues with semiparametric estimators:
  - The asymptotic arguments can be somewhat fragile.
  - Trimming is often necessary.
  - Can still have sensitivity to bandwidth choice, kernel choice, etc. in practice.
  - It is not always clear if the first stage estimator is good enough.