Advanced Econometrics

University of Warsaw

Jerzy Mycielski

Doctoral School of Social Sciences, 2026

Nonparametric regression

- When estimating models using parametric methods, we assume that the functional form of the regression function $E(y|X) = m(X, \theta)$ is known and the estimation concerns the unknown parameter θ.
- Is it possible to estimate a regression function without assuming a specific functional form of the regression function?
- The methods we use in such a case are referred to as non-parametric methods.
- One regressor nonparametric regression model:

$$y = m(x) + e$$

$$\mathbb{E}(e|X) = 0$$

$$\mathbb{E}\left(e^2 \middle| X\right) = \sigma^2(X)$$

- Binned means estimator:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} 1\{|X_i - x| < h\} Y_i}{\sum_{i=1}^{n} 1\{|X_i - x| < h\}}$$

for some h, and bins centered in $\{x_1, x_2, \ldots, x_K\}$

Nadaraya-Watson estimator

- Nadaraya-Watson regression estimator $m(X) = m(x)$

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

  where h is bandwidth.

- Nadaraya-Watson estimator can be derived as the solution of the following minimization problem

$$Y = m(X) + e \simeq m(x) + e$$

$$\underset{m}{argmin} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) (Y_i - m)^2$$

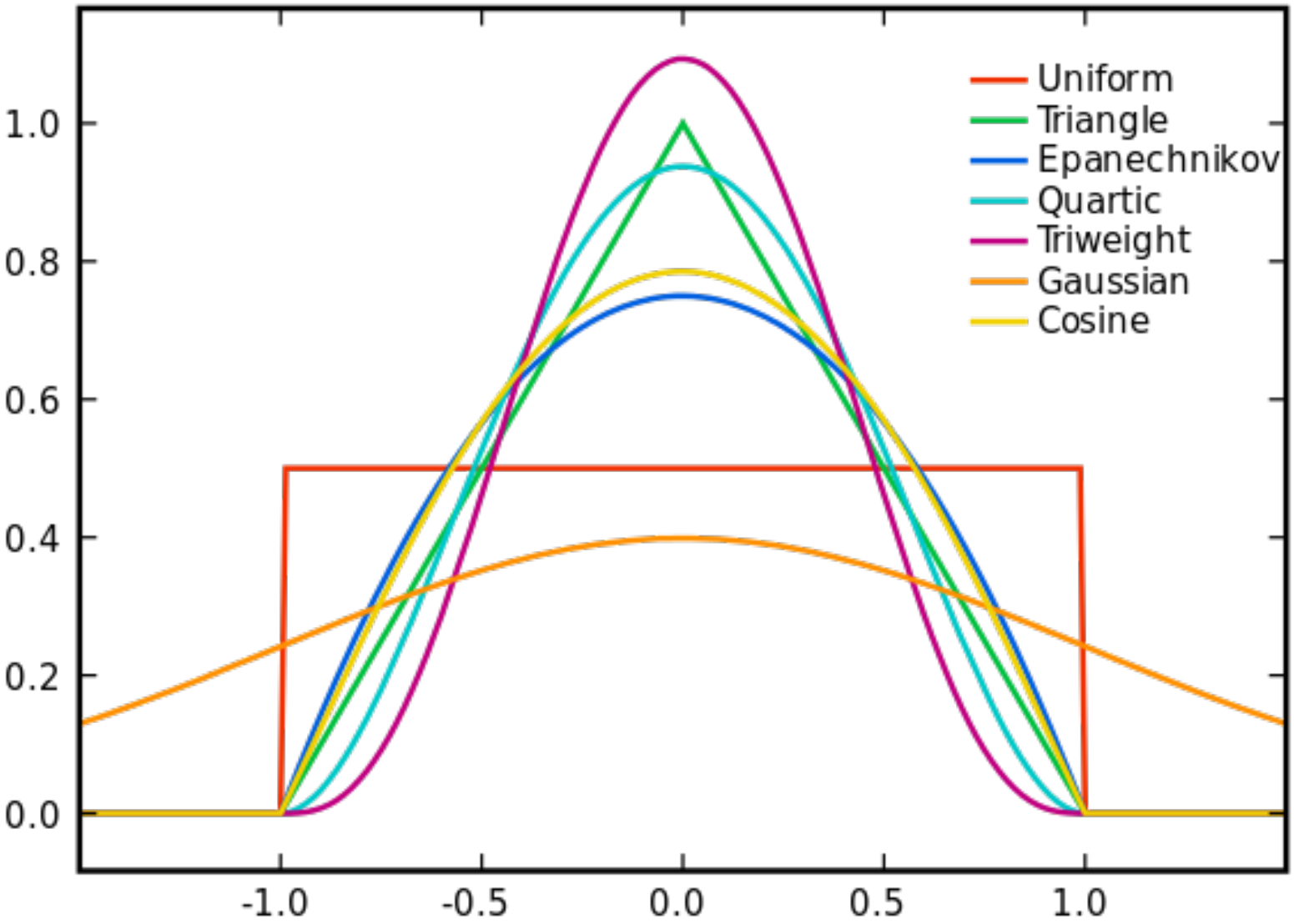
Kernel Function

- Properties of the kernel function $K(u)$:
  - $0 \leq K(u) \leq \overline{K}(u) < \infty$
  - $K(u) = K(-u)$
  - $\int_{-\infty}^{\infty} K(u)\, du = 1$
  - $\int_{-\infty}^{\infty} |u|^r K(u)\, du < \infty$ for positive r
- Sometimes we add normalization:
  - $\int_{-\infty}^{\infty} |u|^2 K(u)\, du = 1$

Examples of normalized kernel functions

Kernel	Formula	R_K
Rectangular	$K(u) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } \lvert u\rvert < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{2\sqrt{3}}$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right)$	$\frac{1}{\sqrt{2\pi}}$
Epanechnikov	$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right) & \text{if } \lvert u\rvert < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$	$\frac{3\sqrt{5}}{25}$
Triangular	$K(u) = \begin{cases} \frac{1}{\sqrt{6}}\left(1 - \frac{\lvert u\rvert}{\sqrt{6}}\right) & \text{if } \lvert u\rvert < \sqrt{6} \\ 0 & \text{otherwise} \end{cases}$	$\frac{\sqrt{6}}{9}$

Graphs of kernel functions



Source: Qin (2026)

Local polynomial estimator

- Local polynomial estimator

$$\begin{aligned} Y &= m(X) + e \\ &\simeq m(x) + m'(x)(X-x) \ldots m^{(p)}(x)\frac{(X-x)^p}{p!} + e \\ &= Z_i(X,x)'\beta(x) + e \end{aligned}$$

$$Z_i(X,x) = \left(1, X_i - x, \ldots, \frac{(X_i-x)^p}{p!}\right),\ \beta(x) = \left(m(x), \ldots, m^{(p)}\right)$$

- Local linear estimator if $p = 1$

$$Y = m(X) + e \simeq m(x) + m'(x)(X-x) + e = Z_i(X,x)'\beta(x) + e$$

Local polynomial estimator

- Local polynomial estimator is found by solving the following minimisation problem:

$$\underset{\beta}{argmin} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Y_i - \beta_0 - \beta_1 (X_i - x) \ldots - \beta_p \frac{(X_i - x)^p}{p!}\right)^2$$

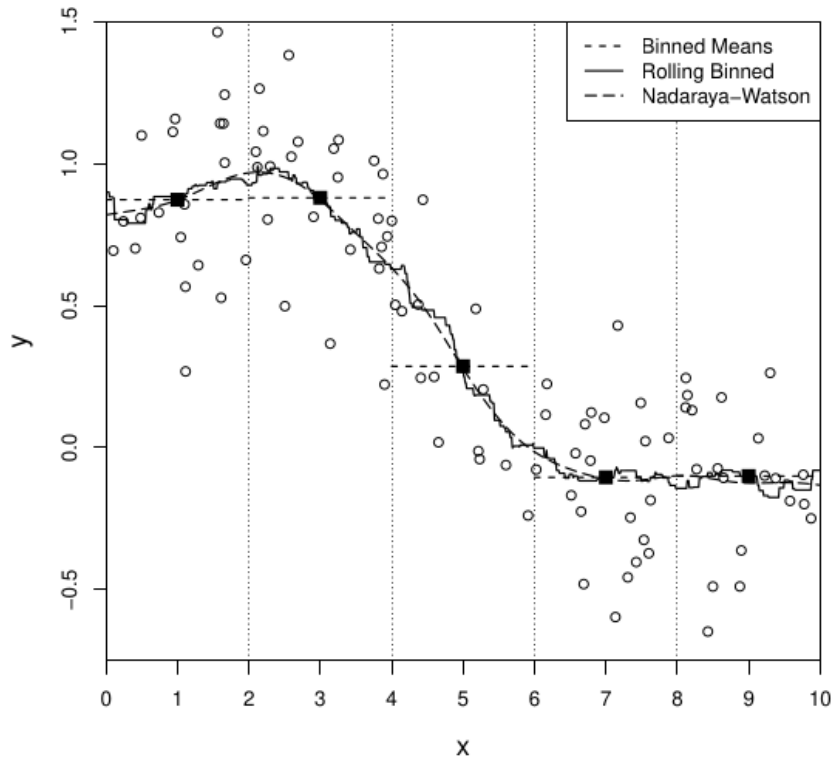
- Solution:

$$\widehat{\beta}_{LP} = \left(\mathbf{Z}'\mathbf{K}\mathbf{Z}\right)^{-1} \mathbf{Z}'\mathbf{K}\mathbf{Y},$$

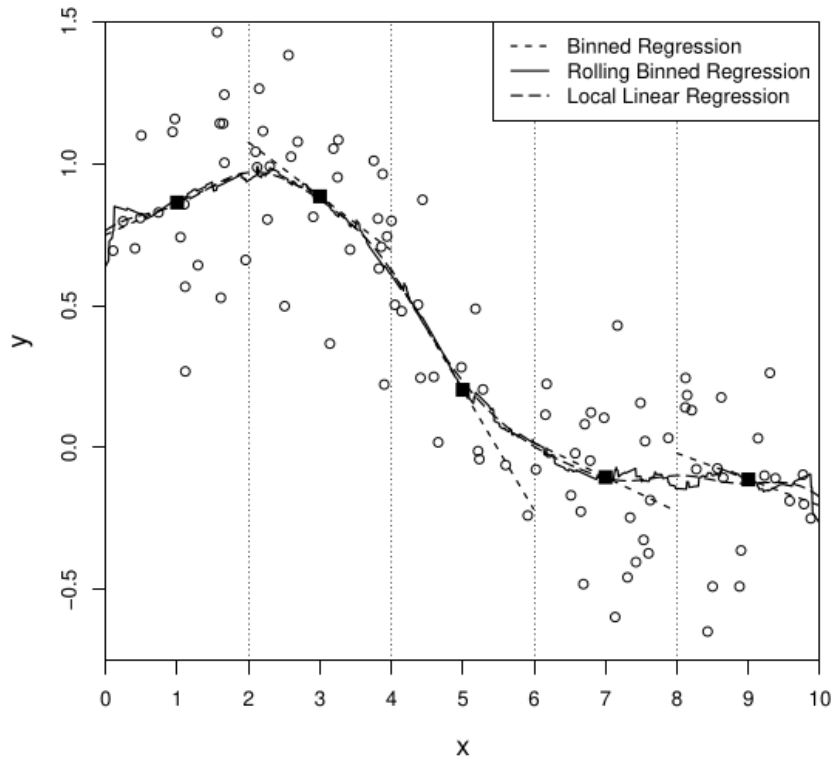
where $\mathbf{K} = diag\left\{K\left(\frac{X_1 - x}{h}\right), ..., K\left(\frac{X_n - x}{h}\right)\right\}$, $\mathbf{Z}$ is the stacked $Z'(X, x)$ and and $\mathbf{Y}$ is the stacked Y_i.

- The most popular special cases:
  - Nadaraya-Watson (NW)
  - Local Linear (LL)

Nadaraya-Watson and Local Linear estimators



(a) Nadaraya-Watson



(b) Local Linear

Figure 19.1: Nadaraya-Watson and Local Linear Regression

Stochastic order symbols, small $o()$

- We will use special notation for factors that converge to zero with some given rate.
- Usually this factors are representing the deterministic or stochastic approximation error.
- Small o notation:

Notation	Meaning
$x_n = o(1)$	$lim_{n\to\infty} x_n = 0$
$x_n = o(a_n)$	$a_n^{-1} x_n = o(1)$
$X_n = o_p(1)$	$plim_{n\to\infty} X_n = 0$
$X_n = o_p(a_n)$	$a_n^{-1} X_n = o_p(1)$

- E.g. for consistent estimator $\hat{\beta} - \beta = o_p(1)$ and $\widehat{\beta} = \beta + o_p(1)$.

Stochastic order symbols, big $O()$

- x_n is bounded uniformly in n if exists $M < \infty$ such that $|x_n| < M$ for all M.
- X_n is bounded in probability if for any $\varepsilon > 0$ there exist constant $M_\varepsilon < \infty$ such that $lim_{n\to\infty} sup \mathbb{P}\left[|X_n| < M_\varepsilon\right] = \varepsilon$.
- If X_n is bounded than intuitively it does not diverge for $n \to \infty$.
- Big O notation

Notation	Meaning
$x_n = O(1)$	x_n is bounded uniformly in n
$x_n = O(a_n)$	$a_n^{-1} X_n = O(1)$
$X_n = O_p(1)$	is bounded in probability
$X_n = O_p(a_n)$	$a_n X_n = O_p(1)$

- E.g. if estimator $\hat{\beta}$ is square root consistent $(\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, \Sigma))$ then $\sqrt{n}\left(\widehat{\beta} - \beta\right) = O_p(1)$, $\widehat{\beta} = O_p\left(n^{-\frac{1}{2}}\right)$.

Conistency of the NW and LL estimators

- Assumptions (1):

1. $h \to 0$,
2. $nh \to \infty$
3. $m(x)$, $f(x)$, and $\sigma^2(x)$ are continuous in neighborhood of x.
4. $f(x) > 0$

Theorem

Suppose assumptions (1) above hold and $m''(x)$ and $f'(x)$ is continuous in neighborhood of x. Then

$$\mathbb{E}\left(\widehat{m}_{NW}(x)|X\right) = m(x) + h^2 B_{NW}(x) + o_p\left(h^2\right) + O_p\left(\sqrt{\frac{h}{n}}\right)$$

$$B_{NW}(x) = \frac{1}{2}m''(x) + f(x)^{-1} f'(x) m'(x)$$

$$\mathbb{E}\left(\widehat{m}_{LL}(x)|X\right) = m(x) + h^2 B_{LL}(x) + o_p\left(h^2\right) + O_p\left(\sqrt{\frac{h}{n}}\right)$$

$$B_{LL}(x) = \frac{1}{2}m''(x)$$

Bandwith, bias and variance

- $B_{NW}(x)$ and $B_{LL}(x)$ are called asymptotic biases. For fixed h they do not vanish even for $n \to \infty$.
- Notice such a bias (smoothing bias) depends on the curvature $m''(x)$ of regression function.

Theorem

For assumptions (1)

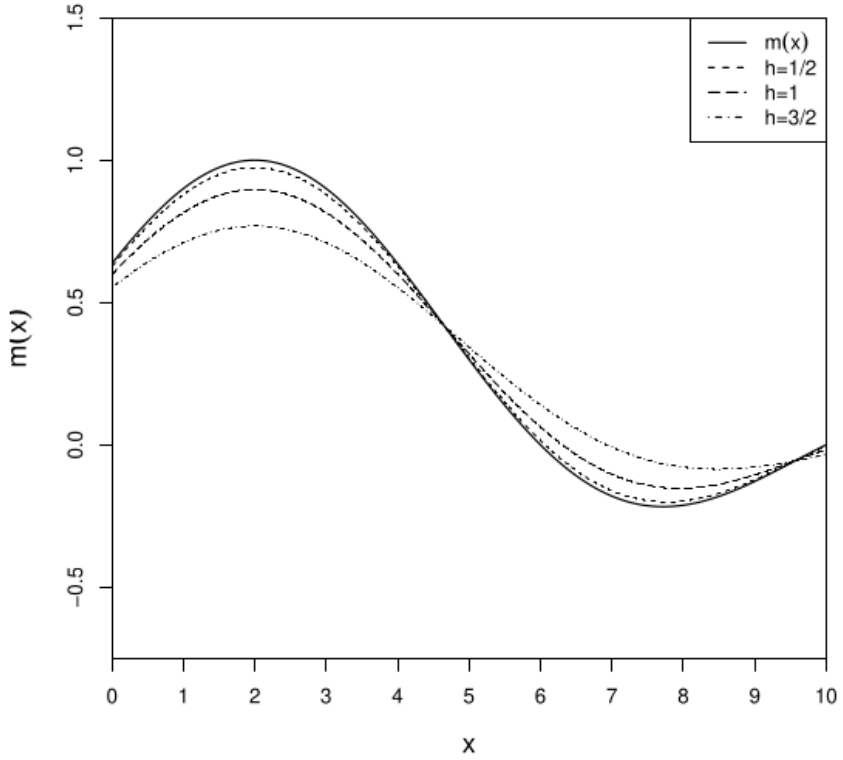
1. $var\left(\widehat{m}_{NW}(x)|\mathbf{X}\right) = \frac{R_K \sigma^2(x)}{f(x)nh} + o_p\left(\frac{1}{nh}\right)$
2. $var\left(\widehat{m}_{LL}(x)|\mathbf{X}\right) = \frac{R_K \sigma^2(x)}{f(x)nh} + o_p\left(\frac{1}{nh}\right)$

where R_K is the kernel roughness

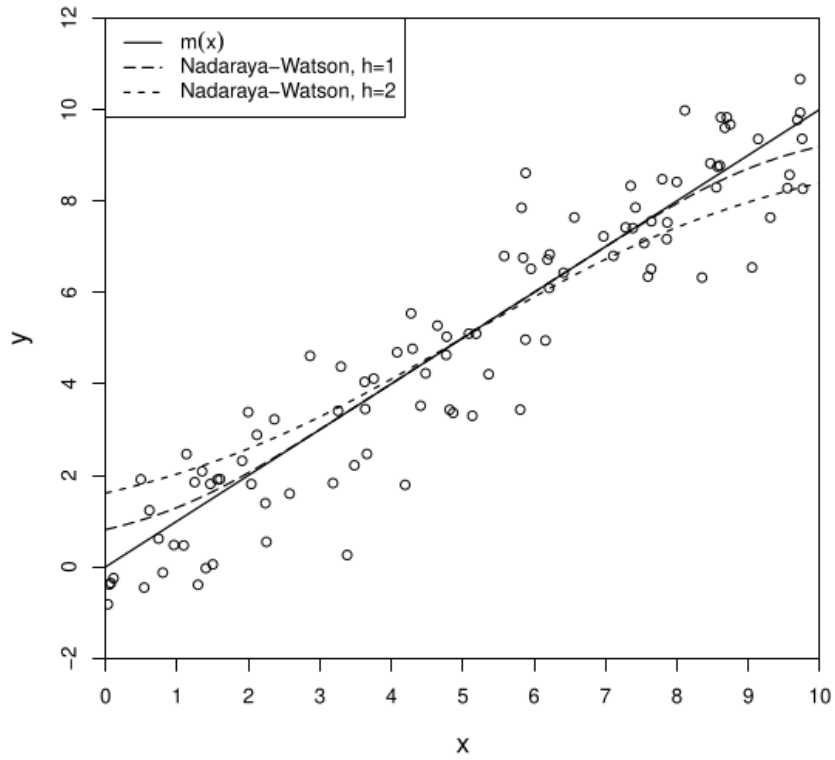
$$R_K = \int_{-\infty}^{\infty} K(u)^2 \, du$$

- For larger h, estimators has larger bias but smaller variance
- What is then the optimal bandwith?

Smoothing bias, boundary bias



(a) Smoothing Bias



(b) Boundary Bias

Source: Hansen (2022)

AMSE and AIMSE

- Asymptotic Mean Squared Error (AMSE) is defined as follows:

$$AMSE\left(x\right) \overset{def}{=} \underbrace{h^4 B\left(x\right)}_{bias^2} + \underbrace{\frac{R_K \sigma^2\left(x\right)}{f\left(x\right) nh}}_{variance}$$

- Asymptotic Integrated Mean Squared Error (AIMSE) is defined as follows:

$$AIMSE \overset{def}{=} \int_S AMSE\left(x\right) f\left(x\right) w\left(x\right) dx = h^4 \overline{B}\left(x\right) + \frac{R_K \overline{\sigma}^2\left(x\right)}{nh}$$

where S is a support of X and

$$\overline{B}\left(x\right) = \int_S B\left(x\right)^2 f\left(x\right) w\left(x\right) dx$$

$$\overline{\sigma}^2 = \int_S \sigma^2\left(x\right) w\left(x\right) dx$$

- An integrable weight function is needed when X has unbounded support to ensure that $\overline{\sigma}^2 < \infty$. Often $w\left(x\right) = 1\left\{\xi_1 \leq x \leq \xi_2\right\}$

Optimal bandwith

- The *AIMSE* measure allows for the trade-off between variance and bias to be taken into account when determining the optimal bandwidth
- The bandwith with minimises *AIMSE* is

$$h_0 = \left(\frac{R_K \overline{\sigma}^2}{4\overline{B}} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

- Inserting h_0 to formula for *AIMSE* we obtain

$$AIMSE_0 = 1.65 \left(R_K^4 \overline{B} \overline{\sigma}^8 \right)^{\frac{1}{5}} n^{-\frac{4}{5}} = O\left(n^{-\frac{4}{5}} \right)$$

- The *AIMSE* of the NW and LL estimators are minimized by the Epanechnikov kernel but the difference with other kernels is minor (1%-3%)
- The formula for $AIMSE_0$ cannot be used directly as $\overline{B}$ and $\overline{\sigma}^8$ are unknown.

Rule of Thumb (ROT), estimation on boundary

- Fan and Gijbels (1996) suggested to preliminary estimate $m(x)$ with OLS using polynomial of order 4

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + e$$

- The estimate of $\widehat{m}''(x) = 2\widehat{\beta}_2 + 6\widehat{\beta}_3 x + 12\widehat{\beta}_4 x^2$
- Moment estimate of $\widehat{B}_{LL} = \mathbb{E}\left[B(x)^2 w(u)\right] = \frac{1}{n}\sum\left(\frac{1}{2}\widehat{m}''(X_i)\right)^2 1\left\{\widehat{\xi}_1 \leq X_i \leq \widehat{\xi}_2\right\}$ for e.g. 5% and 95% quantiles $\widehat{\xi}_1$, $\widehat{\xi}_2$ of the X distribution
- Assume that error term is homoscedastic $\sigma^2(x) = \widehat{\sigma}^2$ so that $\overline{\sigma}^2 = \widehat{\sigma}^2\left(\widehat{\xi}_2 - \widehat{\xi}_1\right)$ and estimate $\widehat{\sigma}^2$ from preliminary estimation
- Plugin into formula for optimal badwith to obtain ROT bandwith
- It can be shown that estimates of m(x) at the boundary of the domain x are characterized by high smoothing bias.
- In such cases, the use of the LL estimator is recommended.

Cross validation bandwidth

- Notice that nonparametric estimators are defined so that for $h \to 0$, $\widehat{m}(X_i) \to Y_i$ and $e_i = Y_i - \widehat{m}(X_i) \to 0$
- This implies that e_i is not a good measure of fit (overfitting)
- $\widetilde{Y}_i = \widehat{m}_{-i}(X_i)$ is the prediction of Y_i based on estimation using all observations except of observation i
- Define leave-one-out residual as $\widetilde{e}_i = Y_i - \widetilde{Y}_i$

$$IMSE_n(h) = \int_S \mathbb{E}\left[\left(\widehat{m}(x,h) - m(x)\right)^2\right] f(x)\, w(x)\, dx$$

- Define

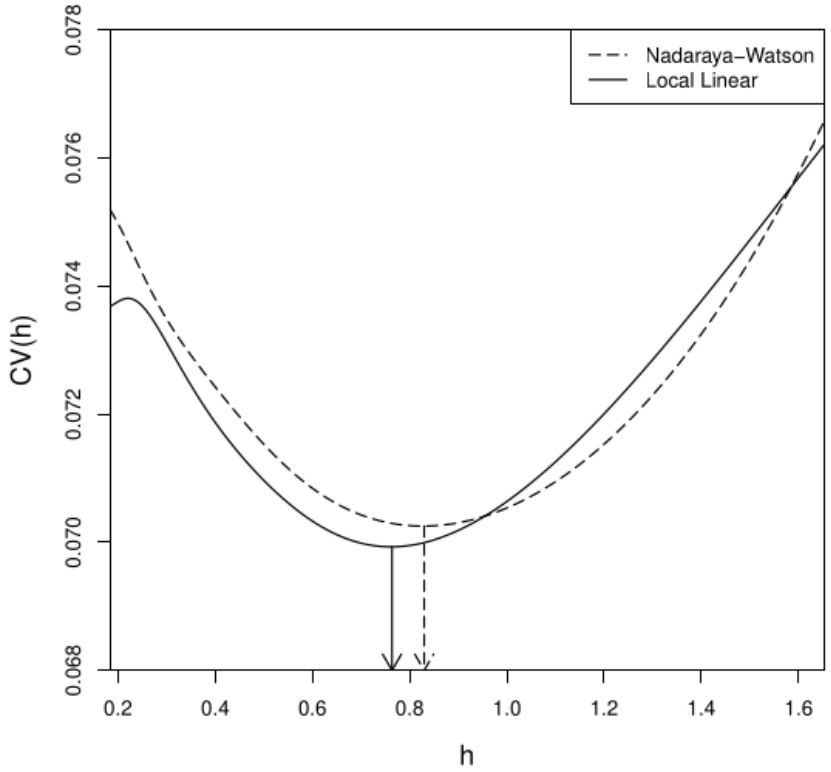
$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{e}_i(h)^2\, w(X_i)$$

- It can be shown that $\mathbb{E}[CV(h)] = \overline{\sigma}^2 + IMSE_{n-1}(h)$
- As $\overline{\sigma}^2$ does not depend on h, minimisation of $E(CV(h))$ is equivalent to minimization of $IMSE_{n-1}(h)$ then

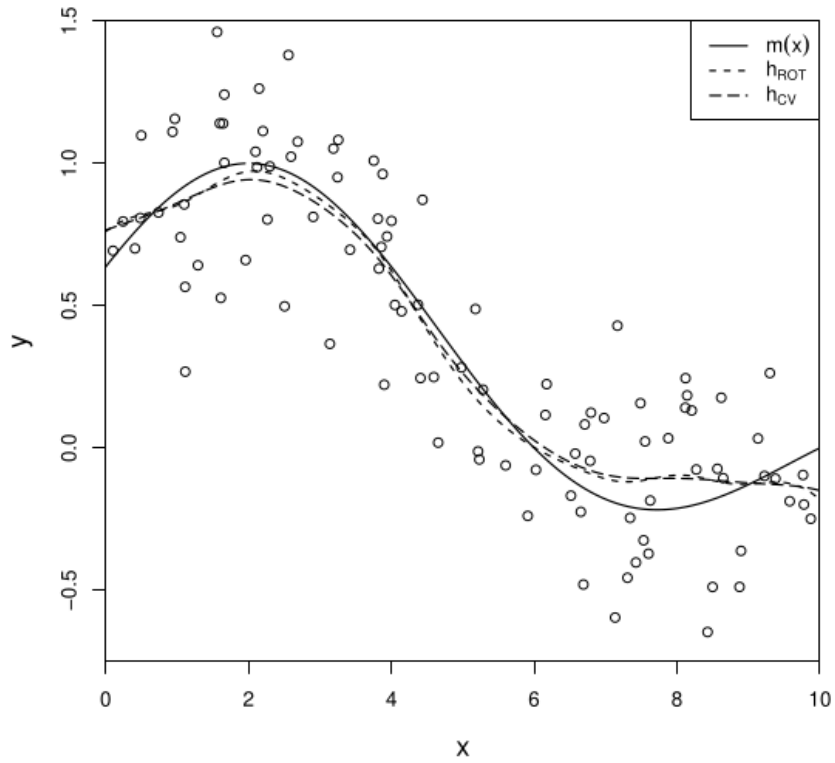
$$h_{CV} = \underset{h \geq h_\ell}{argmin}\ CV(h)$$

- Solution to this minimisation problem is found numerically

ROT and CV bandwith



(a) Cross-Validation Criterion



(b) Nonparametric Estimates

Source: Hansen (2022)

Asymptotic properties of nonparametric estimators

- Under Assumptions (1) $\widehat{m}_{NW}(x) \xrightarrow{p} m(x)$ and $\widehat{m}_{LL}(x) \xrightarrow{p} m(x)$

Theorem

If $\mathbb{E}(|e|^r | X = x) = \overline{\sigma}^2 < \infty$ *for some* $r > 2$ *and* $nh^5 = O(1)$ *then*

$$\sqrt{nh}\left(\widehat{m}_{NW}(x) - m(x) - h^2 B_{NW}(x)\right) \xrightarrow{d} N\left(0, \frac{R_K \sigma^2(x)}{f(x)}\right)$$

$$\sqrt{nh}\left(\widehat{m}_{LL}(x) - m(x) - h^2 B_{LL}(x)\right) \xrightarrow{d} N\left(0, \frac{R_K \sigma^2(x)}{f(x)}\right)$$

- Notice rate of convergence is $\sqrt{nh}$ rather than $\sqrt{n}$. It is slower as $h \to 0$.
- Therefore nh can be intepreted as effective sample size.
- If rate of convergence of h is faster then optimal $n^{-\frac{1}{5}}$ then $h = o\left(n^{-\frac{1}{5}}\right)$ and bias terms in above theorem can be ignored.

Variance estimation and confidence bounds

- Nonparametric variance model

$$\sigma^2(x) = var(e|X = x) = E\left(e^2 \middle| X = x\right)$$

- This can be estimated nonparametrically e.g. with *NW* estimator

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \widetilde{e}_i^2}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

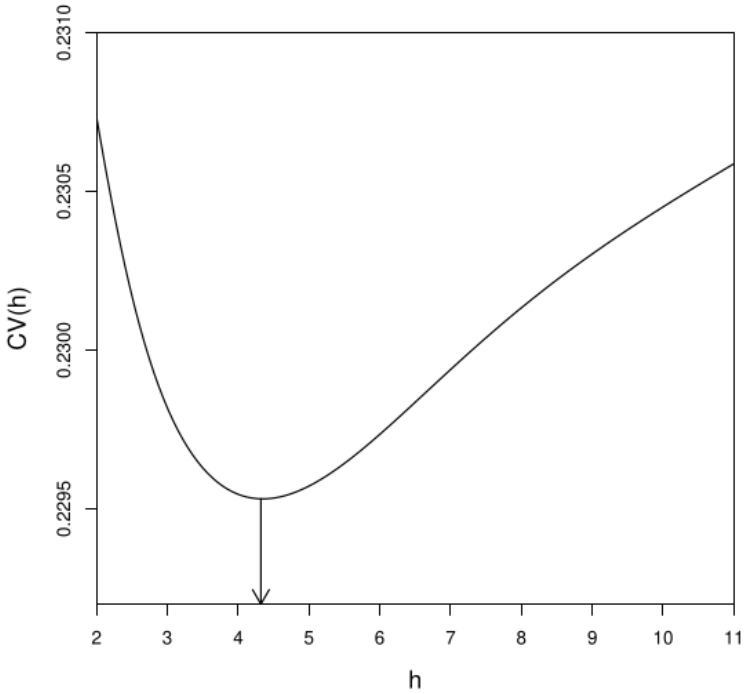
- Variance estimator

$$V_{\widehat{m}(x)} = \frac{R_K \widehat{\sigma}^2(x)}{nh\widehat{f}(x)}$$

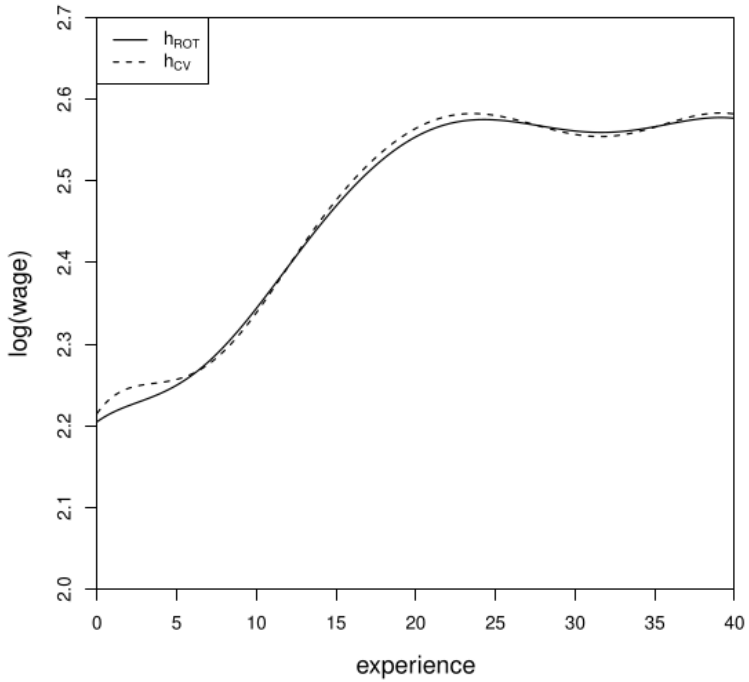
where $\widehat{f}(x)$ is nonparametric density estimator e.g. $\widehat{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right)$ and b is bandwidth

- 95% pointwise confidence interval: $\widehat{m}(x) \pm 1.96\sqrt{V_{\widehat{m}(x)}}$

log(wage) regression on experience



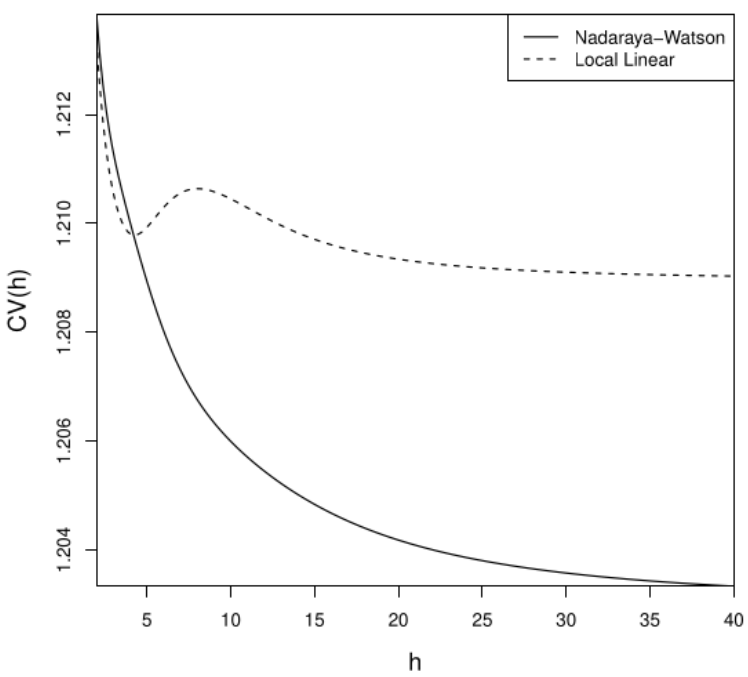
(a) Cross-Validation Criterion



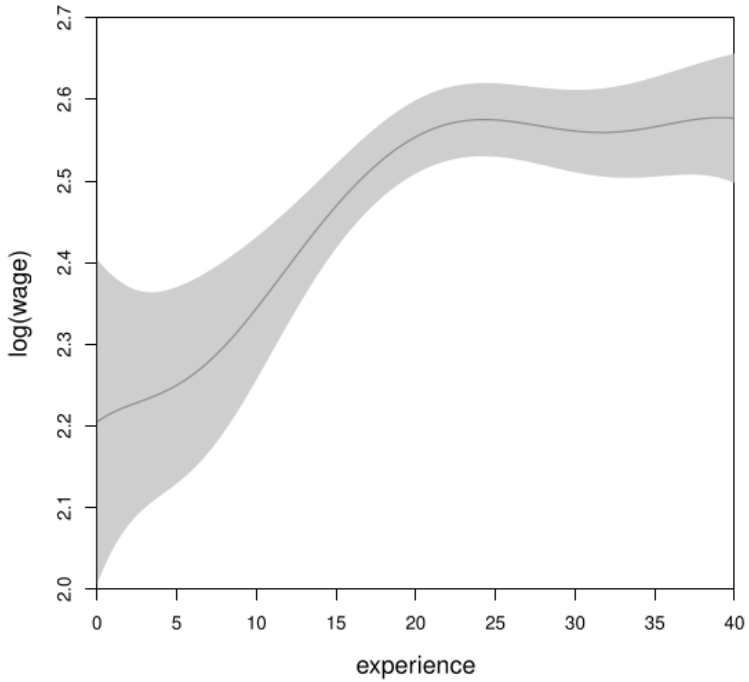
(b) Local Linear Regression

Source: Hansen (2022)

Confidence bands construction



(a) Cross-Validation for Conditional Variance



(b) Regression with Confidence Bands

Source: Hansen (2022)

Multiple regressors

- Vector valued $X = (X_1, \ldots, X_d)'$
- Kernel weights for observation i:

$$K_i(x) = K\left(\frac{X_{1i} - x_1}{h_1}\right) K\left(\frac{X_{2i} - x_2}{h_2}\right) \ldots K\left(\frac{X_{di} - x_d}{h_d}\right)$$

- Nadaraya-Watson kernel estimator:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K_i(x) Y_i}{\sum_{i=1}^{n} K_i(x)}$$

- Local linear estimator $\widehat{m}(x) = \widehat{\alpha}(x)$

$$\begin{pmatrix} \widehat{\alpha}(x) \\ \widehat{\beta}(x) \end{pmatrix} = (\mathbf{Z'KZ})^{=1} \mathbf{Z'KY}$$

  where $\mathbf{K} = diag(K_1(x), K_2(x), \ldots, K_d(x))$
- Choice of the bandwith is based on cross-validation criterion, where leave-one-out residuals $\widetilde{e}_i$ are defined in the same way as in one regressor case.
- However, the minimisation problem is more complicated as we maximise over $h_1, h_2, \ldots, h_d$.

Curse of dimensionality

- For vector-valued X:

$$AIMSE = h^4 \int_S \left(\sum_{i=1}^{d} B_j(x) \right)^2 f(x) w(x) dx + \frac{R_K^d}{nh^d} \int_S \sigma^2(x) w(x) dx$$

- The bias is of order h^4 but variance is of order $\left(nh^d\right)^{-1}$ and then $AIMSE$ depends on d.
- It can be shown that bandwith h which minimises $AIMSE$ is of order $h \sim n^{-\frac{1}{4+d}}$.
- For such a bandwith $AIMSE = O\left(n^{-\frac{4}{4+d}}\right)$
- Therefore the rate for convergence decreases with the number of regressors.
- The reason for this is that with more dimensions, the number of observations that are close to a given value of x is inherently smaller.

Partially linear regression, Robinson (1988)

- Assume that

$$Y_i = m(X) + Z'\beta + e$$

$$E(e|X, Z)$$

- The conditional mean is separable between X and Z (no nonparametric interactions)
- Expectation of structural equation with respect to X is:

$$E(Y|X) = m(X) + E(Z|X)'\beta$$

- Substracting this from structural model we obtain

$$Y_i - E(Y|X) = (Z - E(Z|X))'\beta + e$$

- Procedure:
  1. Regress nonparametrically Y_i on X and Z on X, obtain fitted values $\widehat{g}_{0i}, \widehat{g}_{1i}, \ldots, \widehat{g}_{ki}$
  2. Regress $Y_i - g_{0i}$ on $Z_1 - \widehat{g}_{1i} \ldots, Z_k - \widehat{g}_{ki}$ and obtain estimate of satandard errors and $\widehat{\beta}$
  3. Use nonparametric regression of $Y_i - Z_i\widehat{\beta}$ on X_i to obtain estimate of $\widehat{m}(x)$ and confidence intervals

Series regression

- Model which is considered:

$$y = m(x) + e$$
$$\mathbb{E}(e|X) = 0$$
$$\mathbb{E}\left(e^2 \middle| X\right) = \sigma^2(X)$$

- Linear series regression is of the form:

$$Y = \tau_1(x)\beta_1 + \tau_2(x)\beta_2 + \ldots + \beta_K\tau_k(x) + e = X_K'\beta_K + e_K$$

where $X_K = X_K(X)$ is a vector of basis transformations $\tau_j(X)$ of X.

- The vector of parameters β_K is usually estimated with *OLS*.
- Often it is difficult to estimate such regression because of high correlation between elements of X_K. In such cases we use ortogonalized polynomials (sample ortogonalization, algebraic orthognal polynomials)
- Estimator of $m(x)$

$$\widehat{m}(x) = X_K'\widehat{\beta}_K$$

Polynomial and spline regressions

- Polynomial regression

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p = X_K' \beta_K$$

  and $X_K(X) = (1, X, X^2, \ldots, X^p)$
- Spline is a piecewise polynomial
- Spline regression

$$m_k(x) = \sum_{j=0}^{p} \beta_j x_j + \sum_{k=1}^{N} \beta_{p+k} (x - \tau_k)^p 1\{x \geq \tau_k\}$$

  where τ_k are locations of knots.
- We can generalize this model to partially linear model by adding linear regressors.
- We would like to minimise standard error:

$$ISE(K) = \int (\widehat{m}_k(x) - m(x))^2 dF(x)$$

Polynomial and spline regressions

- It can be shown that under mild assumptions and $n, K \to \infty$

$$ISE(K) = O_p\left(K^{-4} + \frac{K}{n}\right)$$

- Then $K \sim n^{\frac{1}{5}}$ optimizes $ISE(K)$ and for such $ISE(K) \leq O_p\left(n^{-\frac{4}{5}}\right)$.
- However, for constant K the estimator is biased.
- Define:

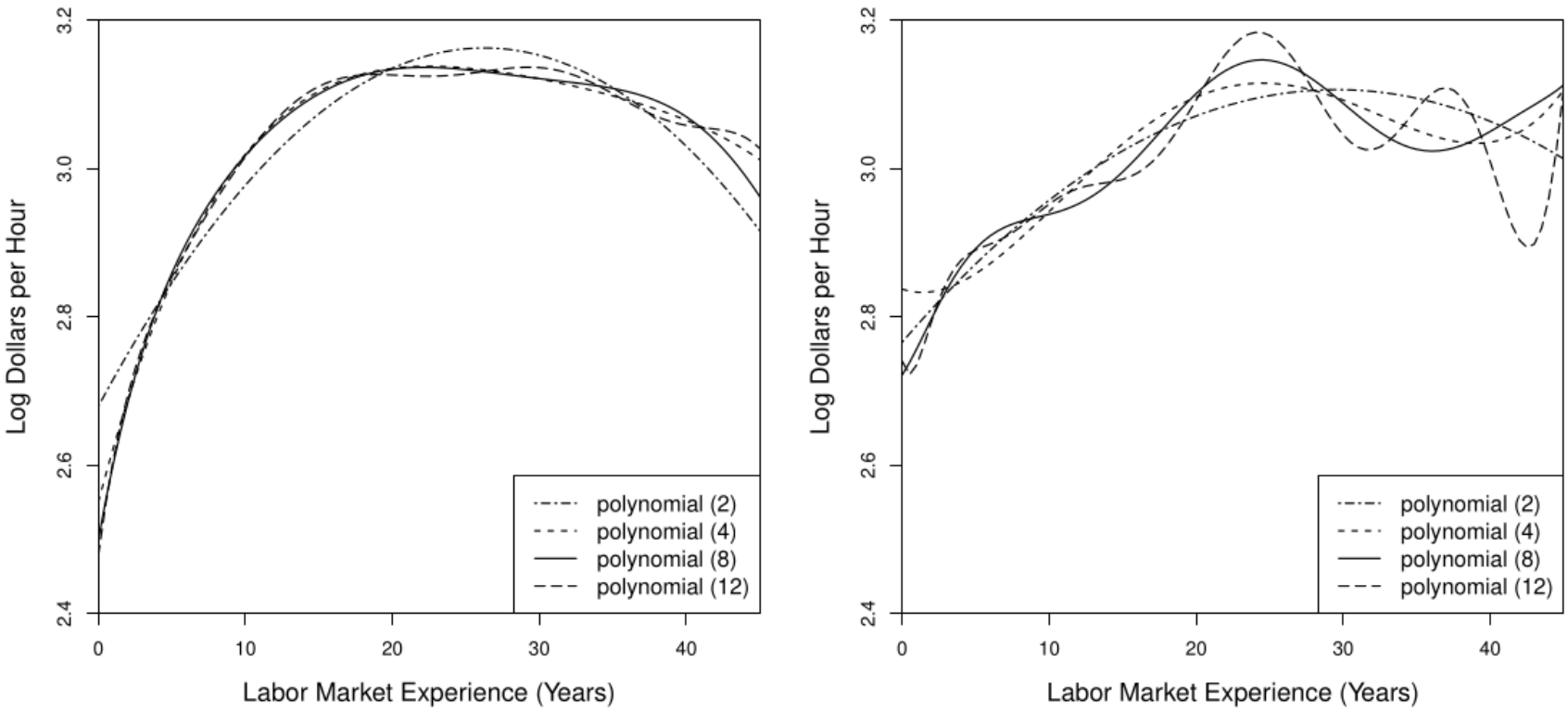
$$\delta_K^{*2} = \inf_{\beta} \sup_{x \in S} \left| m(x) - X_K' \beta \right|$$

- It is possible to prove that given that if approximation error δ_K^* converges fast enough so that $n\delta_K^{*2} \to 0$

$$\frac{\sqrt{n}\left(\widehat{m}_k(x) - m(x)\right)}{V_K^{\frac{1}{2}}(x)} \xrightarrow{d} N(0,1)$$

- In practice K is choosen using cross-validation criterion $CV(K) = \sum_{i=1}^{n} \widetilde{e}_{Ki}^2$, that is we choose such K which is minimising variance of leave-one-out residuals.

Polynomial regression



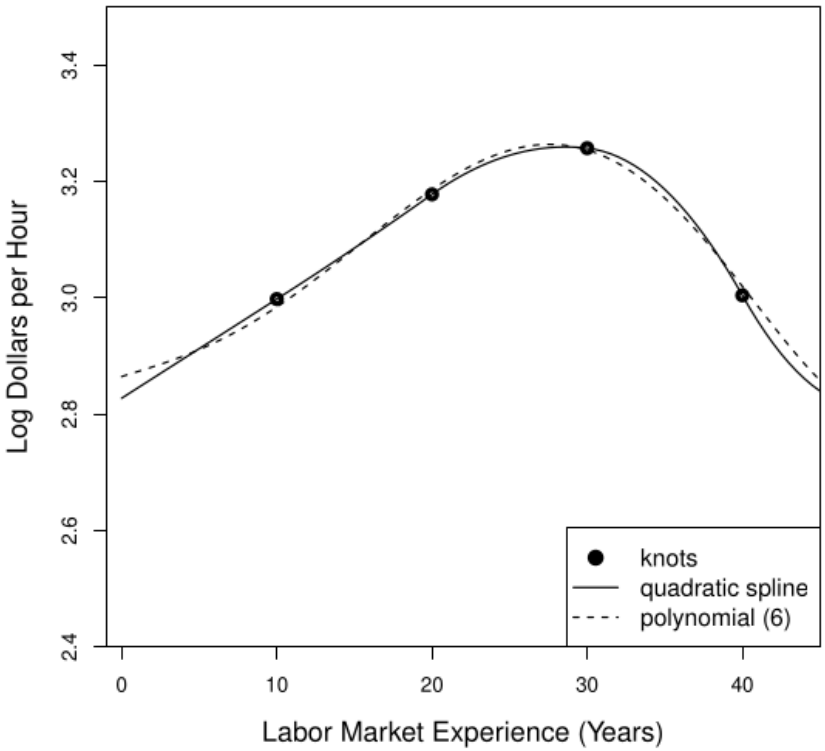
(a) White Women

(b) Black Women

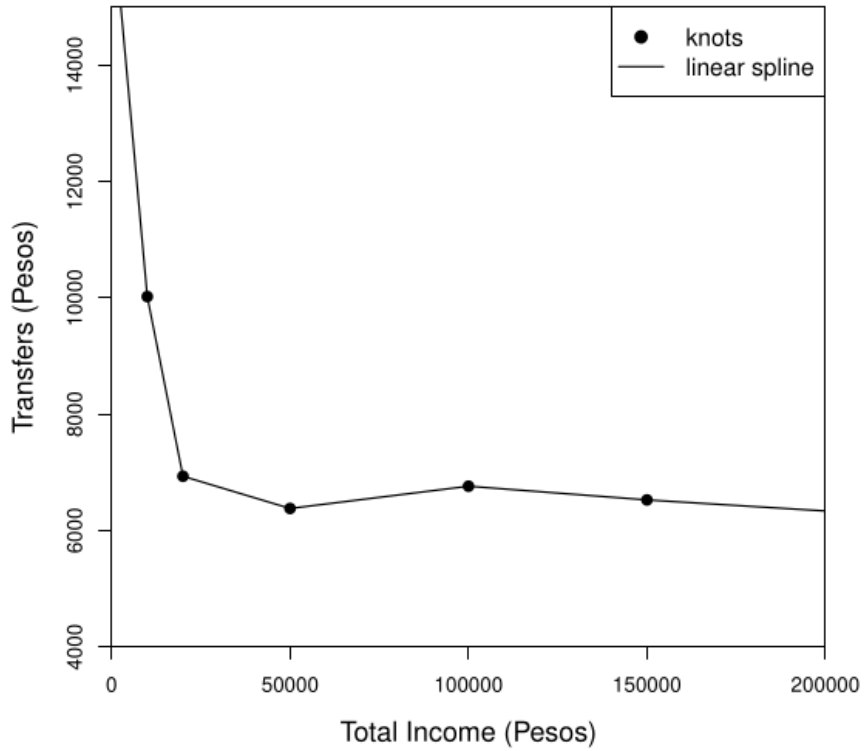
Figure 20.1: Polynomial Estimates of Experience Profile, College-Educated Women

Source: Hansen (2022)

Spline regression



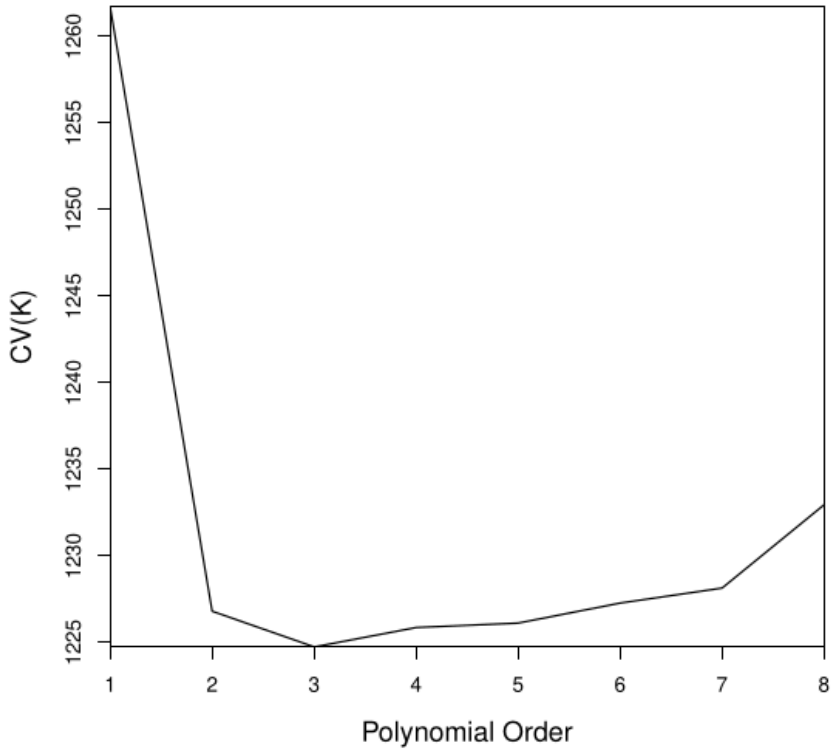
(a) Experience Profile



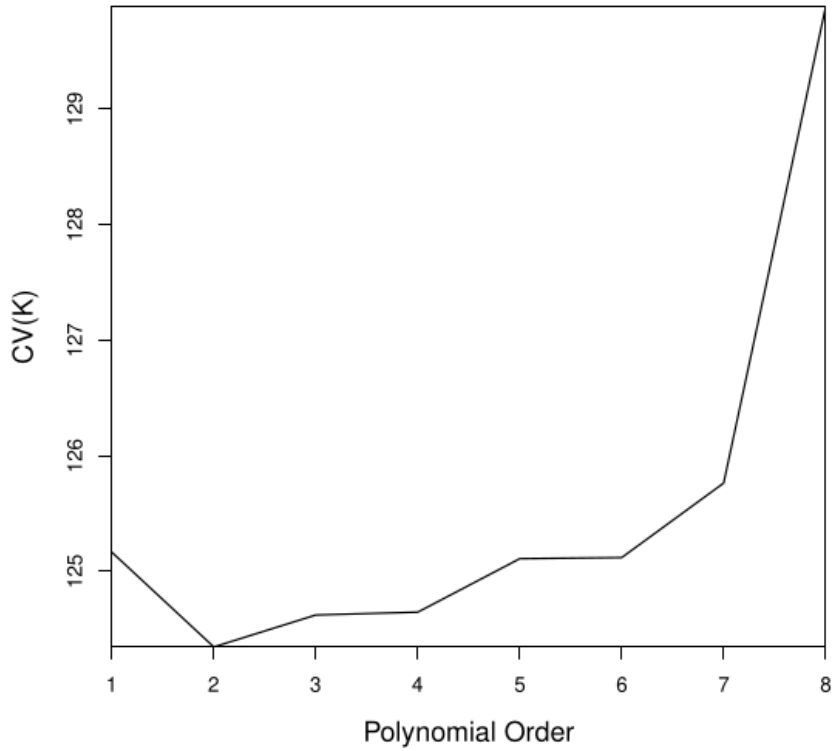
(b) Effect of Income on Transfers

Source: Hansen (2022)

Polynomial regression, CV criterion



(a) White Women



(b) Black Women

Source: Hansen (2022)

Confidence bounds, polynomial regression

Source: Hansen (2022)

References I

- Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN: 9780412983214. URL: https://books.google.pl/books?id=BM1ckQKCXP8C.
- Hansen, B. (2022). *Econometrics*. Princeton University Press. ISBN: 9780691235899.
- Qin, Zean (2026). URL: https://zean.be/articles/a-simple-introduction-of-kernel-density-estimation/.
- Robinson, P. M. (1988). "Root-N-Consistent Semiparametric Regression". In: *Econometrica* 56.4, pp. 931–954. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912705 (visited on 04/20/2026).