ATS Estimation & Diagnostics for Nongaussian Models

ATS = Average-Transform-Smooth

- Average observations by taking means
- Transform the averages
- Smooth the transformed averages by a nonparametric or parametric modeling

This has elements of methods that go back many decades in statistics

Variance stabilizing transformations when the variance is a function of the mean

Example: Taking the square root of Poisson observations with changing means makes the variance approximately constant

ATS Estimation & Diagnostics for Nongaussian Models

But we want to Gaussianize the data as well, also to make analysis simpler, especially model diagnostics

Averaging the data helps this as well as transformation

Work on fundamentals: can we make things Gaussian

Cleveland, W. S. and Mallows, C. L. and McRae J. E.
ATS Methods: Nonparametric Regression for Nongaussian Data
Journal of the American Statistical Association, Volume = 88, 821-835, 1993

ATS for Power Spectra

Example: Non-parametric estimation of the power spectrum from the periodogram ordinates $I(f_k)$ at the Fourier frequencies f_k

Ordinates divided by the power spectrum, $S(f_k)$, are close to independent chi-squared with 2 df

Take means of periodogram ordinates at successive blocks of frequencies of of size b

Take logs which leaves us with $\log(\bar{I}(j)) - \log(S\bar{f}_j)$, nicely additive

ATS for Power Spectra

What is the smallest value of b that will do a reasonable job of gaussianizing the log block means

The answer is $b = 4$, which is a very exciting result

Typically, in practice, makes the bias of averaging tolerable

We can treat estimation as a gaussian regression and, say, smooth with loess, and do regression diagnostics

The model is $\log(\bar{I}(j)) = \log(S\bar{f}_j)$ + ϵ_j where errors ϵ_j are i.i.d. gaussian

We can use the full spectrum of regression diagnostics for nonparametric estimation of the power spectrum

The residuals are $\hat{\epsilon}_j = \log(\bar{I}(j)) - smooth(\log(\bar{I}(j))$

The ed Method for Nonparametric Univariate Density Estimation with Diagnostic Checking

(e)stimation and (d)iagnostics

Collaboration with Ryan Hafen

Density estimation goes back to the 50's

Mostly kernel density estimation

A seminal paper by Murray Rosenblatt in 1956

David Scott has made major contributions to multivariate density estimation

But, we need much more powerful methods of diagnostic checking to see if the density patterns are faithfully represented

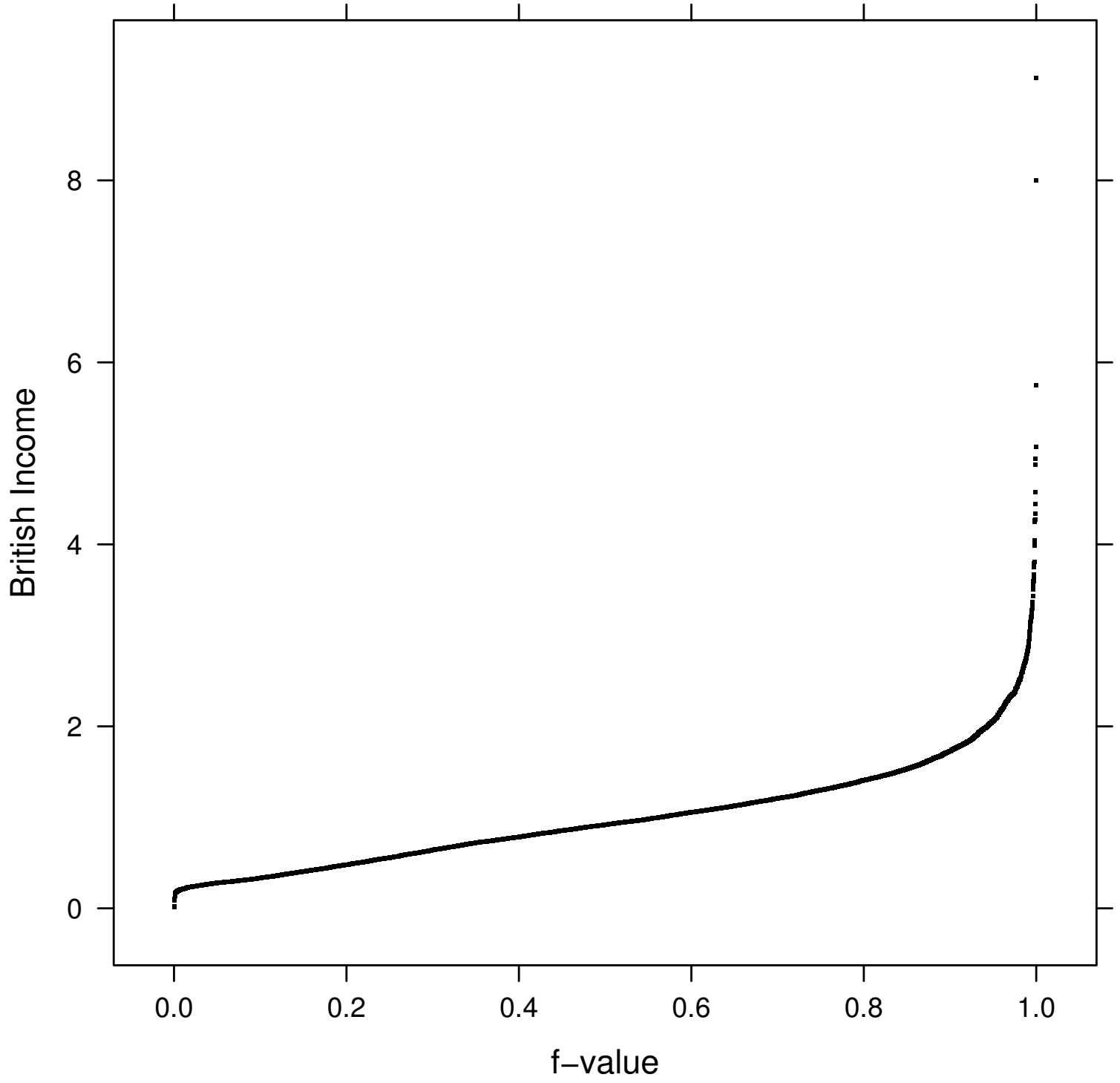
The ed Method for Density Estimation

Enables taking a model building approach

As with power spectrum estimation, turns nonparametric density estimation into a gaussian regression analysis

And this enables the power of gaussian regression diagnostics

Normalized British Income Data (7201 observations)



Order Statistics and Their Gaps

x_j normalized pounds sterling (nps) for j = 1 to $m = 7162$, ordered from smallest to largest

Order statistic κ-gaps:

$$\begin{aligned} g_1^{(\kappa)} &= x_{\kappa+1} - x_1 \\ g_2^{(\kappa)} &= x_{2\kappa+1} - x_{\kappa+1} \\ g_3^{(\kappa)} &= x_{3\kappa+1} - x_{2\kappa+1} \\ &\vdots \end{aligned}$$

For $\kappa = 10$:

$$\begin{aligned} g_1^{(10)} &= x_{11} - x_1 \\ g_2^{(10)} &= x_{21} - x_{11} \\ g_3^{(10)} &= x_{31} - x_{21} \\ &\vdots \end{aligned}$$

Gaps have units nps

Number of observation in each interval is κ

Balloon Densities

Gaps: $g_i^{(\kappa)} = x_{i\kappa+1} - x_{(i-1)\kappa+1}, \quad i = 1, 2, \ldots, n$

$$b_i^{(\kappa)} = \frac{\kappa/m}{g_i^{(\kappa)}} \quad \frac{\text{fraction of observations}}{\text{nps}}$$

$g_i^{(\kappa)}$ is positioned at the midpoint of the gap interval $\left[x_{(i-1)\kappa+1}, x_{i\kappa+1}\right]$

$$x_i^{(\kappa)} = \frac{x_{i\kappa+1} + x_{(i-1)\kappa+1}}{2} \quad \text{nps}$$

Now κ is fixed and we think of $g_i^{(\kappa)}$ as a random variable

A Very Attractive Property of the Log Balloon Estimate

$$y_i^{(\kappa)} = \log(b_i^{(\kappa)}), i = 1, \ldots, n$$

Distributional Properties: The "Theory"

"Approximately" independent and distributed like a constant plus the log of a chi-squared distribution with 2κ degrees of freedom

$$\begin{aligned} \mathsf{E}(y_i^{(\kappa)}) &= \log f(x_i^{(\kappa)}) + \log \kappa - \psi_0(\kappa) \\ \mathsf{Var}(y_i^{(\kappa)}) &= \psi_1(\kappa) \end{aligned}$$

ψ_0 = digamma function ψ_1 = trigamma function

Log Balloon Densities

Start with the log balloon densities as “the raw data”

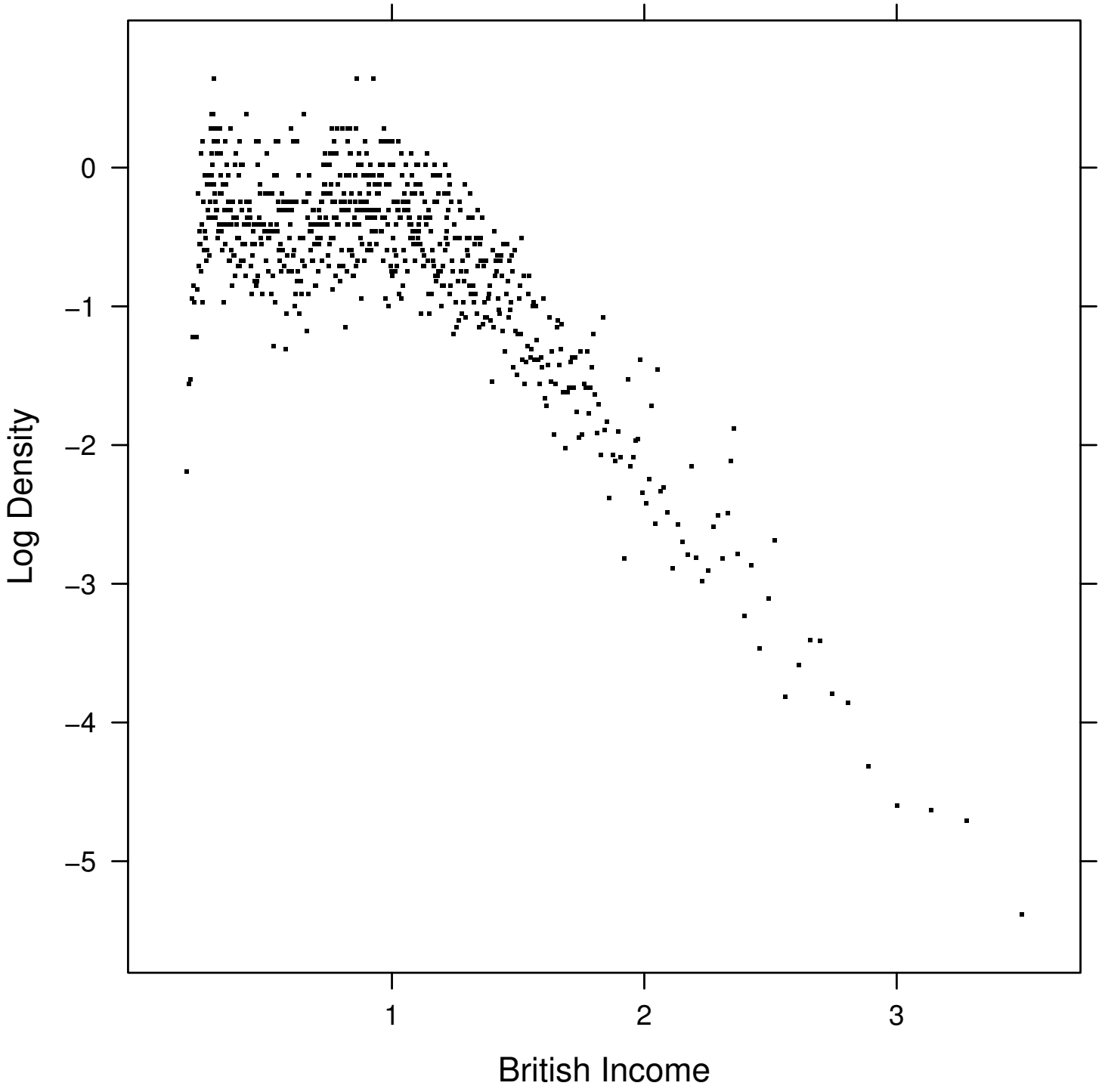
Two considerations in the choice of κ

(1) small enough that there is as little distortion of the density as possible by the averaging that occurs

(2) large enough that $y_i^{(\kappa)}$ is approximately normal

- $\kappa = 10$ is quite good and $\kappa = 20$ nearly perfect (in theory)
- we can give this up and even take $\kappa = 1$ but next steps are more complicated

Log Balloon Densities vs. Gap Midpoints $\kappa = 10$



Smooth

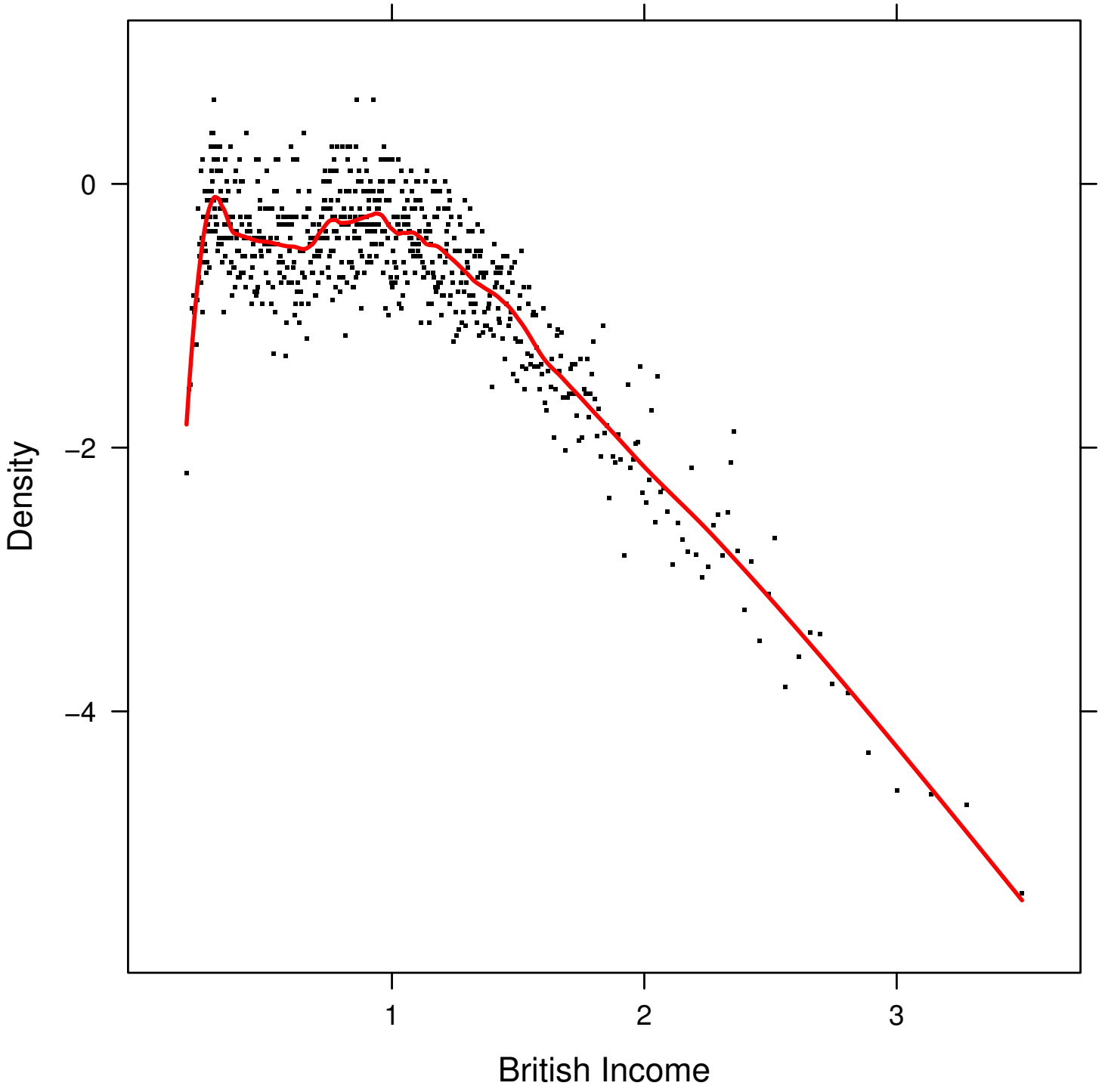
Smooth $y_i^{(\kappa)}$ as a function of $x_i^{(\kappa)}$ using loess

Fit polynomials locally of degree δ

Bandwidth parameter: $0 < \alpha \leq 1$

Fit at x uses the $[\alpha n]$ closest points to x, the neighborhood of x

Smooth Log Balloon Densities Using Loess



Three Tuning Parameters for Loess Fit

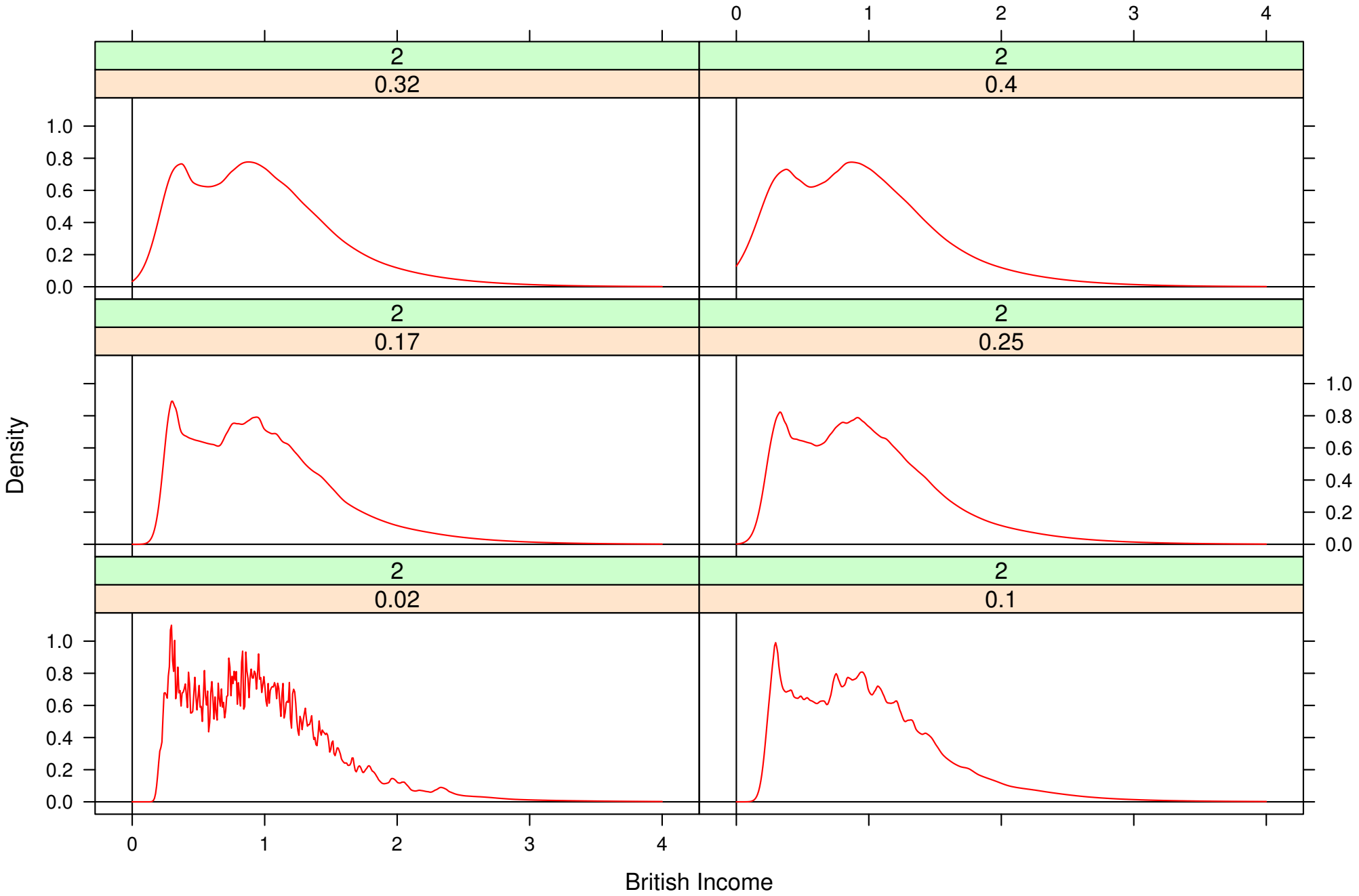
κ: gap length

α: bandwidth

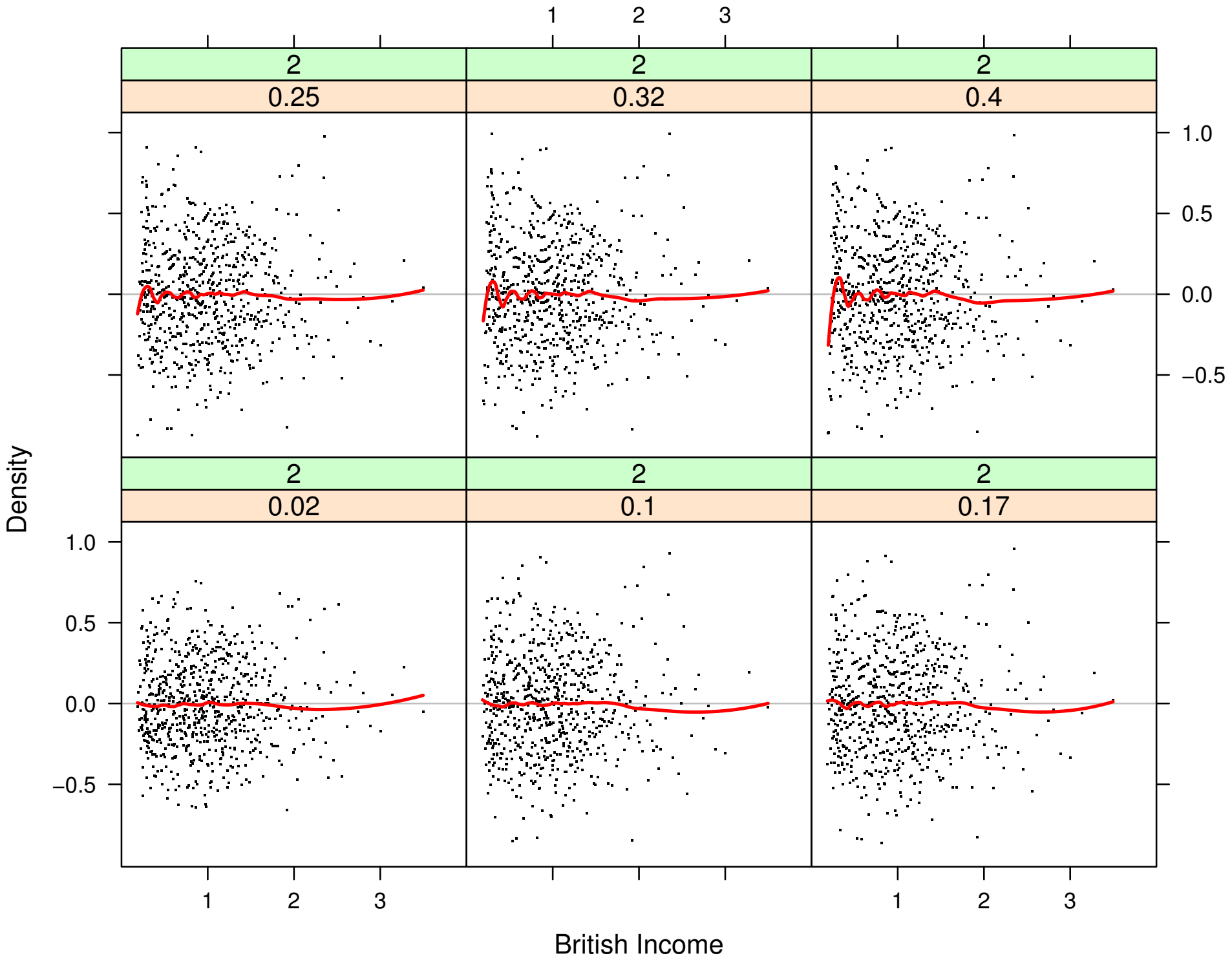
δ: degree of polynomial in local fitting

Use a model selection criterion like C_p or AIC to help chose values

Fitted Values vs. Income



Residuals vs. Income



Model Selection for British Incomes

From Cp plot, plots of residuals, and plots of fits

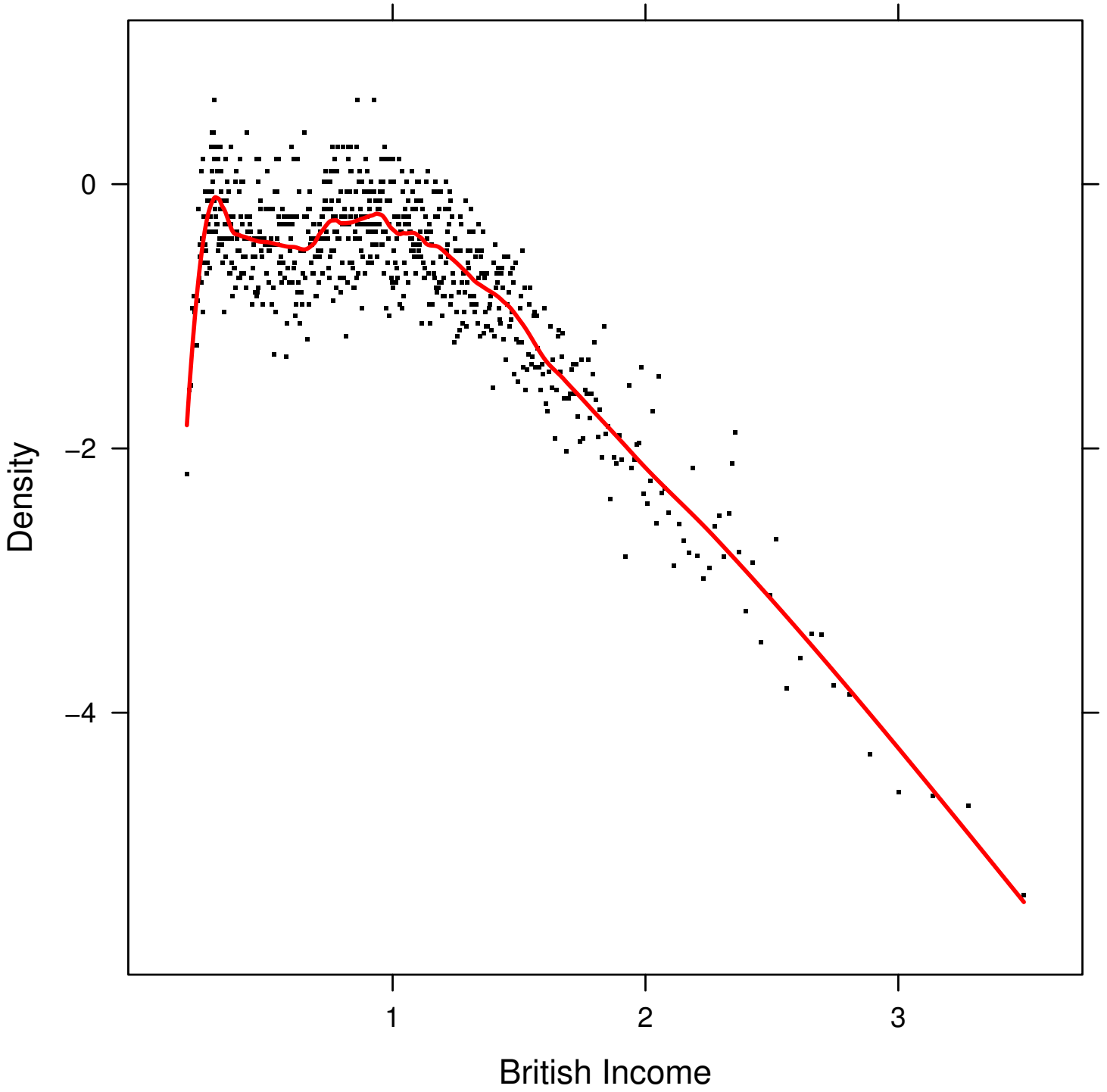
Gap size: $\kappa = 10$

Polynomial degree: $\delta = 2$

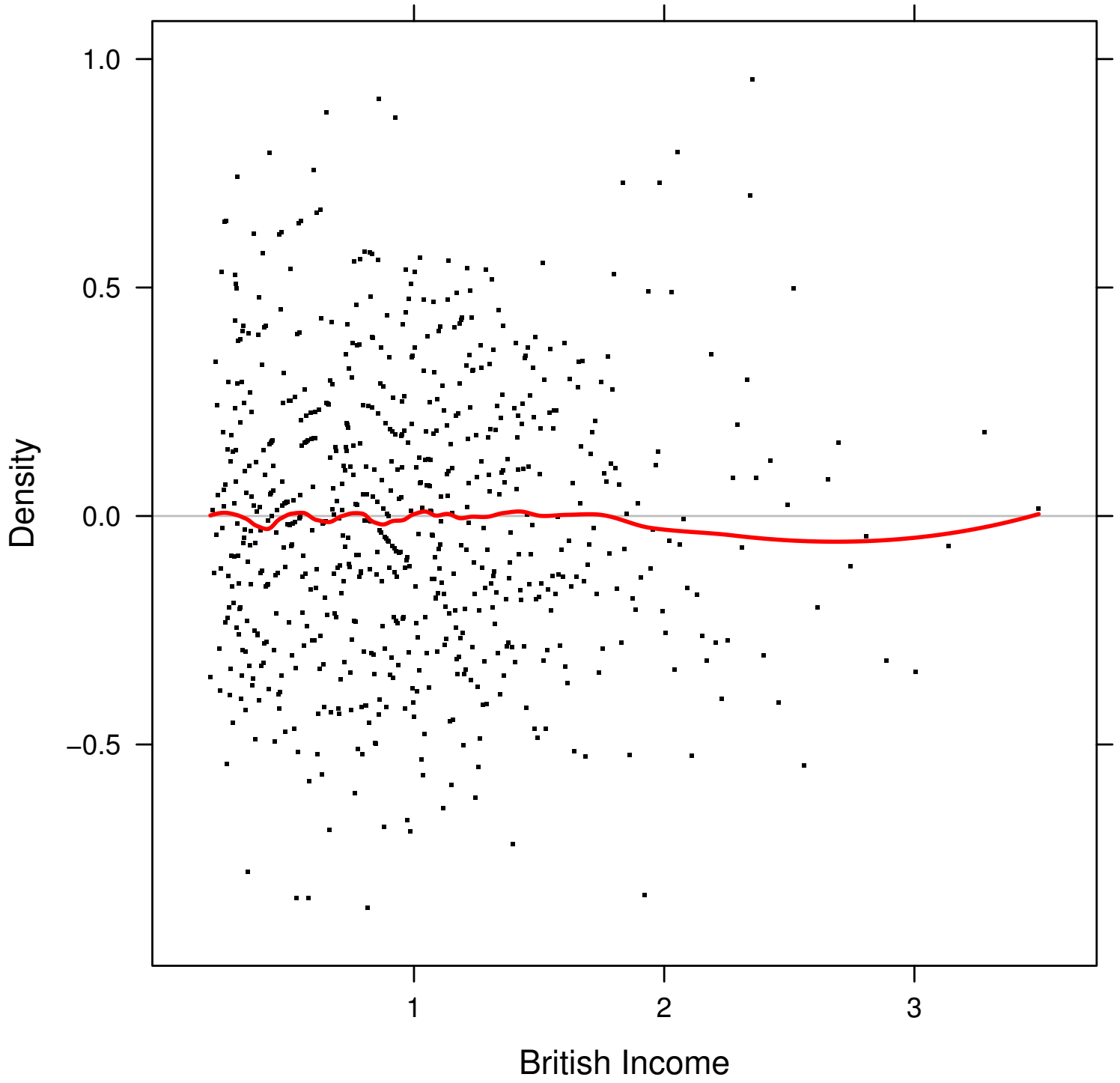
Bandwidth parameter: $\alpha = 0.16$

Equivalent degrees of freedom: $\nu = 19$

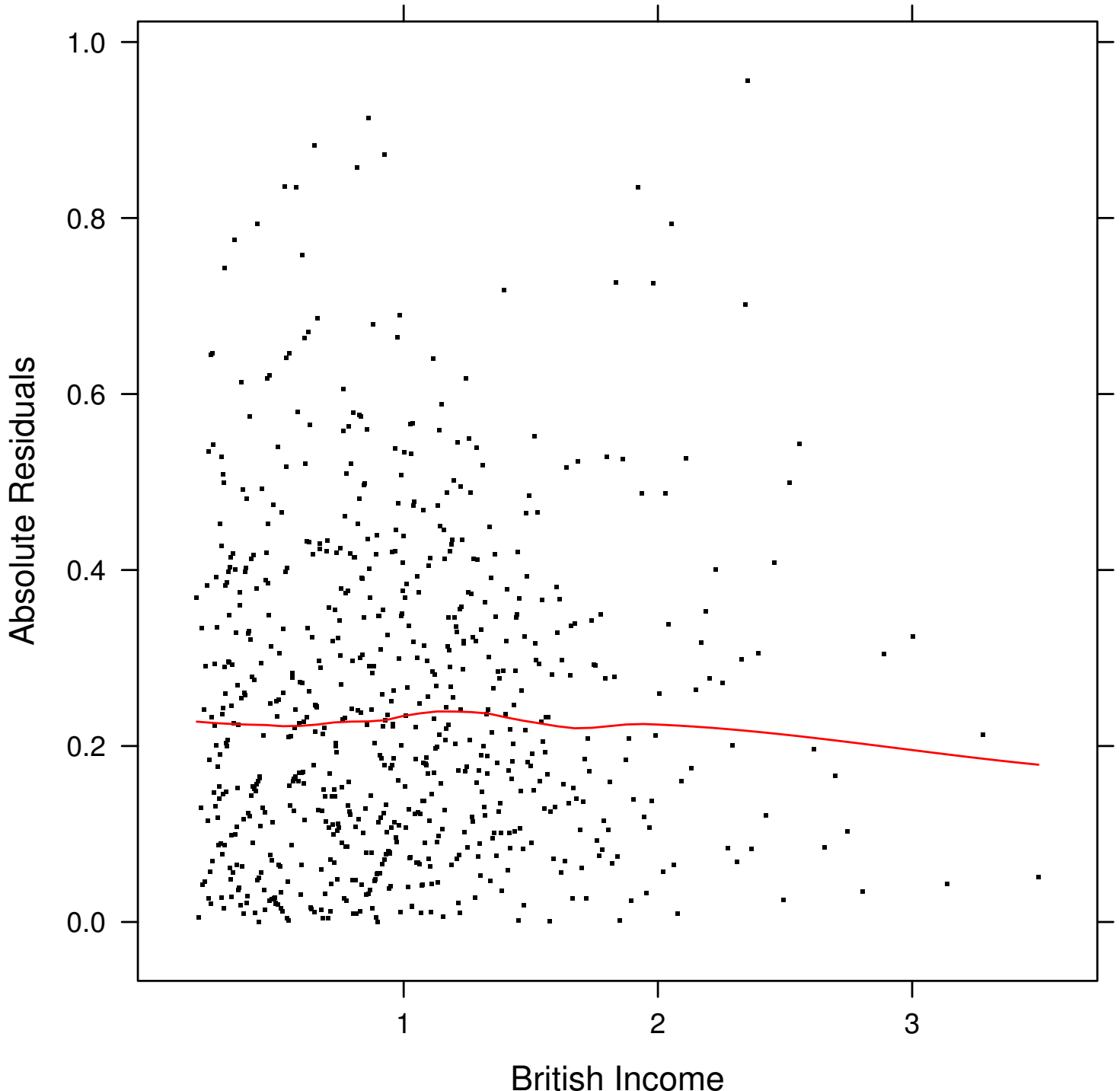
ed Log Density Estimate for Income: Fit vs. Income



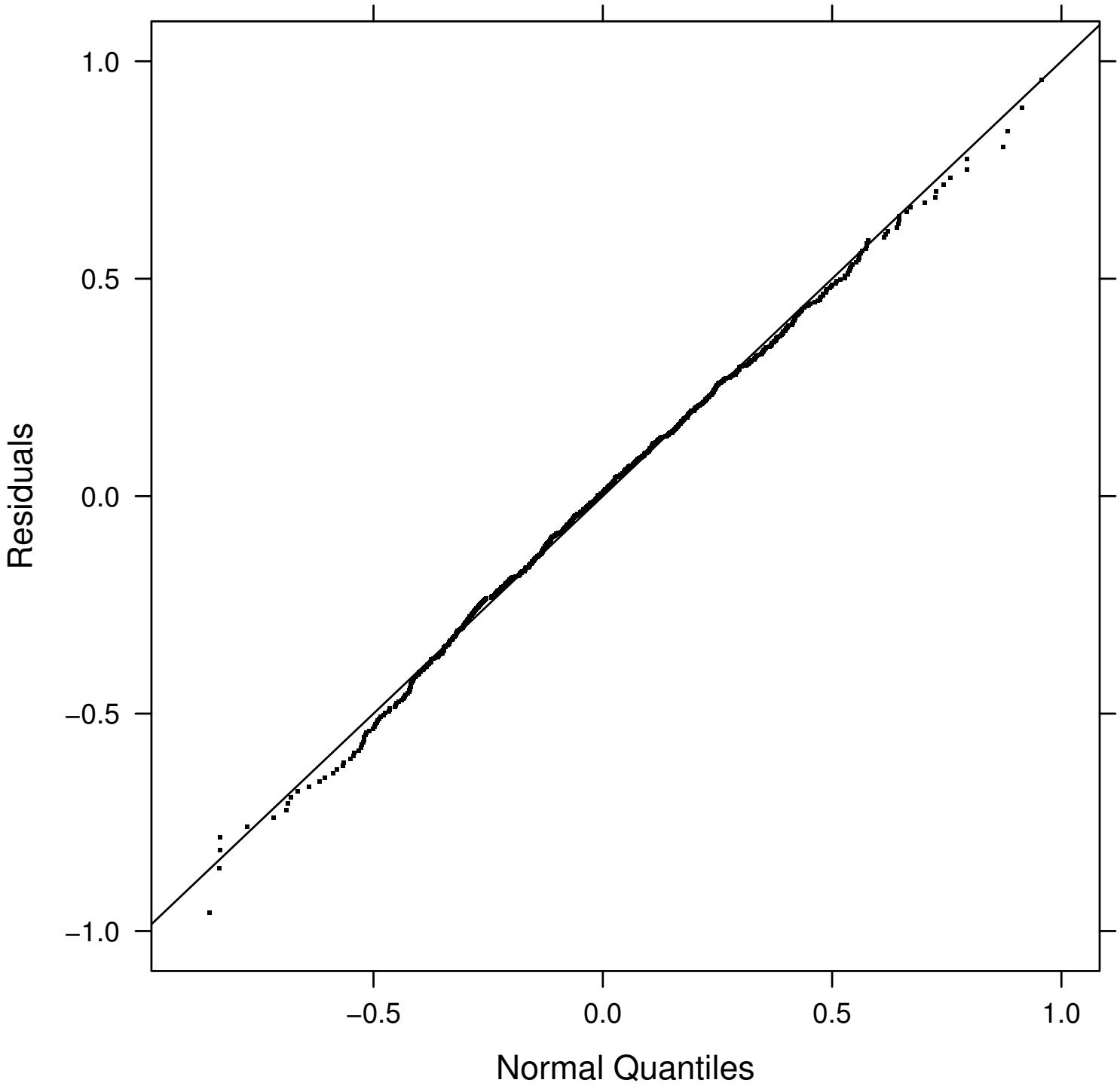
ed Log Density Estimate for Income: Residuals vs. Income



ed Log Density Estimate for Income: Absolute Residuals vs. Income



Normal Quantile Plot of Residuals



ed Log Density Estimate for Income: 99% Pointwise Confidence Intervals