Automated Learning and Data Visualization

William S. Cleveland

Department of Statistics
Department of Computer Science
Purdue University

Methods of Statistics, Machine Learning, and Data Mining

Mathematical methods

- summary statistics
- models fitted to data
- algorithms that process the data

Visualization methods

- displays of raw data
- displays of output of mathematical methods

Mathematical Methods

Critical in all phases of the analysis of data:

- from initial checking and cleaning
- to final presentation of results

Automated learning: adapt themselves to systematic patterns in data

Can carry out predictive tasks

Can describe the patterns in a way that provides fundamental understanding

Different patterns require different methods even when the task is the same

Visualization Methods

Critical in all phases of the analysis of data:

- from initial checking and cleaning to
- final presentation of results

Allow us to learn which patterns occur out of an immensely broad collection of possible patterns

Visualization Methods in Support of Mathematical Methods

Typically not feasible to carry out all of the mathematical methods necessary to cover the broad collection of patterns that could have been seen by visualization

Visualization provides immense insight into appropriate mathematical methods, even when the task is just prediction

Automatic selection of best mathematical methods

- model selection criteria
- training-test framework
- risks finding best from of a group of poor performers

Mathematical Methods in Support of Visualization

Typically not possible to understand the patterns in a data set just displaying the raw data

Must also carry out mathematical methods and then visualize

- the output
- the remaining variation in the data after adjusting for output

Tactics and Strategy

Mathematical methods exploit the tactical power of the computer:

- an ability to perform massive mathematical computations with great speed

Visualization methods exploit the strategic power of the human:

- an ability to reason using input from the immensely powerful human visual system

The combination provides the best chance to retain the information in the data

Machine Learning Algorithm Deep Blue vs. Human Kasparov

Why was Kasparov so distressed about the possibility that
IBM was cheating by allowing a human to assist the algorithm?

IBM Machine Learning System Deep Blue vs. Human Kasparov

Why was Kasparov so distressed about the possibility that
IBM was cheating by allowing a human to assist the algorithm?

He knew he had no chance to beat a human-machine combination

The immense tactical power of the IBM machine learning system

The strategic power, much less than his own, of a grand master

Conclusion

MATHEMATICAL METHODS

&

VISUALIZATION METHODS

ARE

SYMBIOTIC

Plan for this Talk

Visualization Databases for Large Complex Datasets

Just a few minutes in this talk

Paper in Journal of Machine Learning Research (AISTATS 2009 Proceedings)

Web site with live examples: ml.stat.purdue.edu/vdb/

Approach to visualization that fosters comprehensive analysis of large complex datasets

Plan for this Talk

The ed Method for Nonparametric Density Estimation & Diagnostic Checking

Current work

- describe here to make the case for a tight coupling of a mathematical method and visualization methods

Addresses the 50 year old topic of nonparametric density estimation

50 years of kernel density estimates and very little visualization for diagnostic checking to see if the density patterns are faithfully represented

A new mathematical method built, in part, to enable visualization

Results

- much more faithful following of density patterns in data
- visualization methods for diagnostic checking
- simple finite sample statistical inference

Visualization Databases for Large Complex Datasets

Saptarshi Guha

Paul Kidwell

Ryan Hafen

William Cleveland

Department of Statistics, Purdue University

Visualization Databases for Large Complex Datasets

Comprehensive analysis of large complex database that preserves the information in the data is greatly enhanced by a visualization database (VDB)

VDB

- many displays
- some with many pages
- often with many panels per page

A large multi-page display for a single display method

- results from parallelizing the data
- partition the data into subsets
- sample the subsets
- apply the visualization method to each subset in the sample, typically one per panel

Time of the analyst

- not increased by choosing a large sample over a small one
- display viewers can be designed to allow rapid scanning: animation with punctuated stops
- Often, it is not necessary to view every page of a display

Display design

- to enhance rapid scanning
- attention of effortless gestalt formation that conveys information rather than focused viewing

Visualization Databases for Large Complex Datasets

Already successful just with off-the-shelf tools and simple concepts

Can be greatly improved by research in visualization methods that targets VDBs and large displays

Our current research projects

- subset sampling methods
- automation algorithms for choosing basic display elements
- display design for gestalt formation

RHIPE

Our approach to VDBs allows embarrassingly parallel computation

Large amounts of computer time can be saved by distributed computing environments

One is RHIPE (ml.stat.purdue.edu/rhipe)

- Saptarshi Guha, Purdue Statistics
- R-Hadoop Integrated Processing Environment
- Greek for “in a moment”
- pronounced “hree pay”

A recent merging of the R interactive environment for data analysis (www.R-project.org) and the Hadoop distributed file system and compute engine (hadoop.apache.org)

Public domain

A remarkable achievement that has had a dramatic effect on our ability to compute with large data sets

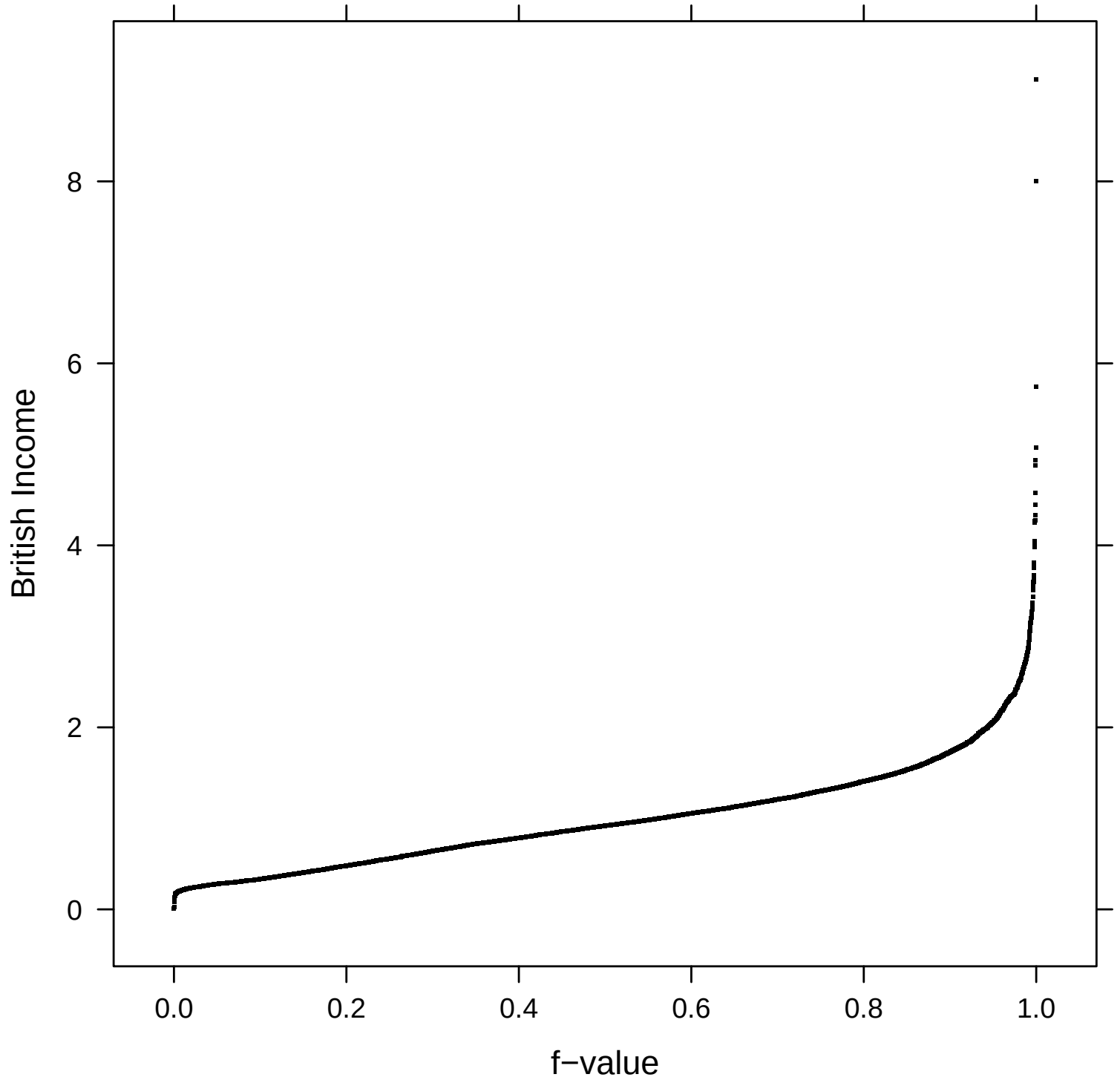
The ed Method for Nonparametric Density Estimation & Diagnostic Checking

Ryan Hafen

William S. Cleveland

Department of Statistics, Purdue University

Normalized British Income Data (7201 observations)



Kernel Density Estimation (KDE): Most-Used Method

Let x_j for j = 1 to m be the observations, ordered from smallest to largest

Fixed-bandwidth kernel density estimate with bandwidth h

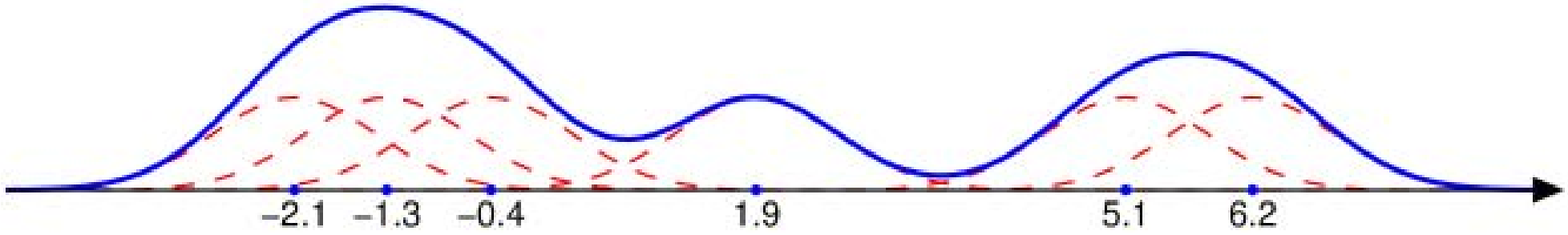
Kernel $K(u) \geq 0, \int_{-\infty}^{\infty} K(u)du = 1$

$$\hat{f}(x) = \frac{1}{mh} \sum_{j=1}^{m} K\left(\frac{x - x_j}{h}\right)$$

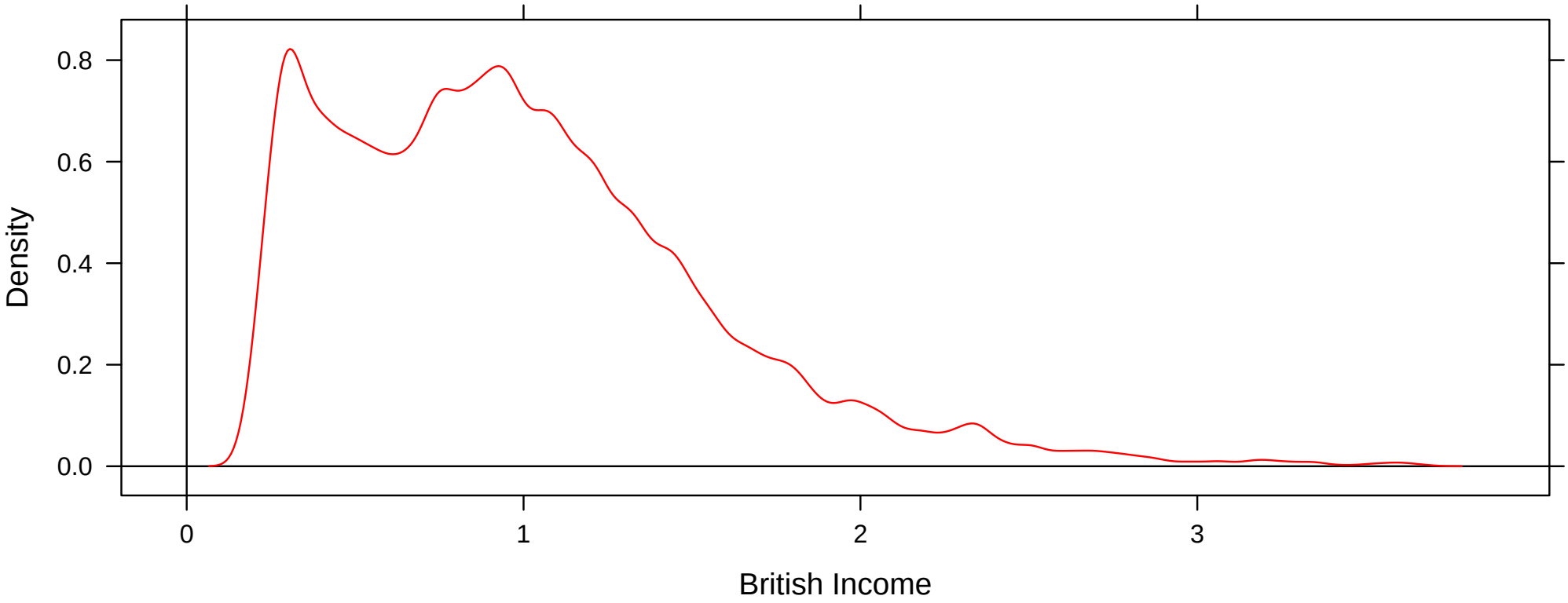
K often unit normal probability density, so h = standard deviation

x_j closer to x adds more to $\hat{f}(x)$ than a further x_j

As the bandwidth h increases, $\hat{f}(x)$ gets smoother



KDE for Incomes: Gaussian Kernel, Sheather-Jones h



Three Problems of Nonparametric Density Estimation: Problem 1

We want

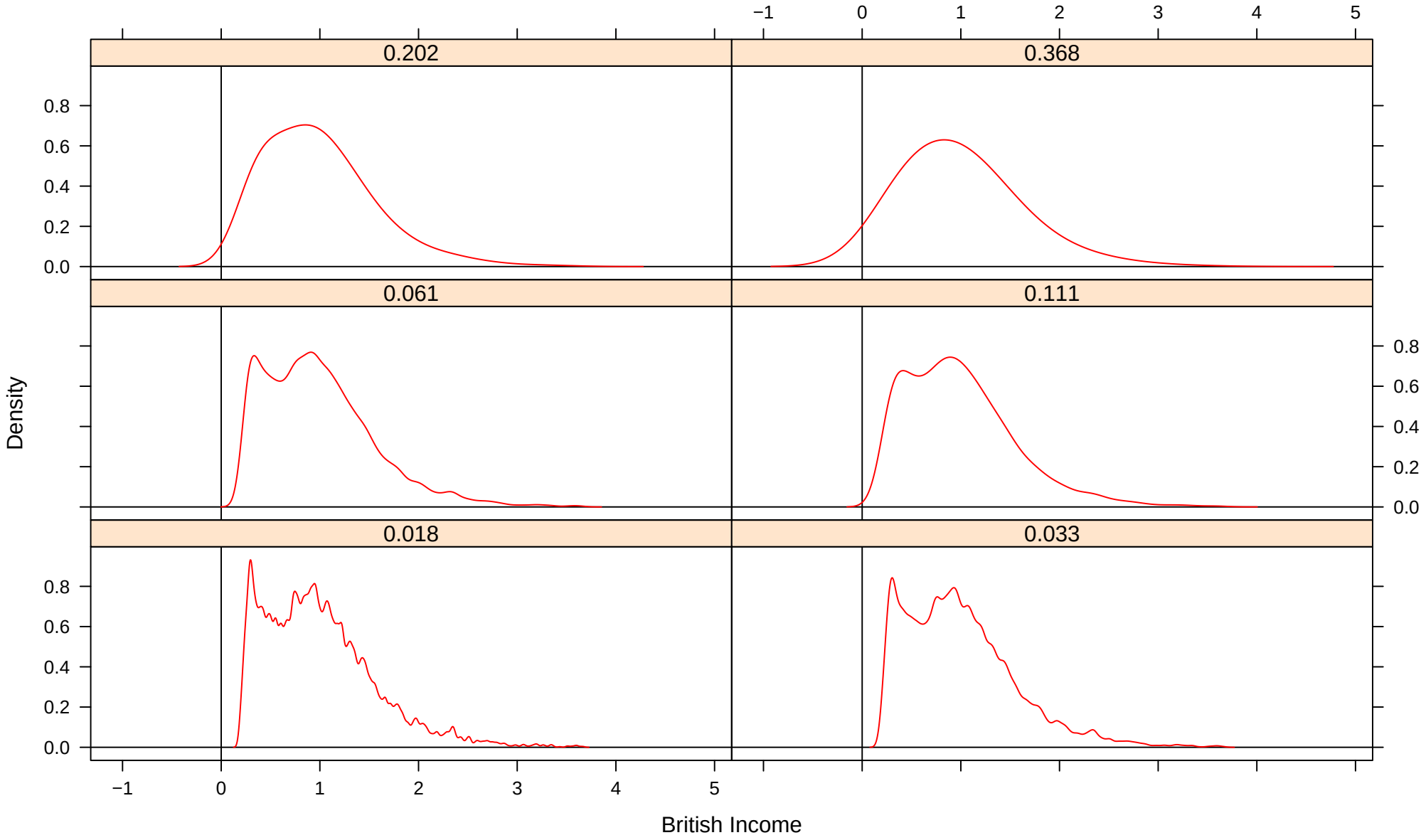
- an estimate to faithfully follow the density patterns in the data
- tools that convince us this is happening with our current set of data

There are a few reasonable things

- plot estimates with different bandwidths, small to large
- model selection criteria
- SiZer (Chaudhuri and Marron)

Problem 1: There does not exist a set of comprehensive diagnostic tools.

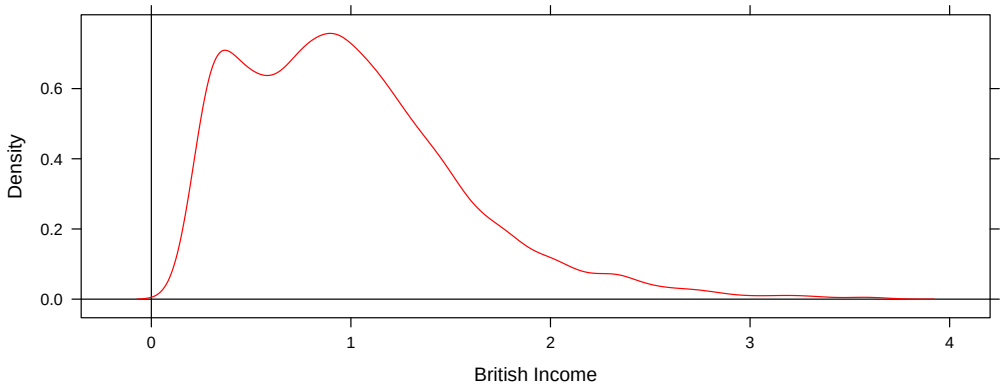
Kernel Density Estimates with Gaussian Kernel for Income Data



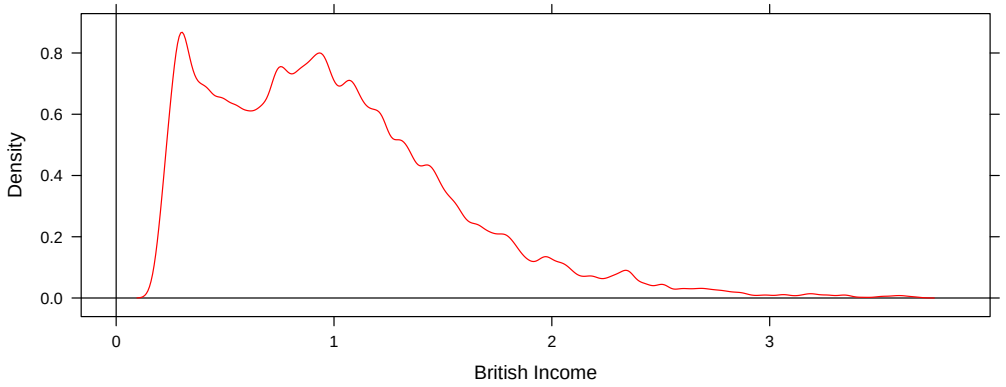
Bandwidth Selection Criterion

Treated as if the criterion automatically provides a good fit to patterns

Silverman



Cross-Validation



There are many criteria

Choosing h replaced by choosing a criterion

Useful, but only a small part of comprehensive diagnosis

Three Problems of Nonparametric Density Estimation: Problem 2

KDEs are simple and can be made to run very fast, which make us want to use them

Price for the simplicity

Discussed extensively in the literature

Three Problems of Nonparametric Density Estimation: Problem 2

$$\mathrm{E}(\hat{\mathrm{f}}(\mathrm{x})) = \int \mathrm{f}(\mathrm{x}-\mathrm{u})\mathrm{K}(\mathrm{u})\mathrm{du}, \quad \mathrm{K}(\mathrm{u}) \geq 0$$

This expected value can be far from $f(x)$

- chop peaks
- fill in valleys
- underestimate density at data boundaries when there is a sharp cutoff in density (e.g., the world's simplest density, a uniform)

General assessment: remedies such as changing h or K with x do not fix the problems, and introduce others

Problem 2: KDEs have much trouble following faithfully the density patterns in data.

The Three Problems of Nonparametric Density Estimation: Problem 3

Problem 3: Very little technology for statistical inference.

Regression Analysis (Function Estimation)

Not the impoverished situation of nonparametric density estimation

For example the following model:

$$y_i = g(x_i) + \epsilon_i$$

- y_i for i = 1 to n are measurements of a response
- x_i is a p-tuple of measurements of p explanatory variables
- ϵ_i are error terms: independent, identically distributed with mean 0

Regression analysis has a wealth of models, mathematical methods, visualization methods for diagnostic checking, and inference technology

The ed Method for Density Estimation

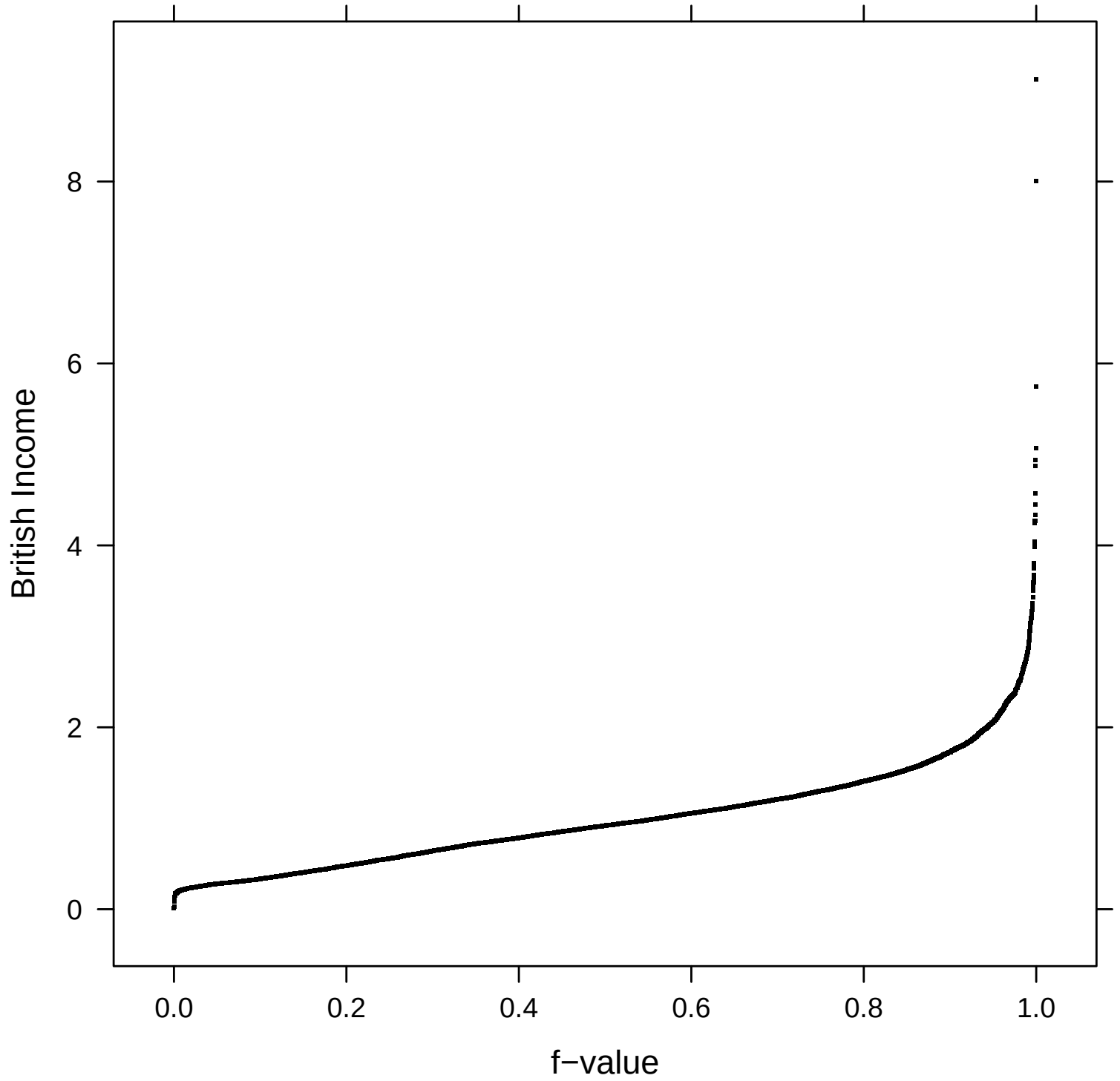
Take a model building approach

Turn density estimation to a regression analysis to exploit the rich environment for analysis

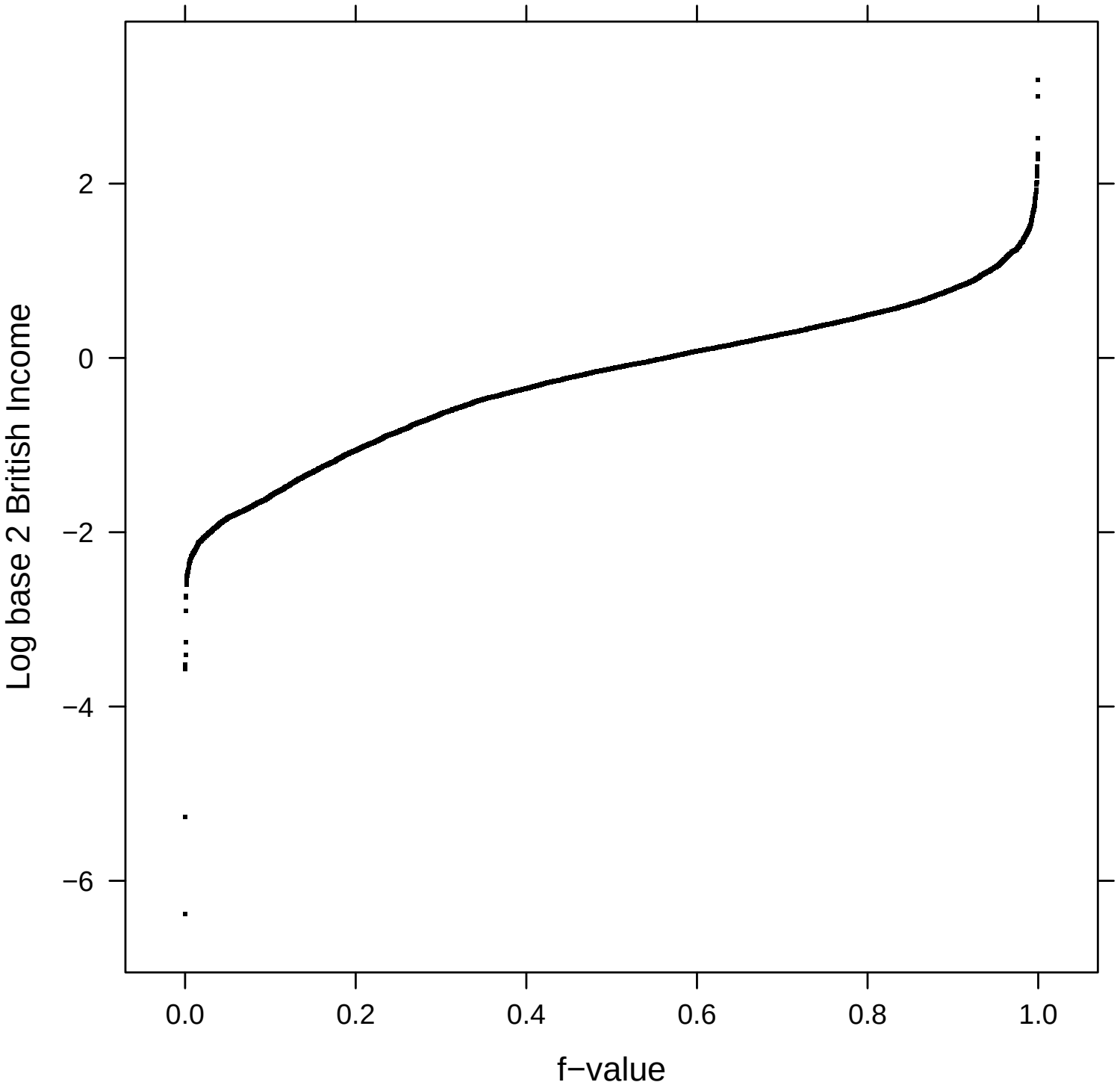
In addition seek to make the regression simple by fostering ϵ_i that have a normal distribution

In fact, it is even more powerful than most regression settings because we know the error variance

British Income Data: The Modeling Begins Right Away



Log British Income Data: The Modeling Begins Right Away



British Income: The Modeling Begins Right Away

Observations used for nonparametric density estimation

- income less than 3.75
- log base 2 income larger than -2.75
- reduces the number of observations by 39 to 7162

Not realistic to suppose we can get good relative density estimates in these tails by nonparametric methods

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

Histogram of British Incomes

Consider a histogram interval with length g

Estimate of density for the interval is

$$\frac{\kappa/m}{g} \quad \frac{\text{fraction of observations}}{\text{nps}}$$

g is fixed and think of κ as a random variable

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

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

$$= \frac{\kappa}{x_{i\kappa+1} - x_{(i-1)\kappa+1}} \quad \frac{\text{fraction of observations}}{\text{nps}}$$

$g_i^{(\kappa)}$ 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

ed Step 1: Log Balloon Densities

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

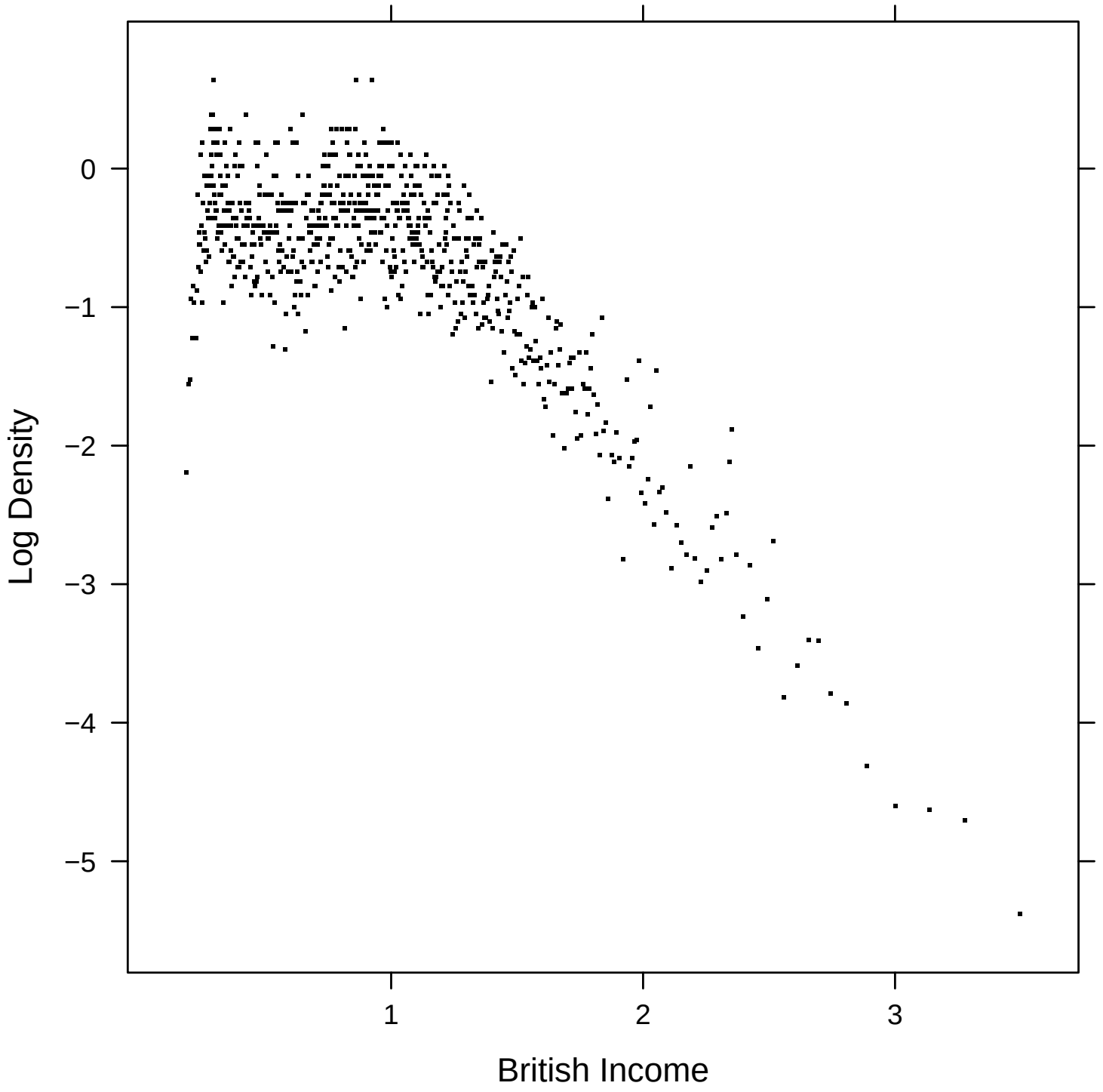
Two considerations in the choice of κ

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

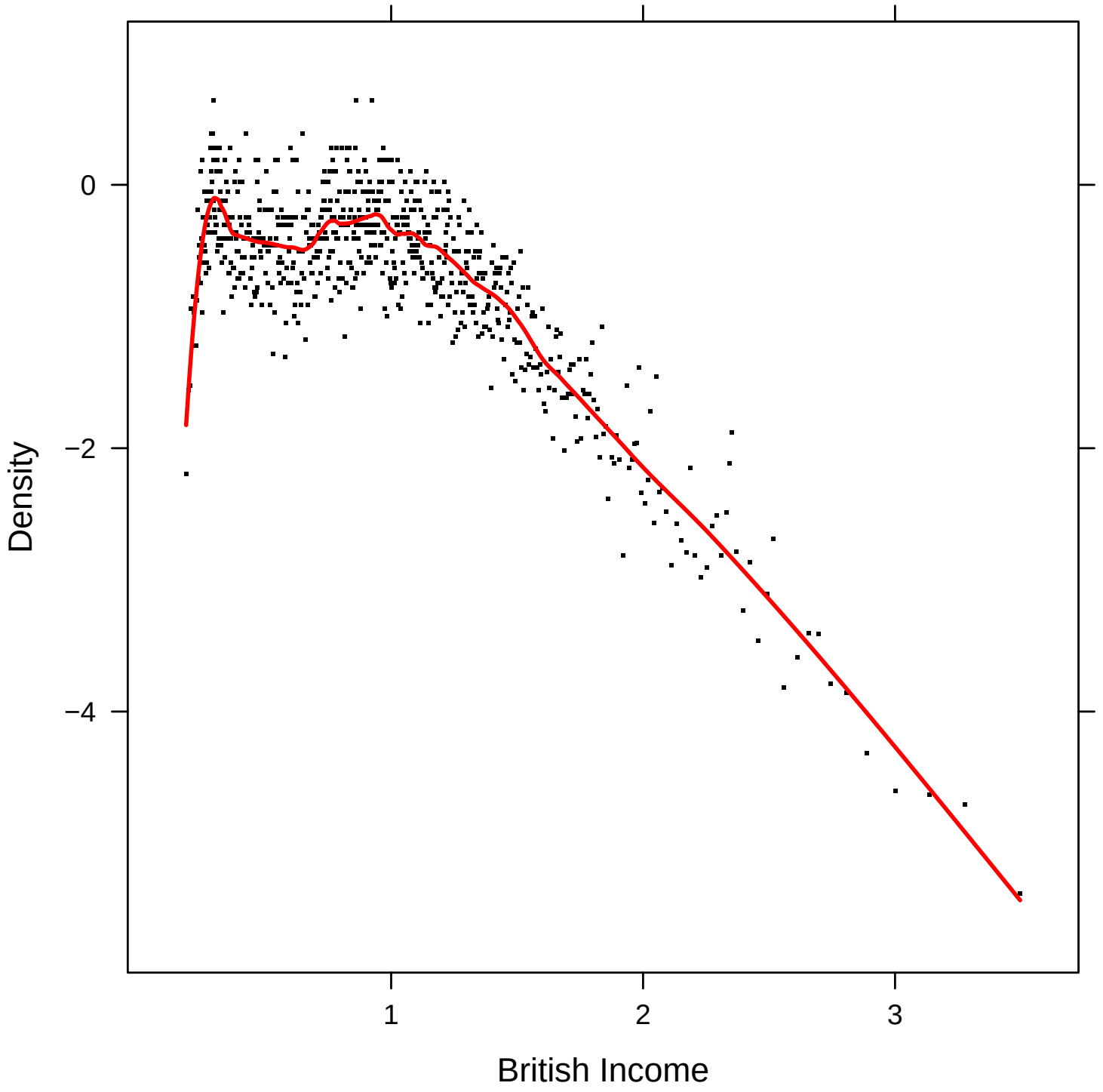
(2) large enough that $y_i^{(\kappa)}$ is approximately normal

- $\kappa = 10$ 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

ed: Log Balloon Densities vs. Gap Midpoints for Income with $\kappa = 10$



ed Step 2: Smooth Log Balloon Densities Using Nonparametric Regression



ed Step 2

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

Fit polynomials locally of degree δ in a moving fashion like a moving average of a time series

Bandwidth parameter $0 < \alpha \leq 1$

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

Weighted least-squares fitting where weights decrease to 0 as distances of neighborhood points increase to the neighborhood boundary

Loess possesses all of the statistical-inference technology of parametric fitting for linear models

ed: Three Tuning Parameters

κ: gap length

α: bandwidth

δ: degree of polynomial in local fitting

Notational Change

Midpoints of gap intervals: $x_i^{(\kappa)} \longrightarrow x_i$

Log balloon densities: $y_i^{(\kappa)} \longrightarrow y_i$

ed Step 2

A Model Building Approach

A starter model for the log balloon densities y_i as a function of gap midpoints x_i

- based on "theory"
- hope for a good approximation to the underlying patterns in the data

$$y_i = y(x_i) + \epsilon_i$$

$y(x) = \log(f(x))$ = log density

- well approximated by polynomials of degree δ locally in neighbors of x determined by α
- expect δ to be 2 or 3 to reach up to the tops of peaks and the bottoms of valleys

ϵ_i

- independent
- identically distributed with mean 0
- distribution well approximated by the normal

Use the comprehensive set of tools of regression diagnostics to investigate assumptions

Model Diagnostics

Important quantities used in carrying out diagnostics

$\hat{y}(x)$ = ed log density fit from nonparametric regression

$\hat{y}_i = \hat{y}(x_i)$ = fitted values at x_i = gap midpoints

$\hat{\epsilon}_i = y_i - \hat{y}_i$ = residuals

$\hat{f}(x) = \exp(\hat{y}(x))$ = ed density estimate

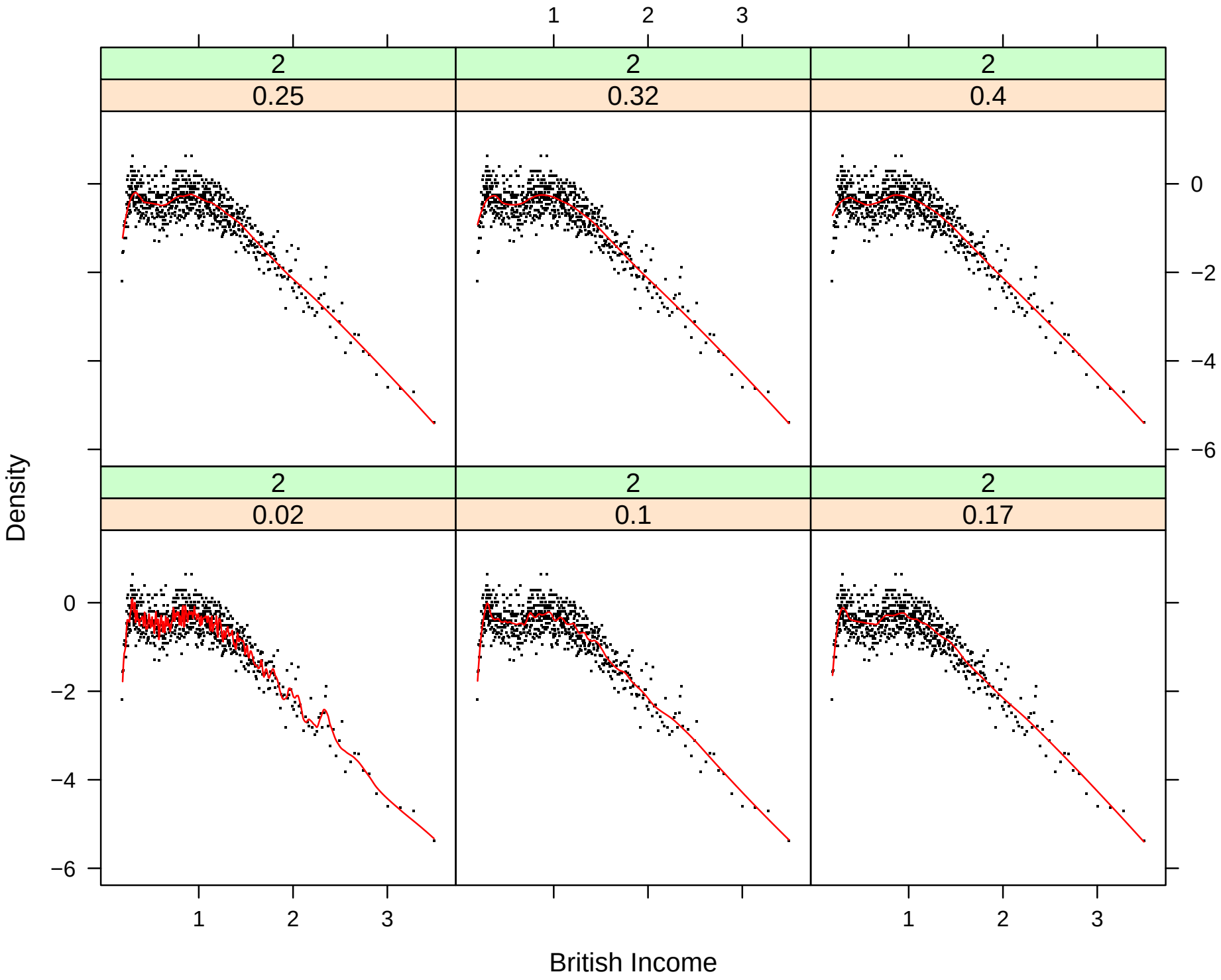
Loess possesses all of the statistical-inference technology of parametric linear regression

$\hat{\sigma}^2_\epsilon$ = estimate of error variance

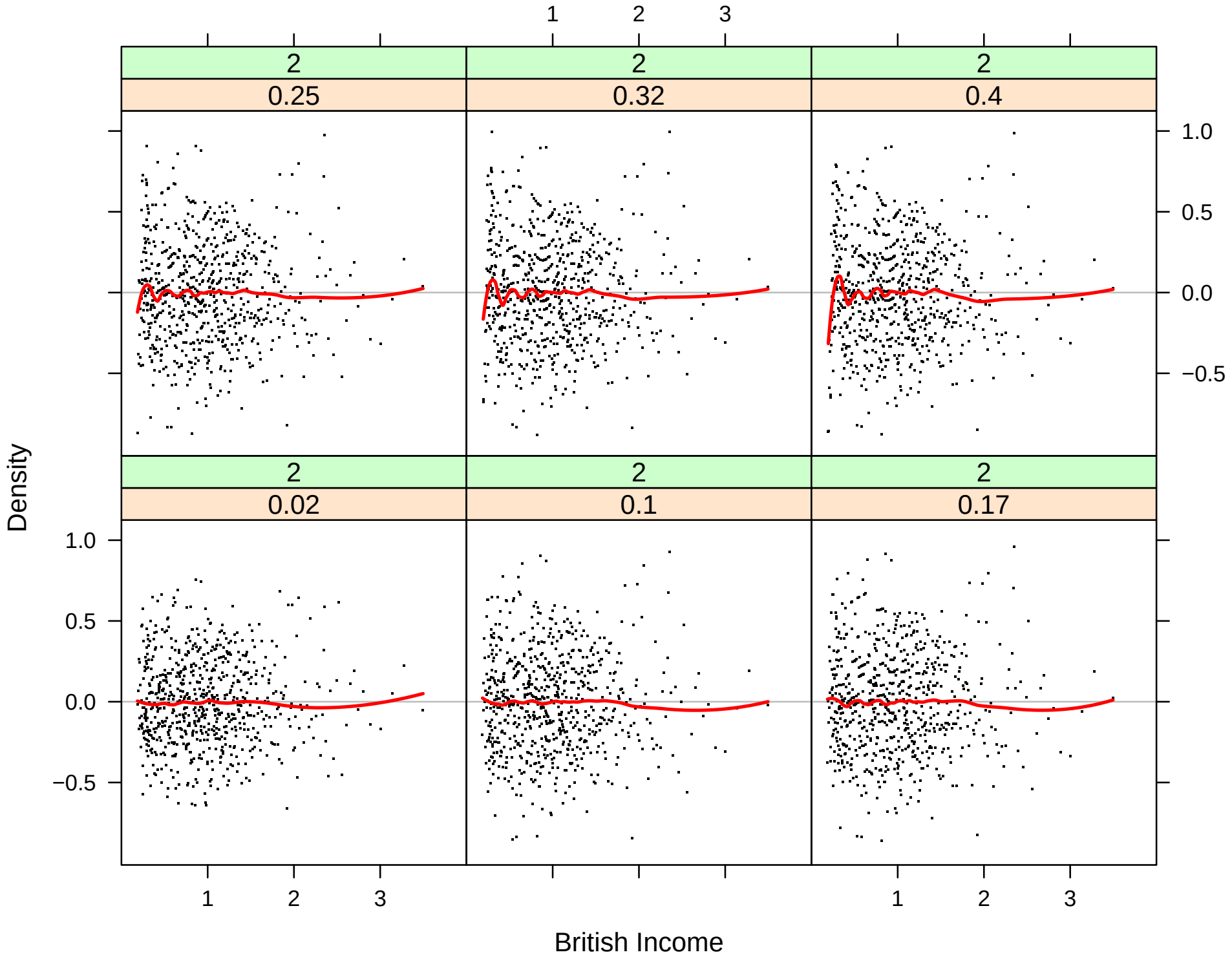
- in theory, variance = $\gamma_1(\kappa)$, which is 0.105 for $\kappa = 10$

ν = equivalent degrees of freedom of the fit (number of parameters)

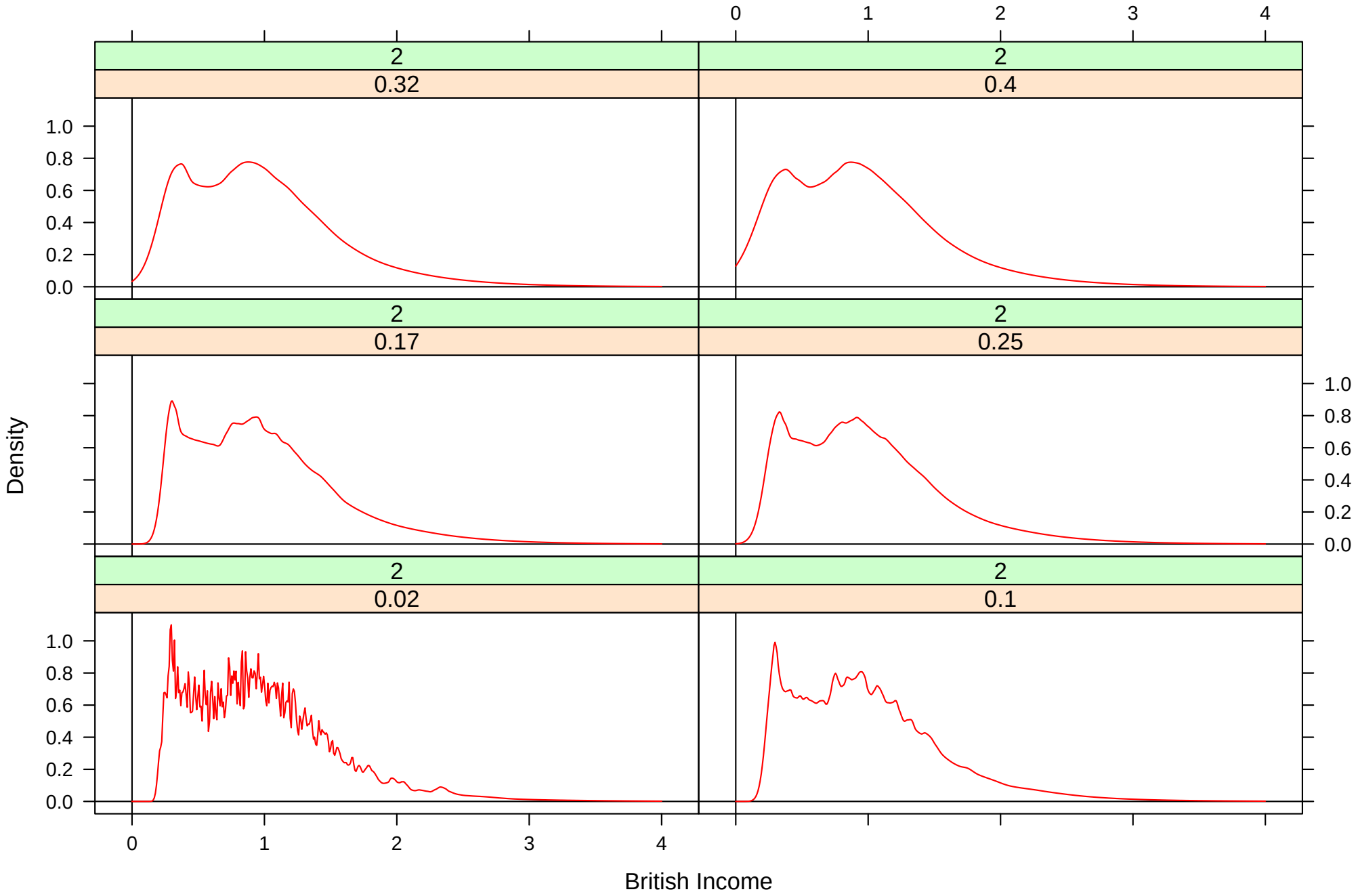
ed Log Raw Density Estimate for Income: Fit vs. Income



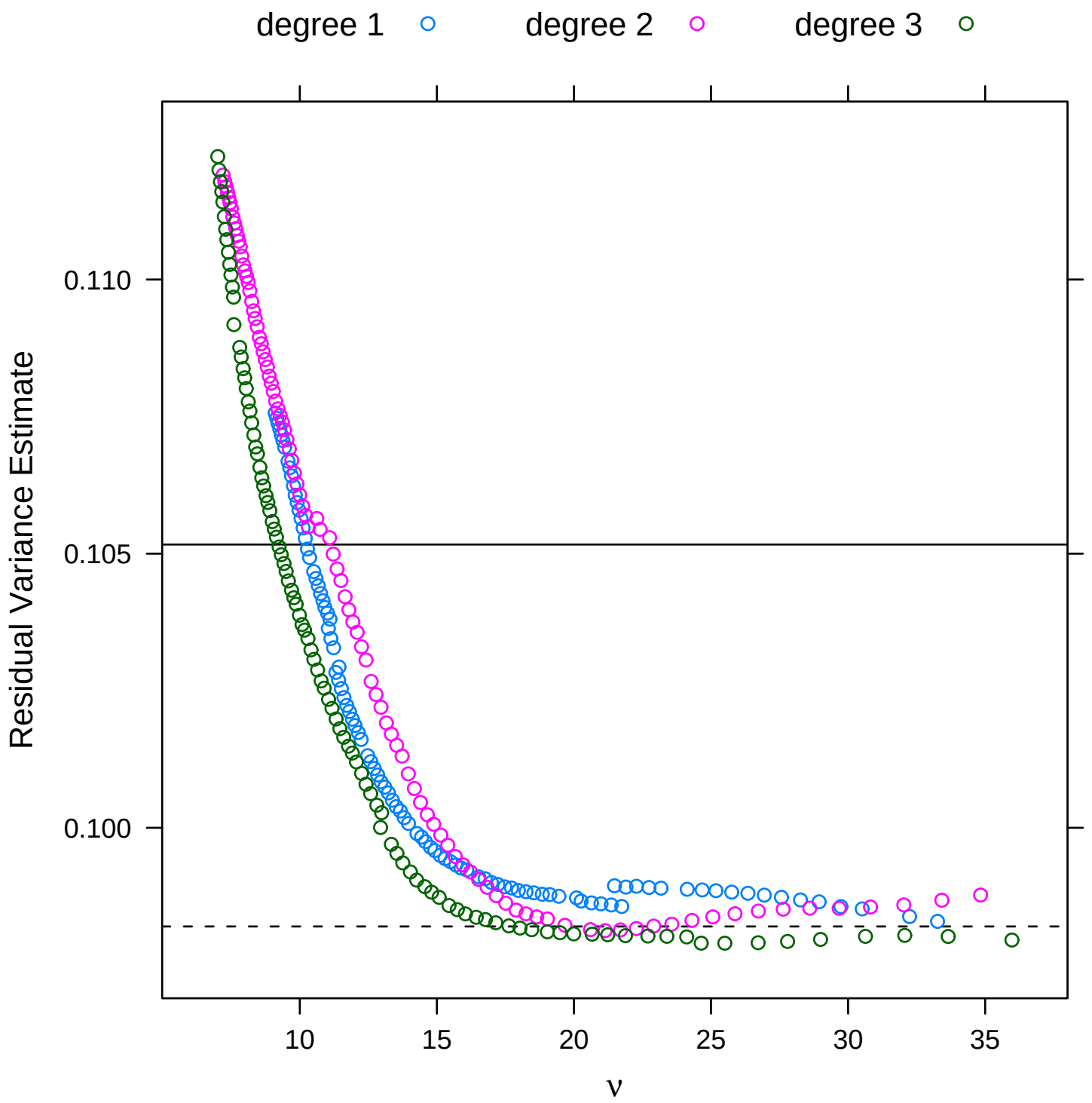
ed Log Density Estimate for Income: Residuals vs. Income



ed Density Estimate for Income: Fit vs. Income



ed Log Density Estimate for Income: Residual Variance vs. Degrees Freedom



Mallows Cp Model Selection Criterion

A visualization tool for showing the trade-off of bias and variance, which is far more useful than just a criterion that one minimizes or is optimizing

M: an estimate of mean-squared error

ν: degrees of freedom of fit

M estimates

$$\frac{\text{Bias Sum of Squares}}{\sigma^2} + \frac{\text{Variance Sum of Squares}}{\sigma^2}$$

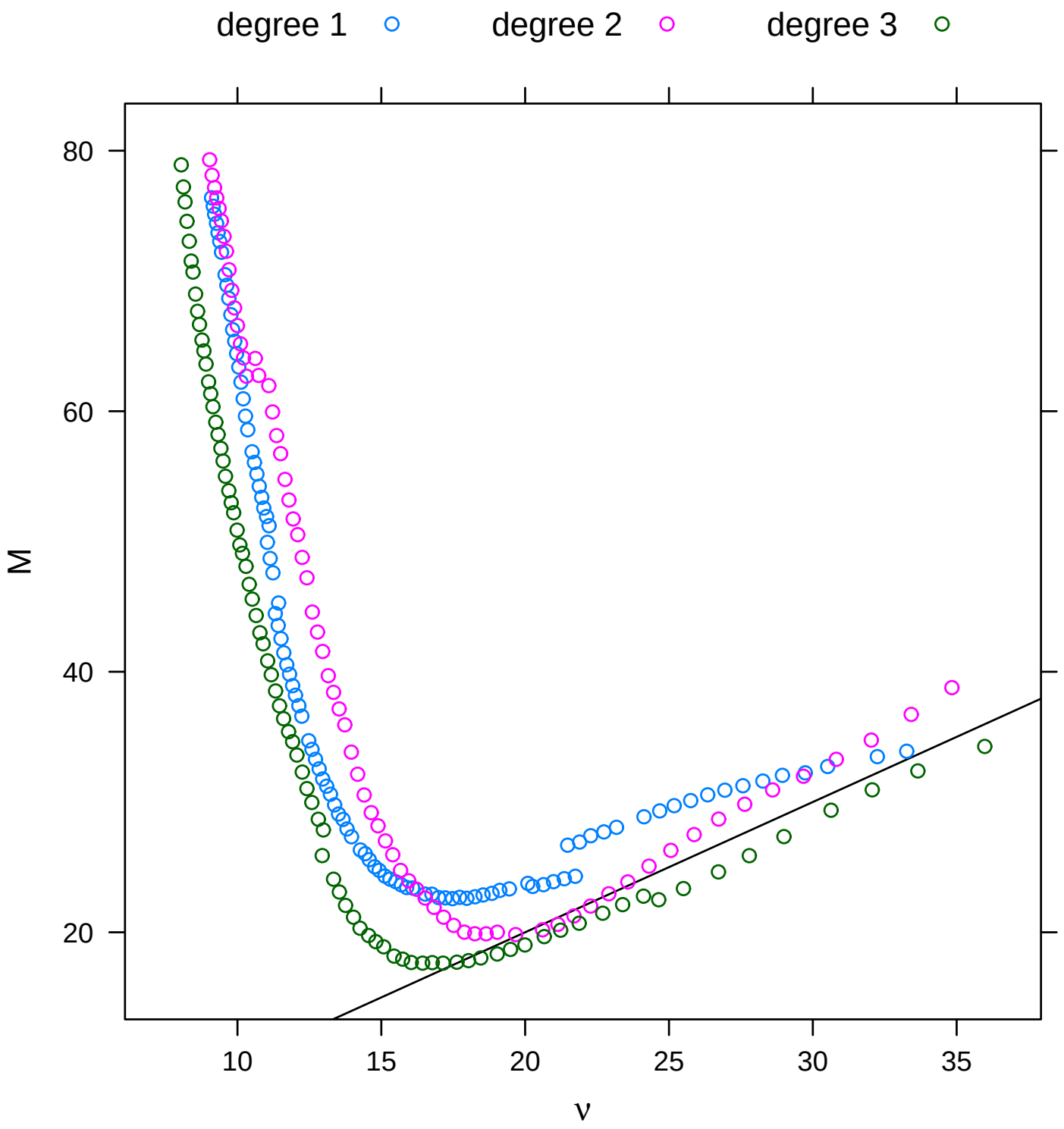
$$M = \frac{\sum_i \hat{\epsilon}_i^2}{\hat{\sigma}^2(\epsilon)} - (n - \nu) + \nu$$

$\mathrm{E}(M) \approx \nu$ when the fit follows the pattern in the data

The amount by which M exceeds ν is an estimate of the bias

Cp: M vs. ν

ed Log Density Estimate for Income: Cp Plot



Model Selection for British Incomes

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

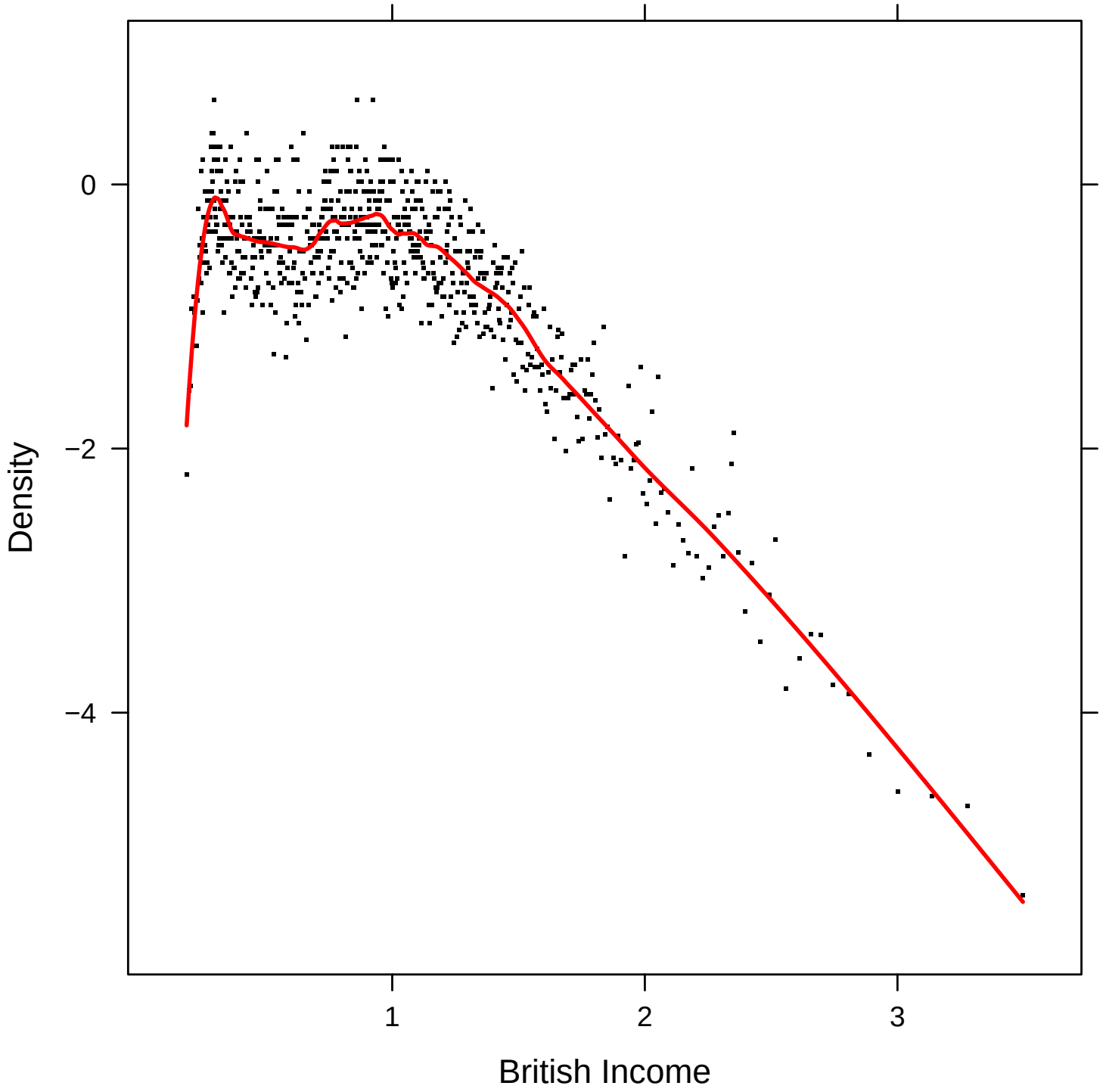
Gap size: $\kappa = 10$

Polynomial degree: $\delta = 2$

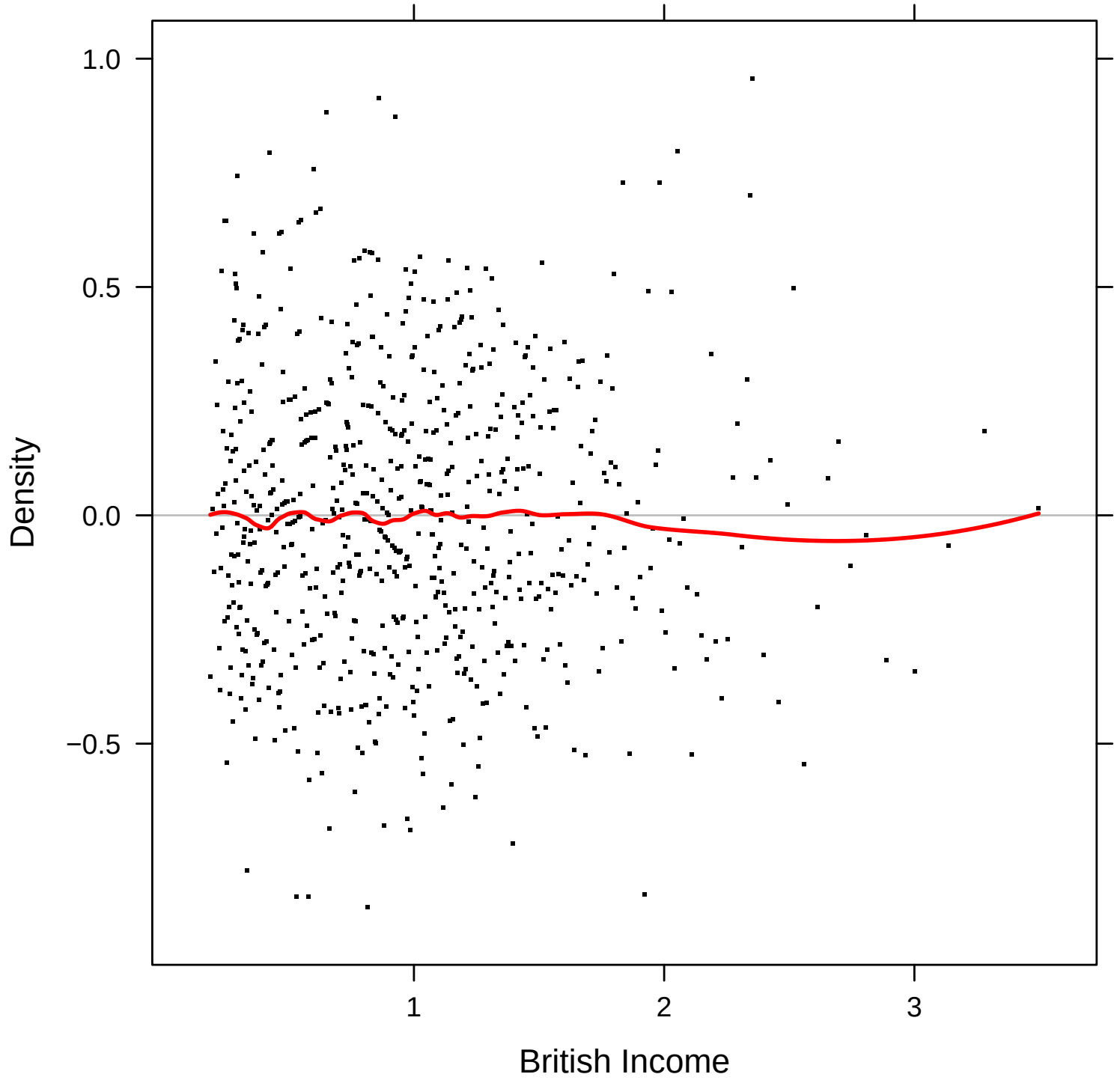
Bandwidth parameter: $\alpha = 0.16$

Equivalent degrees of freedom: $\nu = 19$

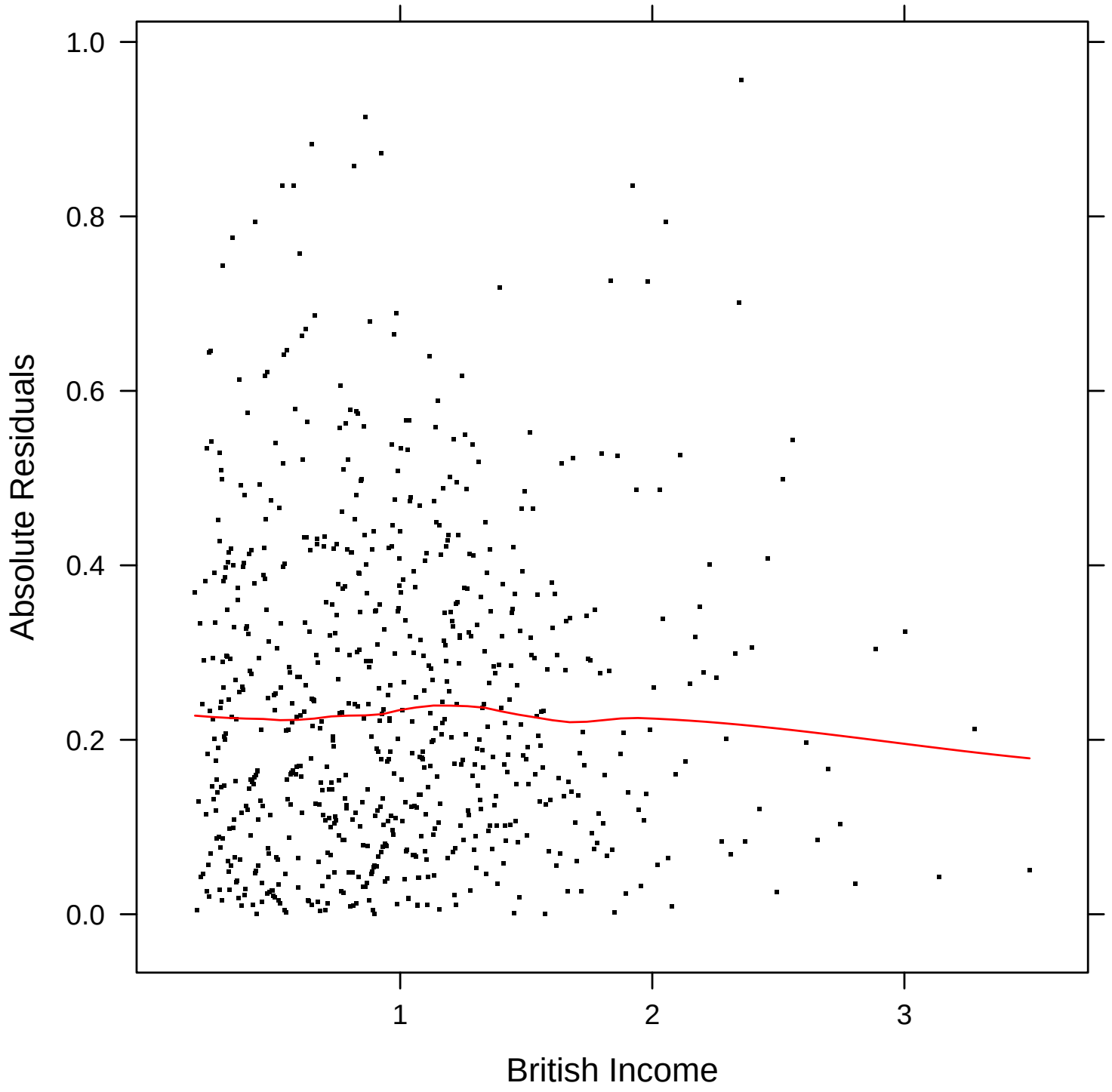
ed Log Density Estimate for Income: Fit vs. Income



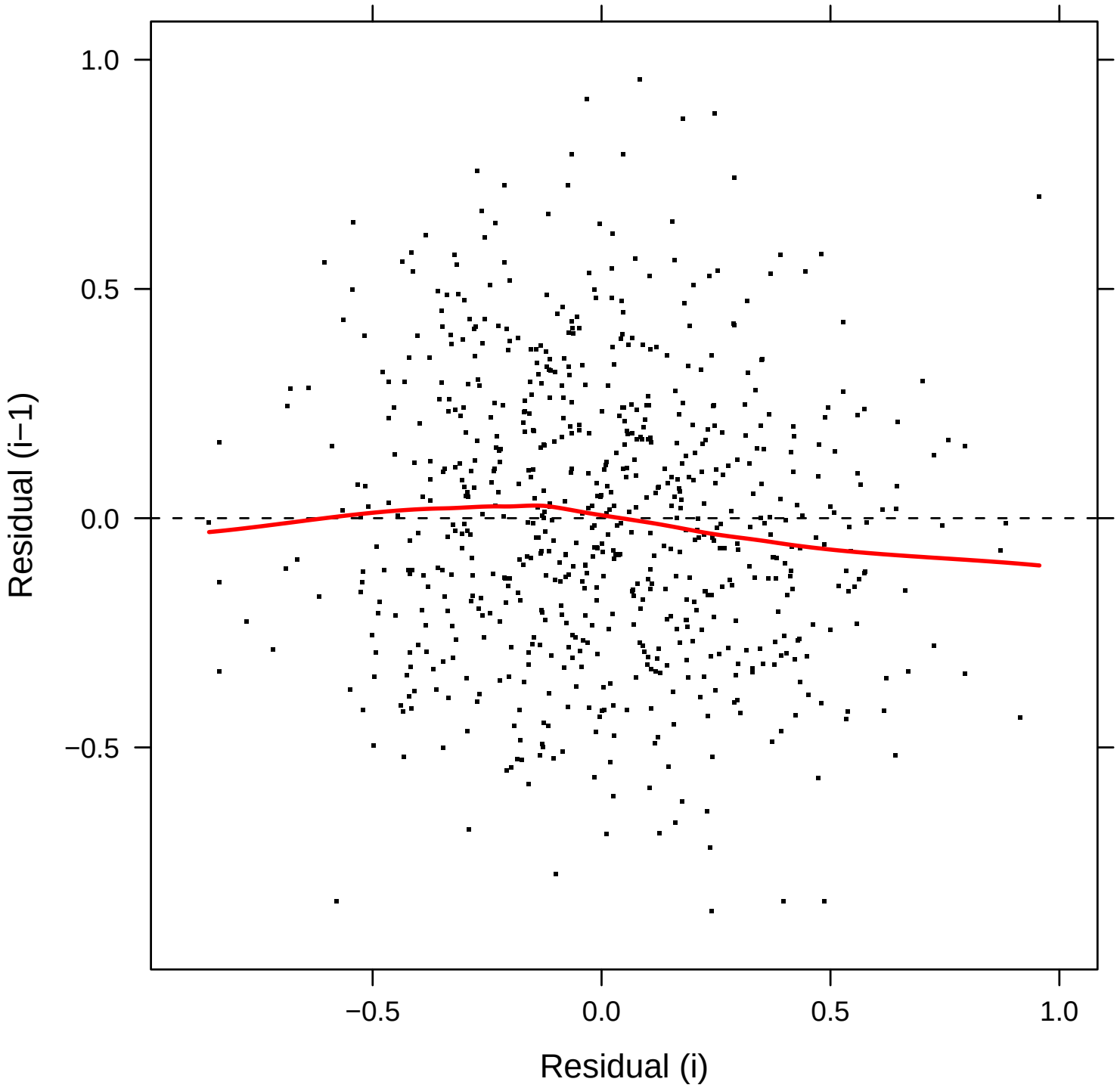
ed Log Density Estimate for Income: Residuals vs. Income



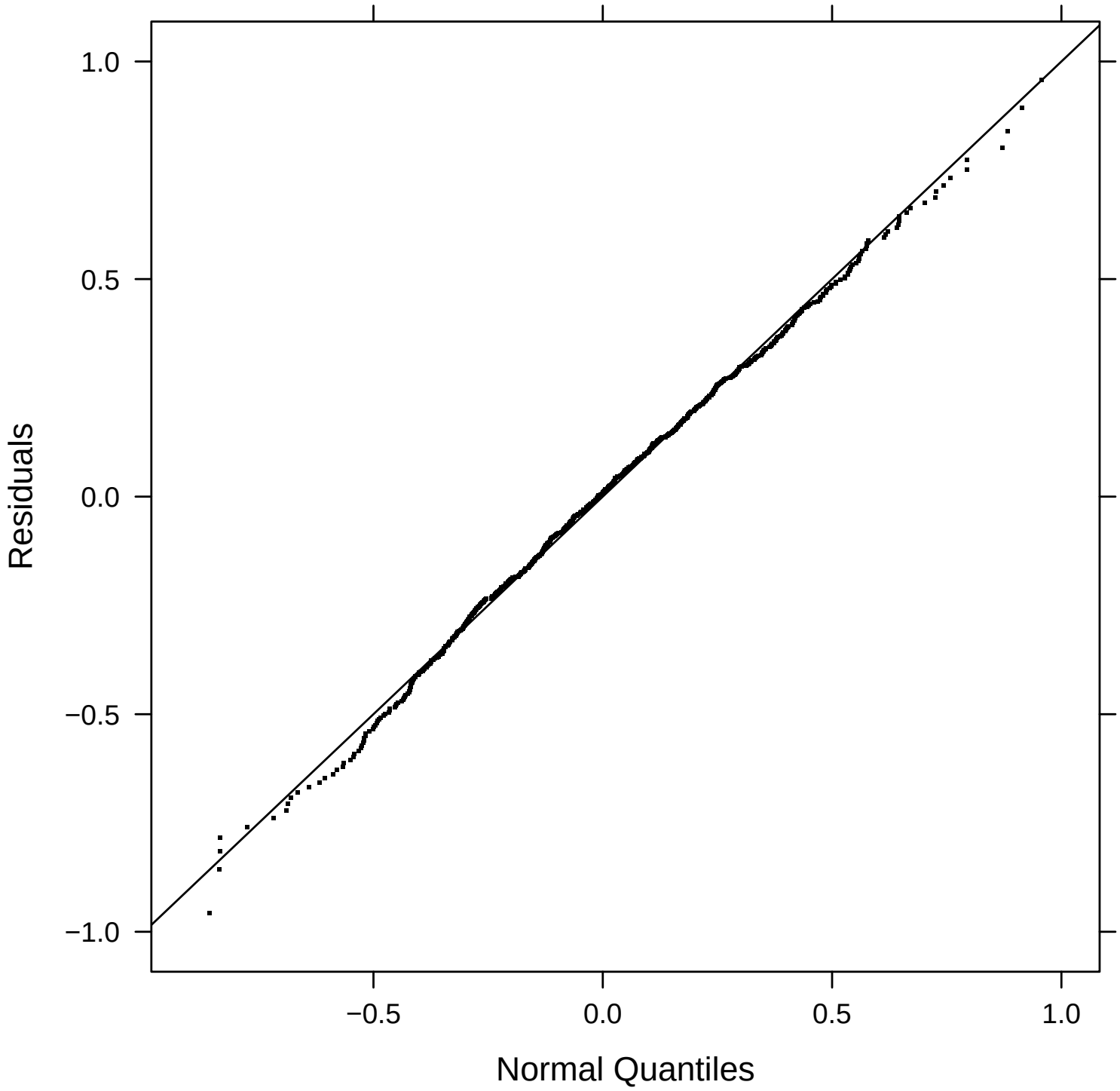
ed Log Density Estimate for Income: Absolute Residuals vs. Income



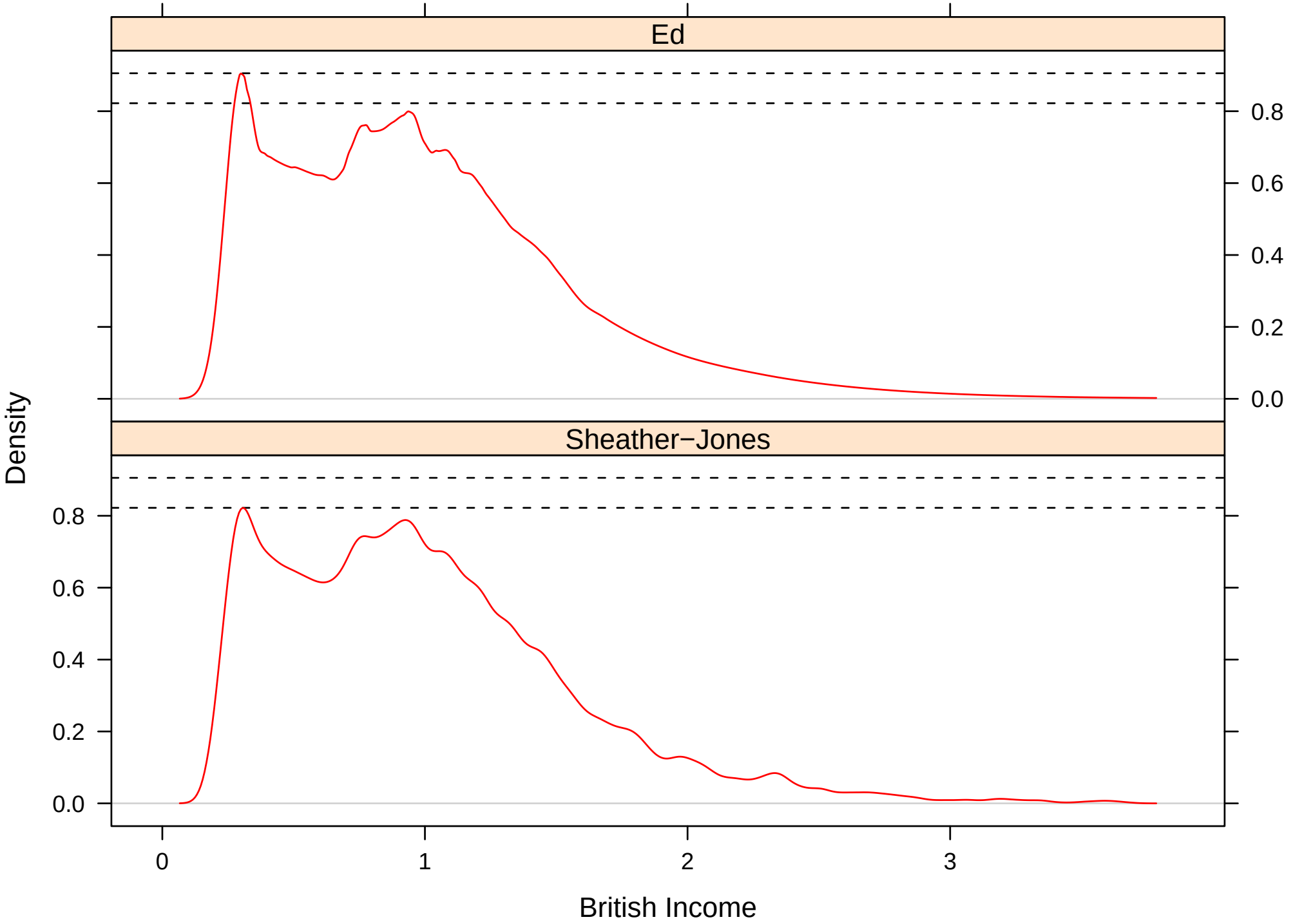
ed Log Density Estimate for Income: Lag 1 Residual Correlation Plot



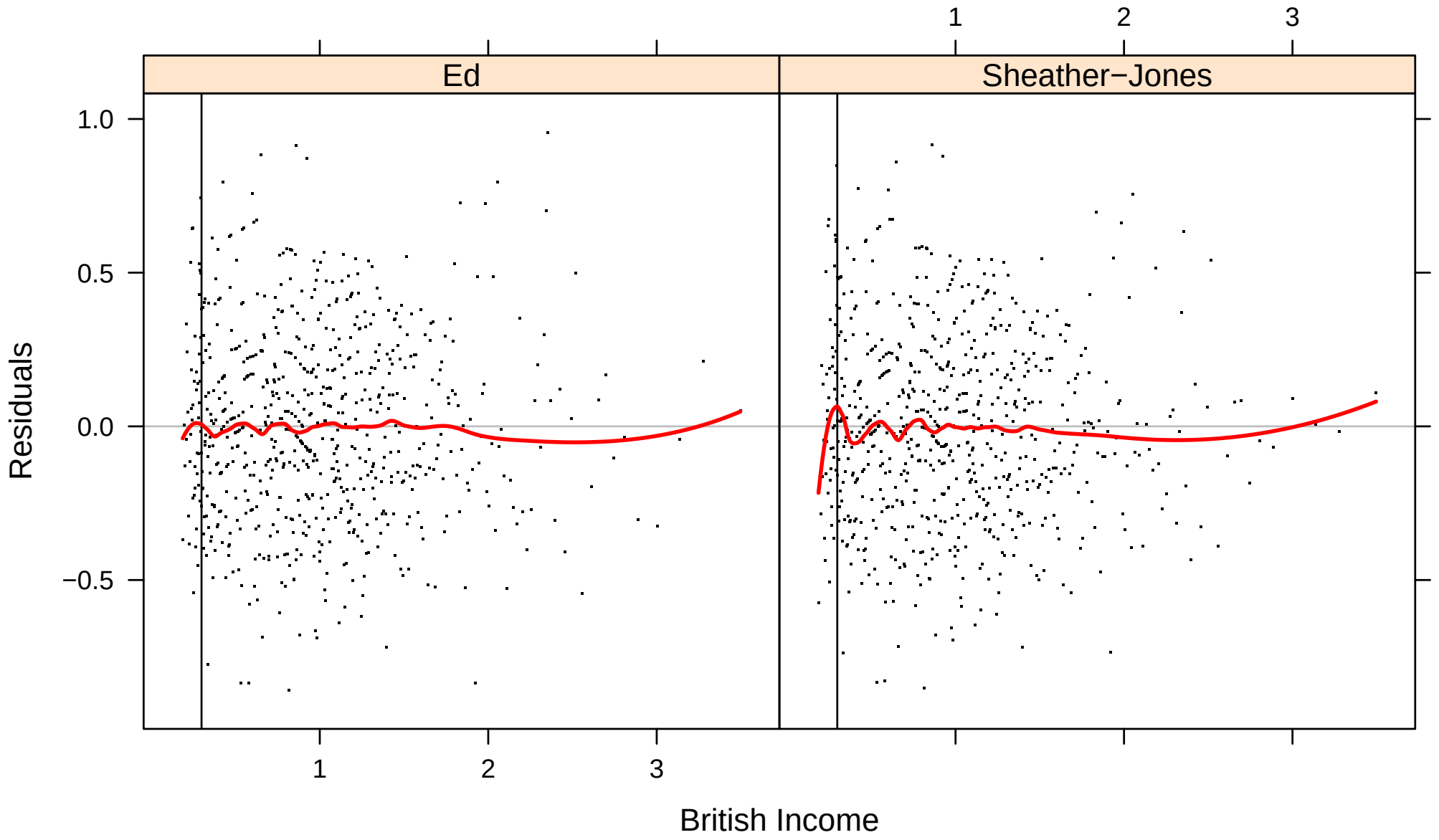
ed Log Density Estimate for Income: Normal Quantile Plot of Residuals



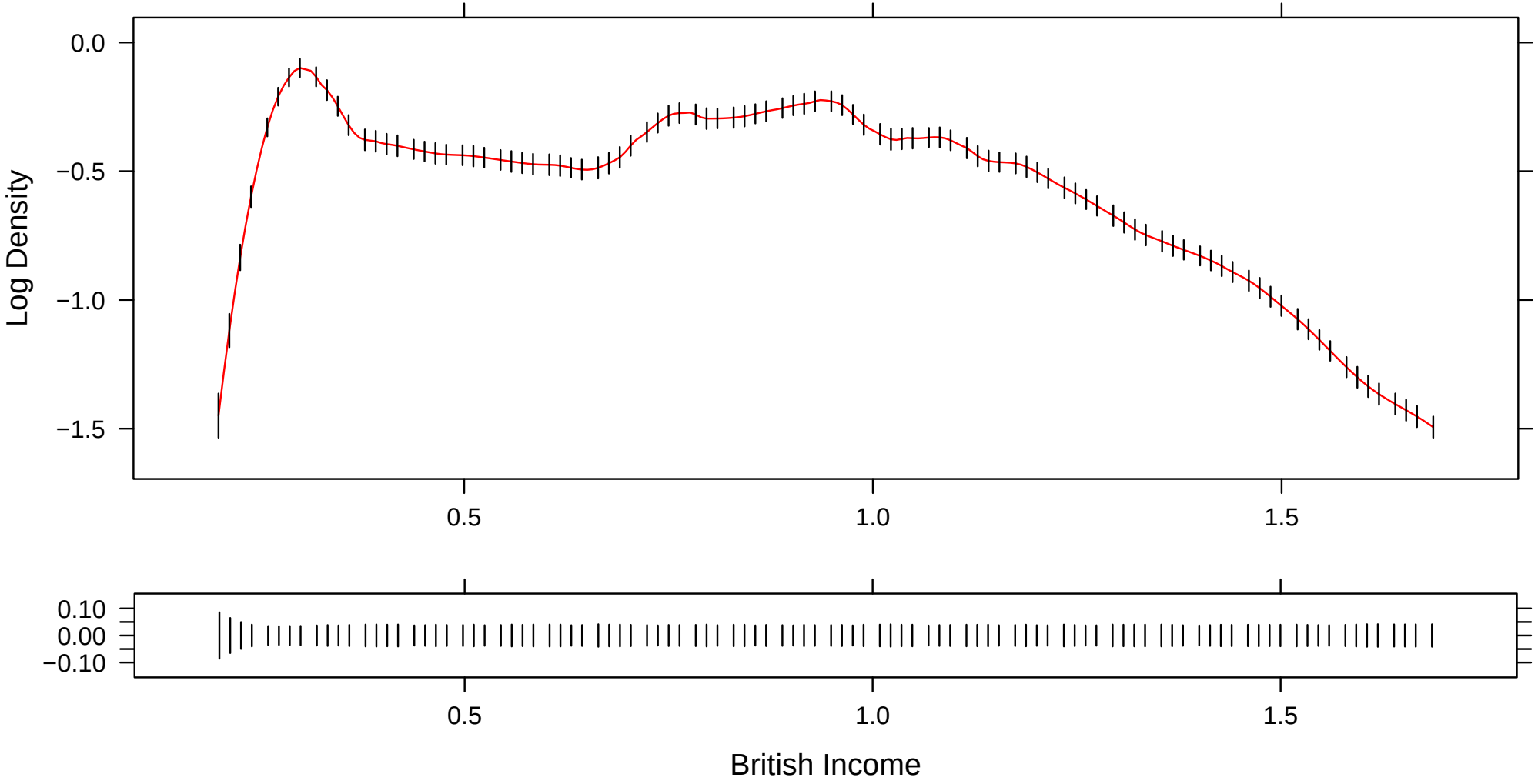
ed Density Estimate and KDE for Income: Fits vs. Income



ed Log Density Estimate and Log KDE for Income: Residuals vs. Income



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



Queuing Delays of Internet Voice-over-IP Traffic

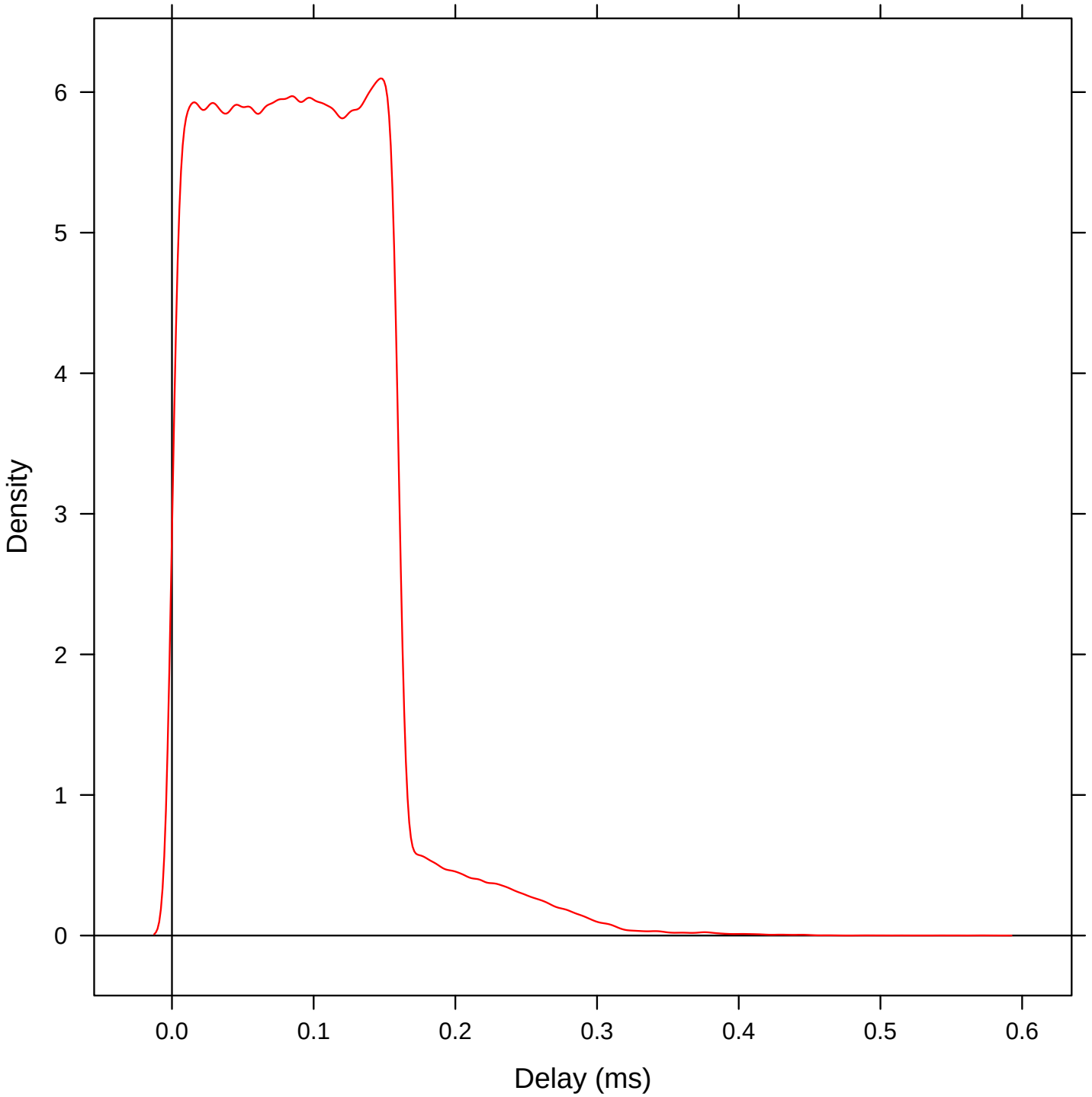
264583 observations

From a study of quality of service for different traffic loads

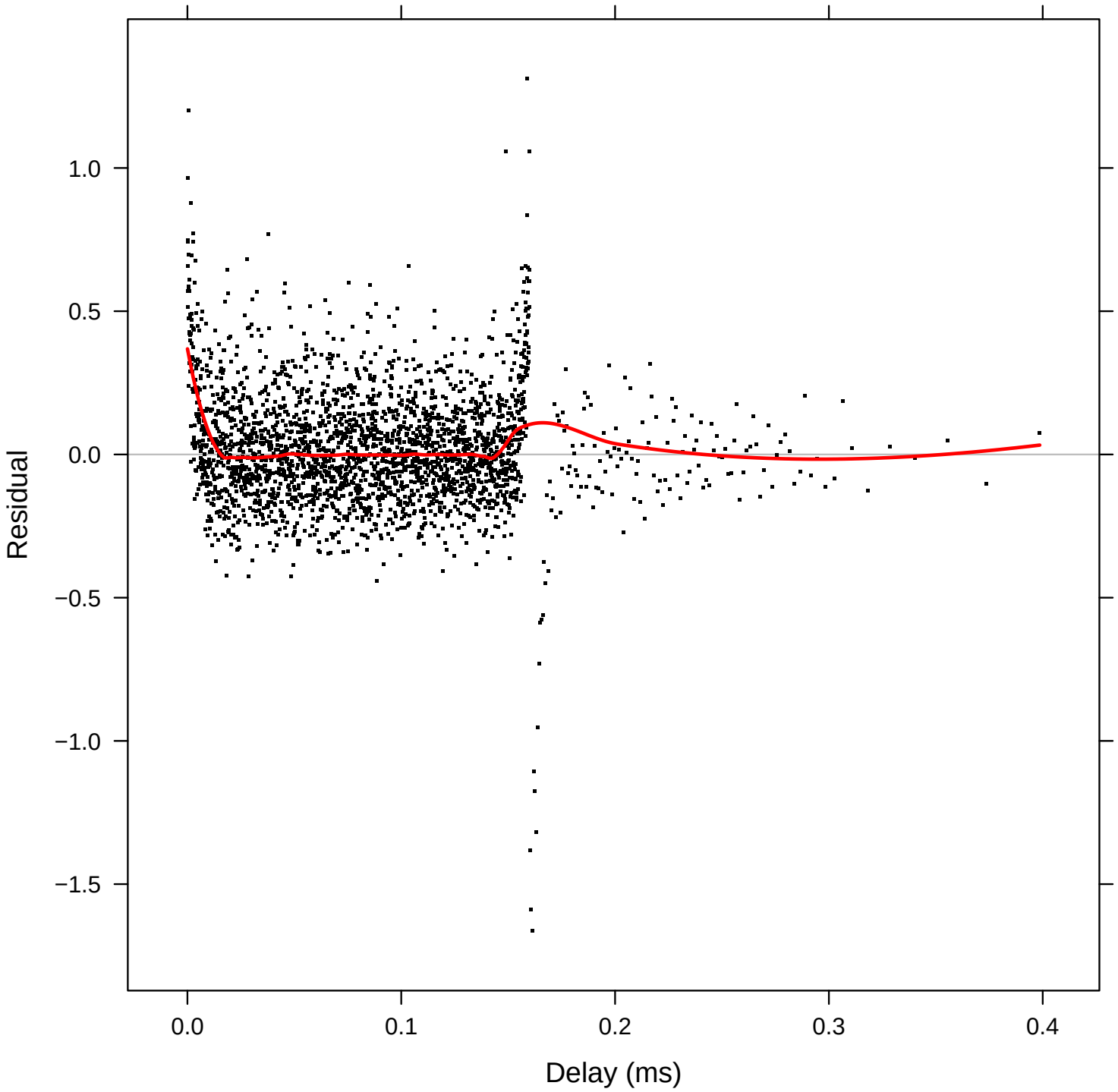
Semi-empirical model for generating traffic

Simulated queueing on a router

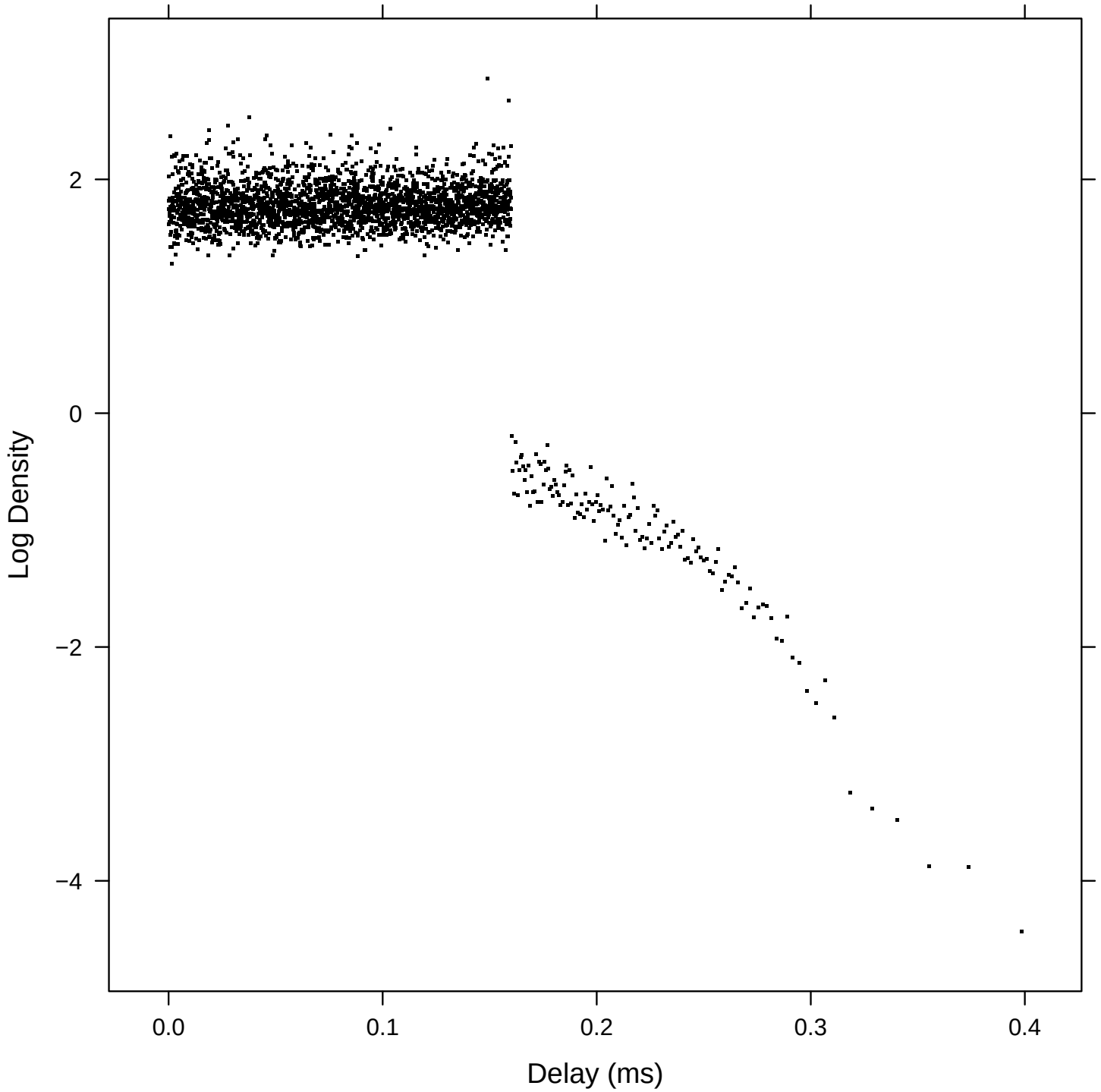
Silverman Bandwidth KDE for Delay: Fit vs. Delay



Silverman Bandwidth Log KDE for Delay: Residuals vs. Delay



ed Log Raw Density Estimates for Delay: $\kappa = 100$ Because of Ties



ed Log Density Estimates for Delay: 3 Independent Fits for Three Intervals

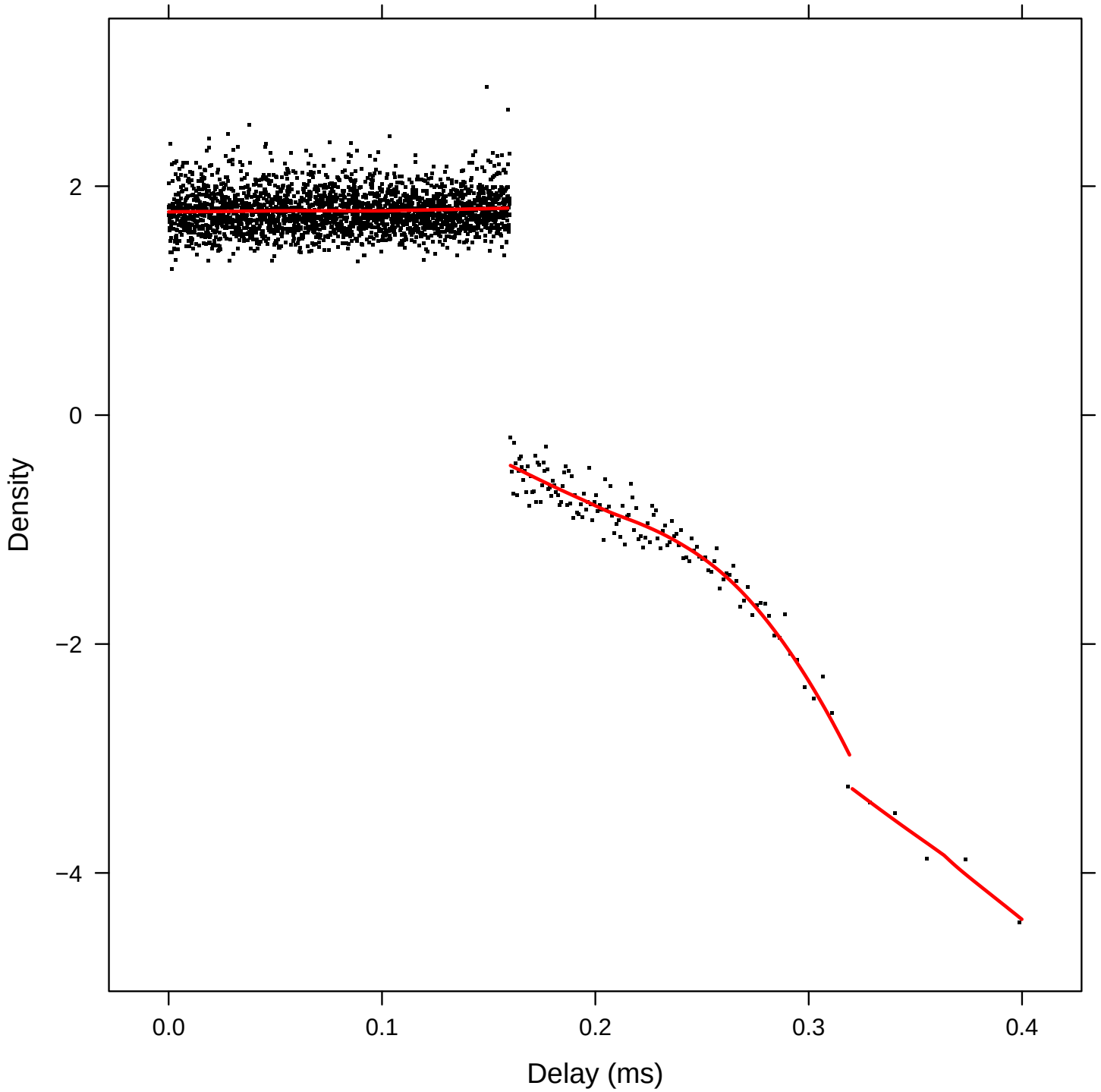
Polynomial degrees δ and bandwidths α for loess fits

Interval 1: $\delta = 1, \alpha = 0.75$

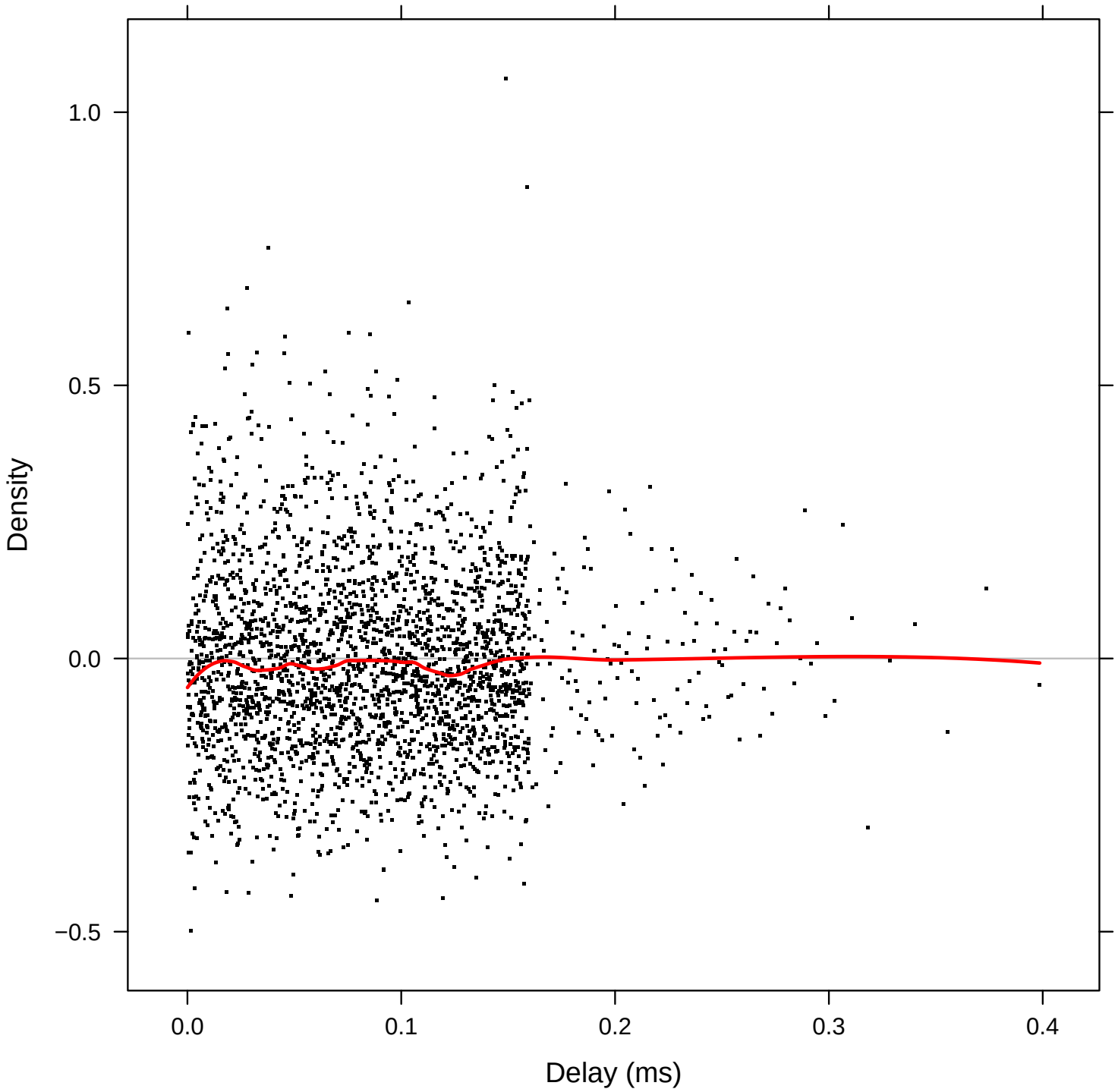
Interval 2: $\delta = 2, \alpha = 0.5$

Interval 3: $\delta = 1, \alpha = 1.5$

ed Log Density Estimates for Delay: 3 Fits vs. Delay



ed Log Density Estimates, 3 Fits for Delay: Residuals vs. Delay



Silverman Bandwidth KDE for Delay: 3 Fits

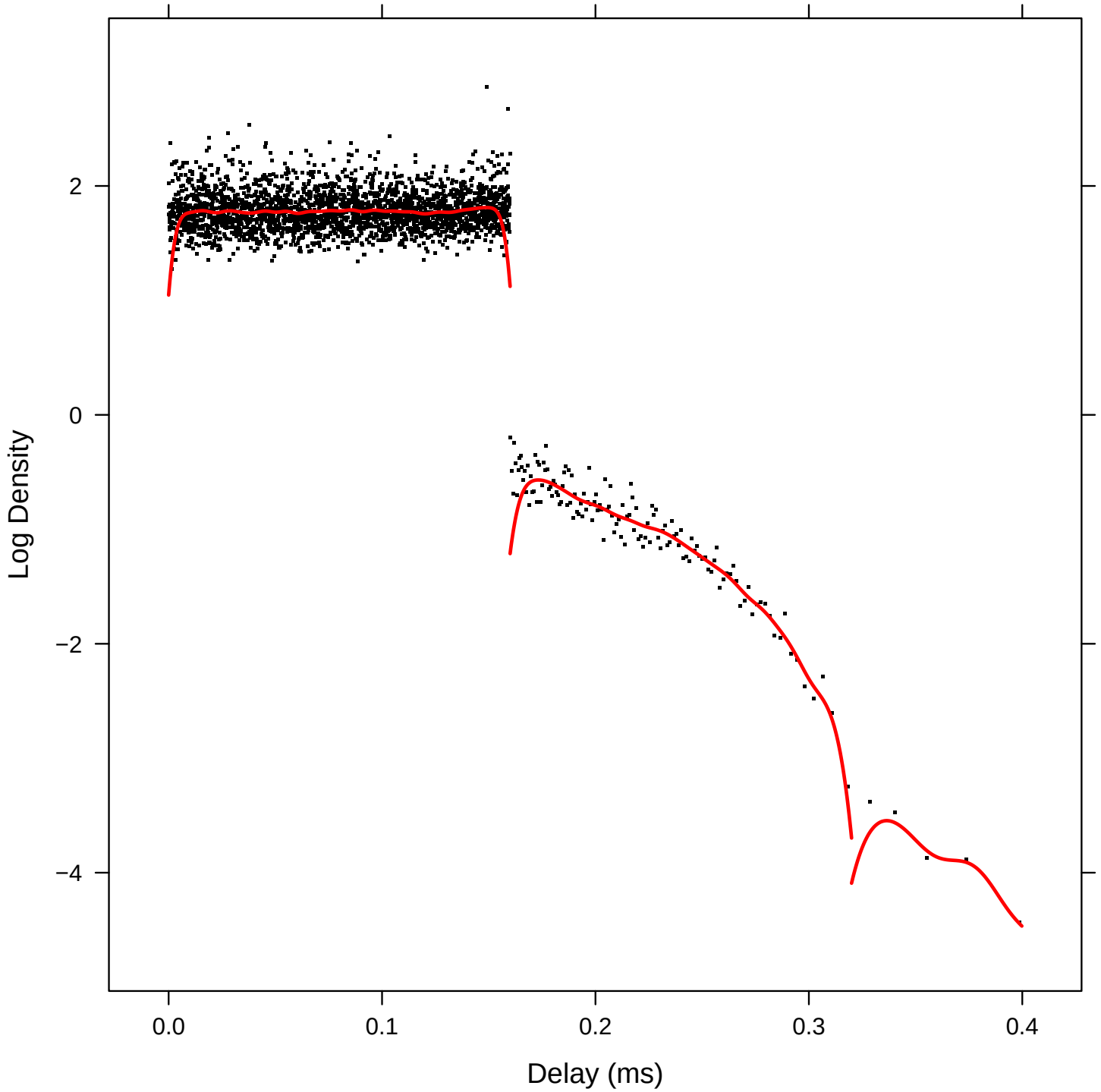
Unit normal kernel so h = standard deviation

Interval 1: h = 0.00346

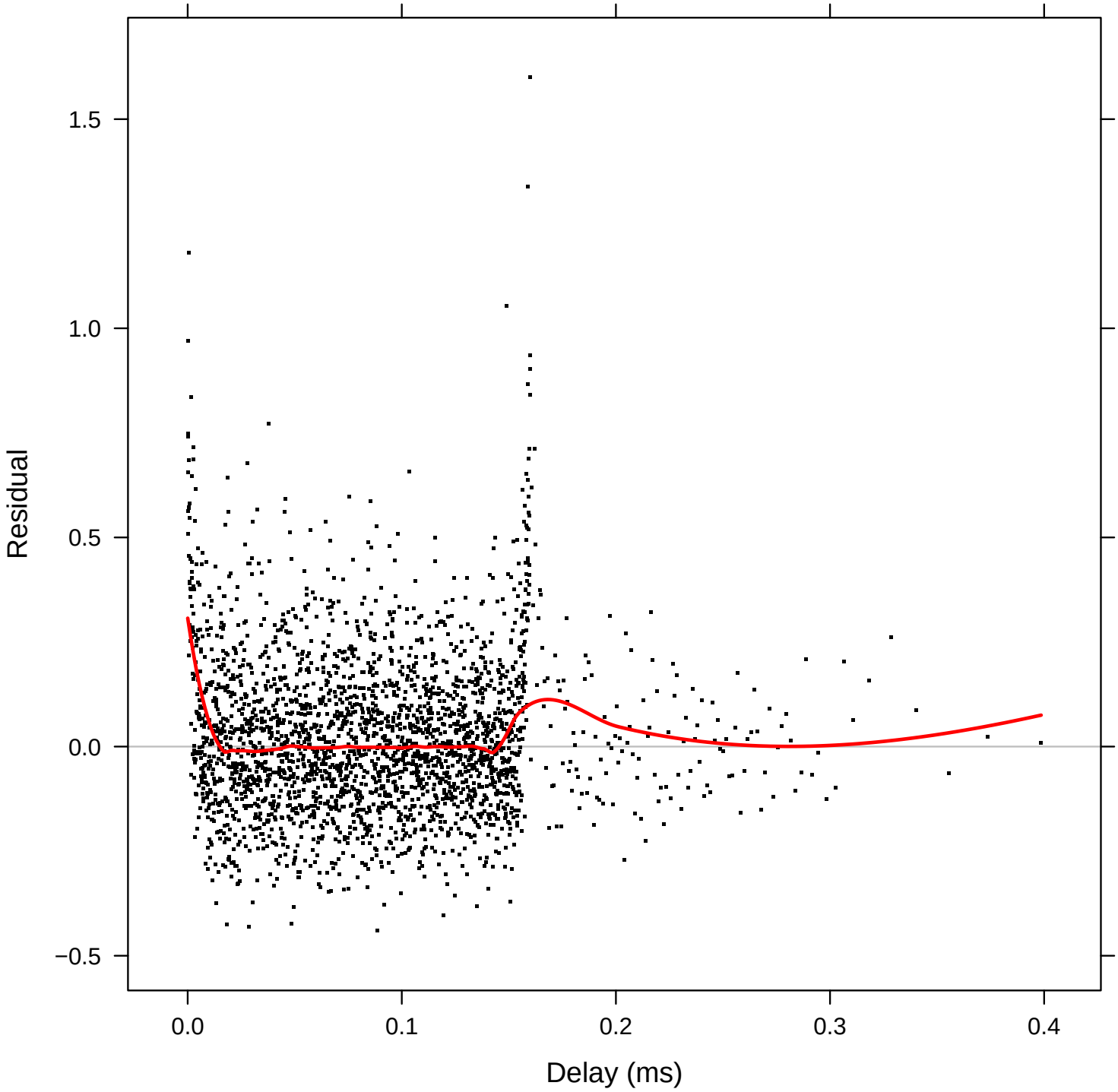
Interval 2: h = 0.00532

Interval 3: h = 0.00971

Silverman Bandwidth KDE for Delay: 3 Fits vs. Delay



Silverman Bandwidth Log KDE, 3 Fits for Delay: Residuals vs. Delay



Computation

We use existing algorithms for loess to get faster computations for ed than direct computation

Still, are the ed computations fast enough, especially for extensions to higher dimensions?

Alex Gray and collaborators have done some exceptional work in algorithms for fast computation of KDEs and nonparametric regression

Nonparametric Density Estimation: Toward Computational Tractability. Gray, A. G. and Moore, A. W. In SIAM International Conference on Data Mining, 2003. Winner of Best Algorithm Paper Prize.

How can we tailor this to our needs here?

RHIPE

ed presents embarrassingly parallel computation

Large amounts of computer time can be saved by distributed computing environments

One is RHIPE (ml.stat.purdue.edu/rhipe)

- Saptarshi Guha, Purdue Statistics
- R-Hadoop Integrated Processing Environment
- Greek for “in a moment”
- pronounced “hree pay”

A recent merging of the R interactive environment for data analysis (www.R-project.org) and the Hadoop distributed file system and compute engine (hadoop.apache.org)

Public domain

A remarkable achievement that has had a dramatic effect on our ability to compute with large data sets