STAT444/844, CM764
Statistical Learning: Function Estimation

Shoja Chenouri
schenouri@uwaterloo.ca
Dept. of Statistics and Actuarial Science
University of Waterloo,

January 6, 2014

Introduction

Overview

- This course is about statistical function estimation, such as estimation of regression functions, probability density functions, and intensity functions.
- We will consider both parametric and nonparametric smoothing methods.
- In parametric case, our main focus will be on variable selection in high-dimensional multiple linear regression including lasso, adaptive lasso, elastic net, group lasso LARS, Dantzig selector, etc.
- In nonparametric case, we will cover many techniques including local polynomials, kernel methods, smoothing splines, wavelets, local likelihood and exponential families, tree based methods, etc.
- Models assessments: cross validation, tuning parameters, AIC, BIC.

Learning Materials

- These slides can be downloaded from Desire2Learn (D2L) site (`https://learn.uwaterloo.ca`)
- Notes from class.
- Instructor's office hours: (M3 3124) 15:30 - 17:30 Mondays or by appointment.
- Teaching Assistant: Garcia (Jiaxi) Liang. His office hours are on Wednesdays 14:00 - 16:00 in M3 3108.

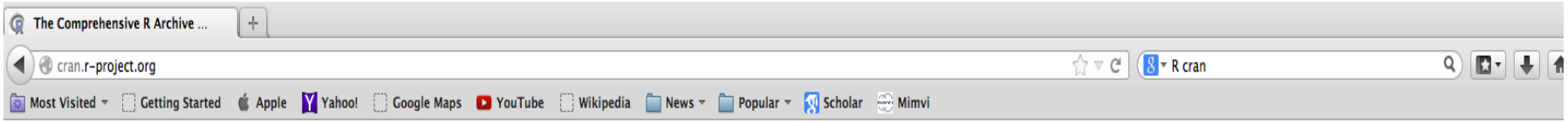
Recommended text books

Here are highly recommended texts on the materials covered in the lectures. For more see the course outline.

- Hastie, T., Tibshirani, R. and Friedman, J. (2009) "*The Elements of Statistical Learning*", 2nd edition, Springer, New York.
- James, G., Witten, D., Hastie, T., Tibshirani, R., (2013). "*An Introduction to Statistical Learning with Applications in* `R`", Springer, New York.
- Wasserman, L. (2006) "*All of Nonparametric Statistics*", Springer, New York.
- Clarke, B., Fokoue, E. and Zhang, H. H. (2009). "*Principles and Theory for Data Mining and Machine Learning*, Springer, New York.
- Sheather, S. J. (2009) "*A Modern Approach to Regression with R*". Springer, New York.

Software for Computation

- One of the objectives of this course is to make you fluent in the computation associated with statistical learning and function estimation.
- `R` will be the language used for computation in this course.
- Assignments will need knowledge of coding in `R`.
- In Exams, you are expected to interpret `R` outputs.



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2013-09-25, Frisbee Sailing) R-3.0.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Submitting to CRAN

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Contributed Documentation

English --- Other Languages

Manuals, tutorials, etc. provided by users of R. The R core team does not take any responsibility for contents, but we appreciate the effort very much and encourage everybody to contribute to this list! To submit, follow the submission instructions on the CRAN main page. All material below is available directly from CRAN, you may also want to look at the list of other R documentation available on the Internet.

Note: Please use the directory listing to sort by name, size or date (e.g., to see which documents have been updated latetly).

English Documents

Documents with more than 100 pages:

- **"Using R for Data Analysis and Graphics - Introduction, Examples and Commentary"** by John Maindonald (PDF, data sets and scripts are available at JM's homepage).
- **"Practical Regression and Anova using R"** by Julian Faraway (PDF, data sets and scripts are available at the book homepage).
- The **Web Appendix** to the book "An R and S-PLUS Companion to Applied Regression" by John Fox contains information about using S (R and S-PLUS) to fit a variety of regression models.
- **"An Introduction to S and the Hmisc and Design Libraries"** by Carlos Alzola and Frank E. Harrell, especially of interest to SAS users, users of the Hmisc or Design packages, or R users interested in data manipulation, recoding, etc. (PDF)
- **"Statistical Computing and Graphics Course Notes"** by Frank E. Harrell, includes material on S, LaTeX, reproducible research, making good graphs, brief overview of computer languaes, etc. (PDF).
- **"An Introduction to R: Software for Statistical Modelling & Computing"** by Petra Kuhnert and Bill Venables (ZIP 3.8MB): A 360 page PDF document of

Course Assessment

- Your final mark will be calculated according to the following weighting:
  - Assignments: 30%,
  - Midterm: 15%,
  - Group project: 15%,
  - Final exam: 40%

- In order to pass the course, you must obtain at least 50% on the final exam as well as an overall final grade of at least 50% according to the above grading scheme.

Introduction

- Data analysis with relatively little prior information.
  - Let $\mathcal{D}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ be a random sample, from an unknown distribution F.

- Classification: For $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathbb{R}^p \times \{0, 1\}$, $i = 1, \ldots n$
  Objective is to estimate $p(\mathbf{x}) = P_{_F}(Y = 1 \,|\, \mathbf{X} = \mathbf{x})$.

- Regression: For $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots n$
  Objective is to estimate $e(\mathbf{x}) = E_{_F}(Y \,|\, \mathbf{X} = \mathbf{x})$.

- Density estimation: For $\mathbf{Z}_i \in \mathbb{R}^p$, , $i = 1, \ldots n$
  objective is to estimate the density function $f_{\mathbf{Z}}(\mathbf{z})$ of $\mathbf{Z}$.

Classification

There are many situations in which classification is required

- **Data:** $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.
- **Explanatory variables:** $\mathbf{X}_i = (X_{i\,1}, \ldots, X_{i\,p})^T \in \mathbb{R}^p$
- **Response variable:** Y_i is the class label. In two class classification problem, it is considered to be either $Y_i \in \{0, 1\}$ or $Y_i \in \{-1, 1\}$.
- Looking for the functional relationship between Y_i and $\mathbf{X}_i$

$$p(\mathbf{x}) = P_{_F}(Y = 1 \,|\, \mathbf{X} = \mathbf{x})$$

Classification Examples

- Netflix Prize: Predict if (how much) someone is going to enjoy a movie based on their movie preferences.
- Spam email: Predict if an email is spam based on relative frequencies of most common words and punctuation marks in the email messages.
- Loan request: When you apply for a loan, the lender will look at your credit, employment, and residence history in order to determine if you qualify for a loan. The decision is made based on the past experience, i.e. how have people with similar histories behaved with respect to paying off their loans.

- Disease diagnonsis: If you are suspected of having a disease the doctor will take your temperature, blood pressure, check various other things probably have some tests done and then, based on all the data collected, determine the presence or absence of the disease. The decision is made based on the past experience, i.e. how have people with similar medical characteristics been shown to have the disease.

When we do a classification, we need to be concerned with how good our method is. In the medical situation, if a person who has a disease is classified as not having it the person could die. On the other hand, if the person who does not have the disease is classified as having it, a needless operation could result.

Regression

- **Data:** $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.
- **Explanatory variables:** $\mathbf{X}_i = (X_{i\,1}, \ldots, X_{i\,p})^T \in \mathbb{R}^p$
  - Can be random with a distribution.
  - Or deterministic chosen by the experimenter.
- **Response variable:** $Y_i \in \mathbb{R}$
- Looking for the functional relationship between Y_i and $\mathbf{X}_i$

$$e(\mathbf{x}) = E_{_F}(Y \,|\, \mathbf{X} = \mathbf{x})\,.$$

Regression: Examples

- Predict tomorrow's stock market price given current market conditions and information.
- Predict the amount of prostate specific antigen (PSA) in the body as a function of a number of different clinical measurements.
- Predict the temperature at any location inside a building using weather data, time, door sensors, etc.
- How does class size affect student participation?
- How much different clinical measurements affect the amount of prostate specific antigen (PSA) in the body.

Parametric Regression

- **Multiple Linear Regression:** We have the strong assumption that

$$Y_i = \alpha + \boldsymbol{\beta}^T \mathbf{X}_i + \epsilon_i = \alpha + \beta_1 X_{i\,1} + \cdots + \beta_p X_{i\,p} + \epsilon_i \,.$$

  - $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma^2)$, where σ^2 unknown.
  - More generally, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. from a symmetric, unimodal distribution.

Nonparametric Regression

Assumes that Y_i is related to $\mathbf{X}_i$ through

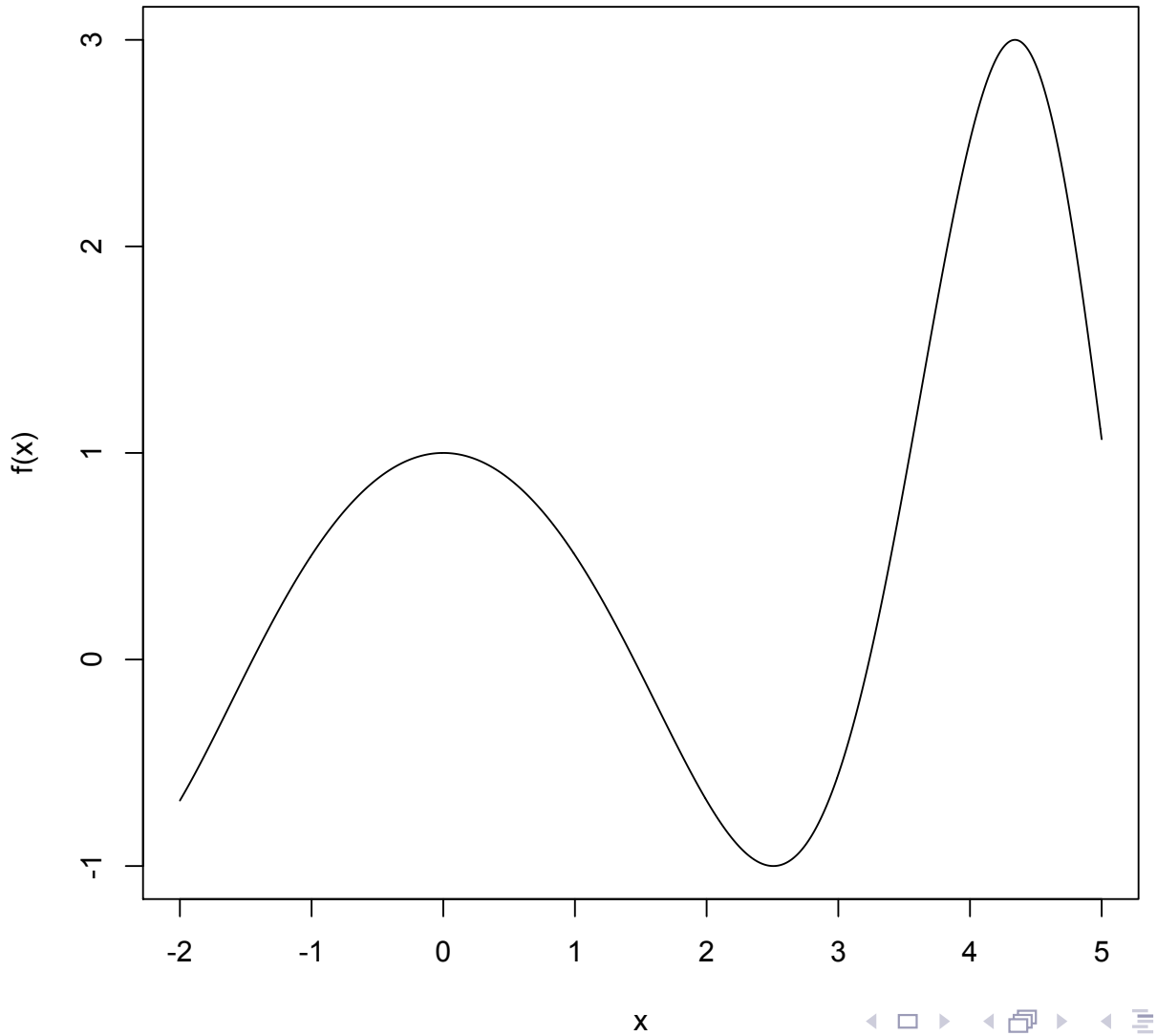
$$Y_i = f(\mathbf{X}_i) + \epsilon_i \,,$$

where f is an unknown function.

- Similar to parametric regressions, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. from a symmetric, unimodal distribution such as $N(0, \sigma^2)$.

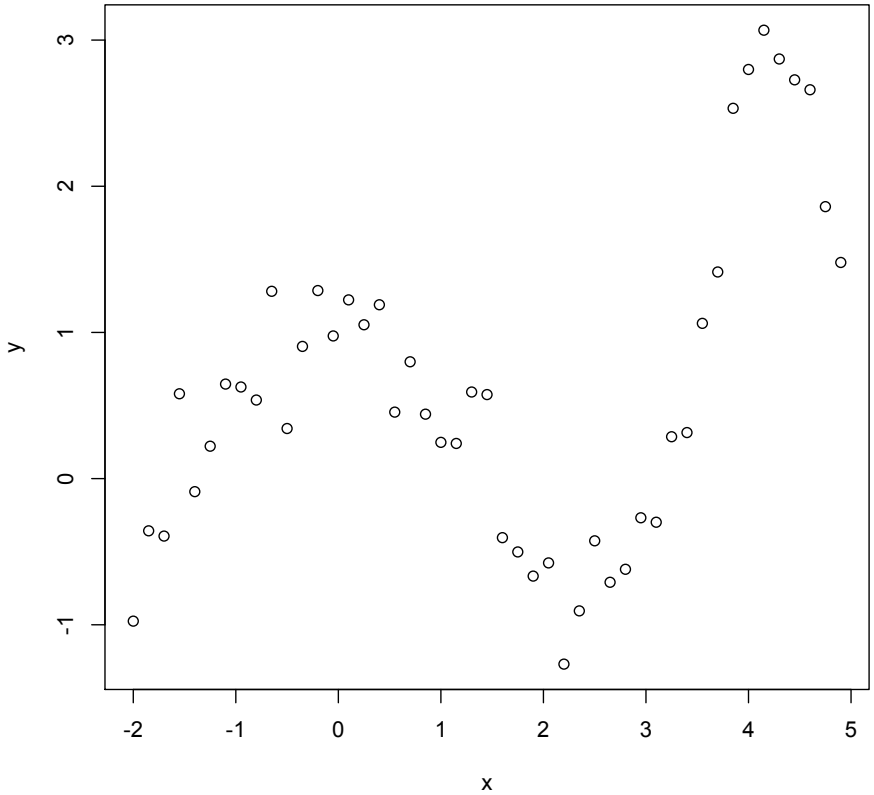
- A central assumption is smoothness of f. It ensures good behaviours such as consistency of estimators $\widehat{f}$.
  - Kernel methods: f belongs to a Sobolev space.
  - Spline methods: f is piecewise continuous.
  - Penalization or regularization methods: Penalize the roughness of f so the data help to determine how wiggly $\widehat{f}$ should be.

True curve $f(x)$



Simulated data

Simulated data with evenly spaced values of x in $[-2, 5]$, $\epsilon \sim N(0,\, 0.25^2)$ and $y = f(x) + \epsilon$.



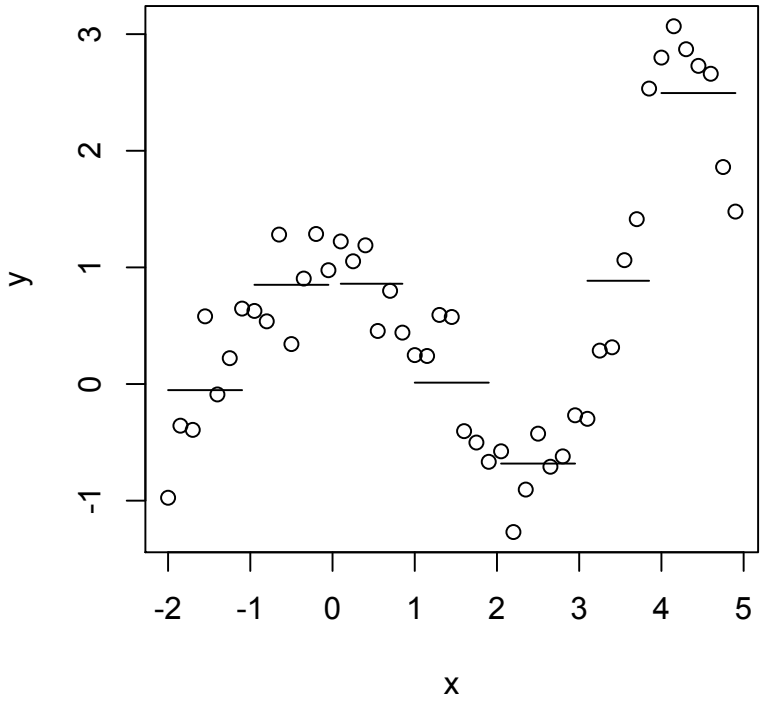
Suppose you don't know $f(x)$, the question is how to recover f from the data given above?

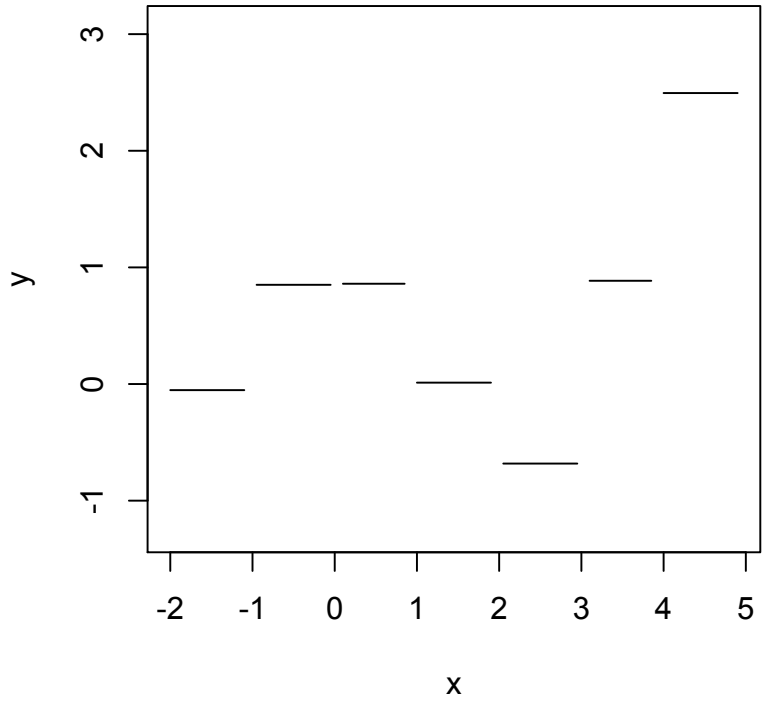
- Partition the real line (range of $\mathbf{X}$) into prespecified disjoint bins.
- In our simulated example

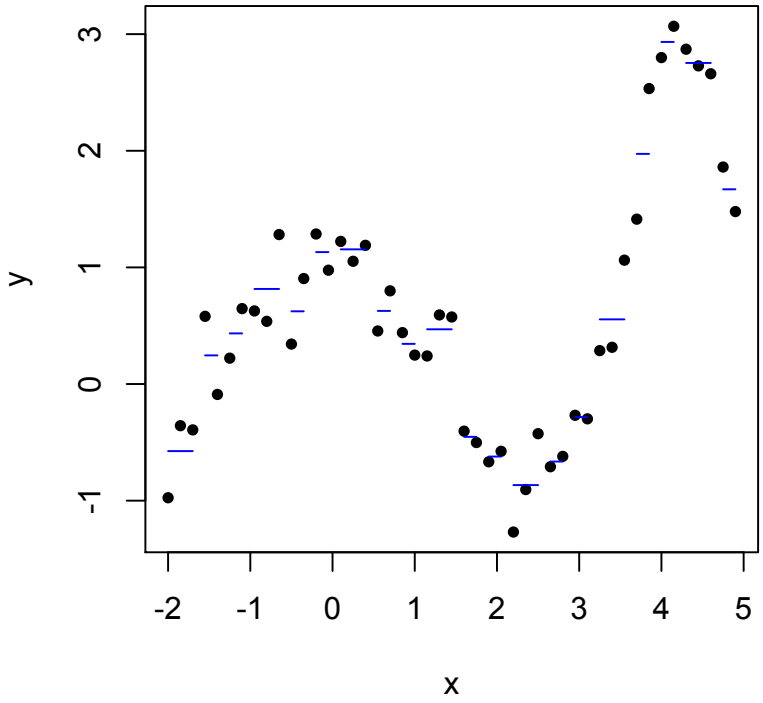
$$(-\infty, -2),\ [-2, -1),\ [-1, 0),\ [0, 1),\ [1, 2),\ [2, 3),\ [3, 4),\ [4, 5),\ [5, \infty)\,.$$

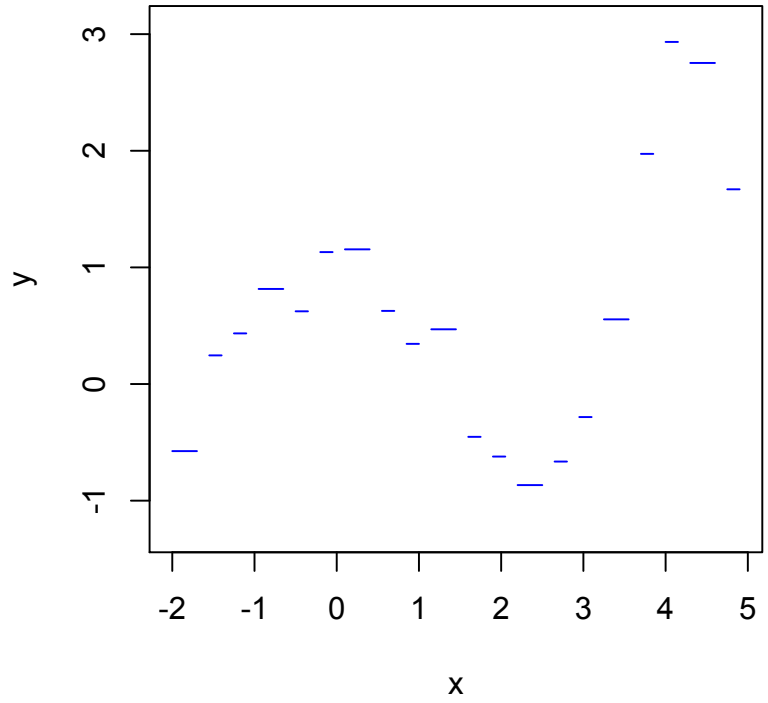
- For bin i, average all Y values for the X values in that bin $[i,\ i+1)$.

$$\widehat{f}(x) = \frac{1}{\#\{X_j \in [i,\ i+1)\}} \sum_{X_j \in [i,\, i+1)} Y_j \qquad \text{if } x \in [i,\ i+1)$$


y
x
-2
-1
0
1
2
3
4
5
-1
0
1
2
3


y
x
-2
-1
0
1
2
3
4
5
-1
0
1
2
3


y
x
-2
-1
0
1
2
3
4
5
-1
0
1
2
3


y
x
-2
-1
0
1
2
3
4
5
-1
0
1
2
3

Why nonparametric regression?

There are several good reasons:

- *As a descriptive statistic:* A data analyst can obtain some insights about the unknown regression function by graphing an estimate of it on the scatter plot.
- *Lack of fit test:* For testing how well a simple parametric model is by comparing of the fits.
- *Semiparametric models:* Flexible adjustments for covariates. For example consider the one-way layout

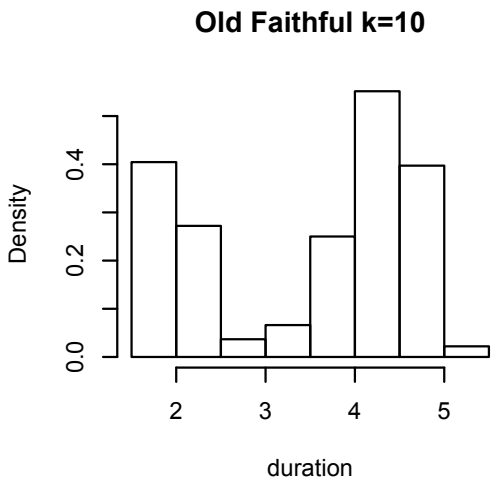
$$Y_{i\,j} = \mu + \alpha_j + f(x_{i\,j}) + \epsilon_{i\,j}\,,$$

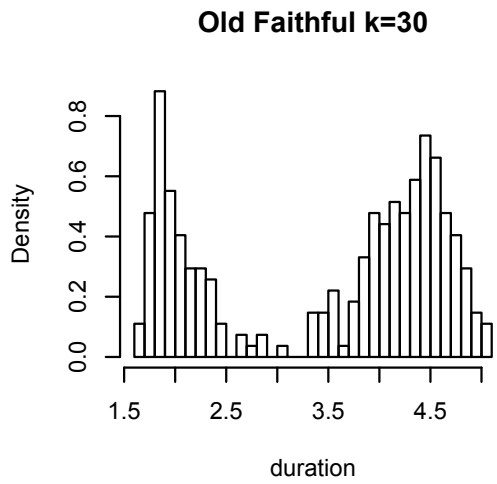
where $Y_{i\,j}$ is the response variable, and the effect of the covariate x could be controlled without imposing a parametric condition.

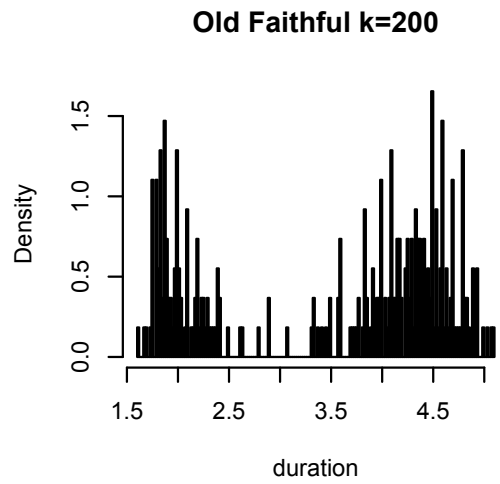
Density Estimation Examples

- Data is from the old Faithful geyser at Yellowstone National park, Wyoming, USA.
- It was observed from August 1st to August 15th 1985.
- During that time, data were collected on the duration of eruptions.
- There were 272 eruptions observed
- The recorded durations are given in R data frame faithful in library MASS.

Histograms of old faithful







Kernel density estimates of old faithful