Data Transformation

Smoothing

Data Transformation

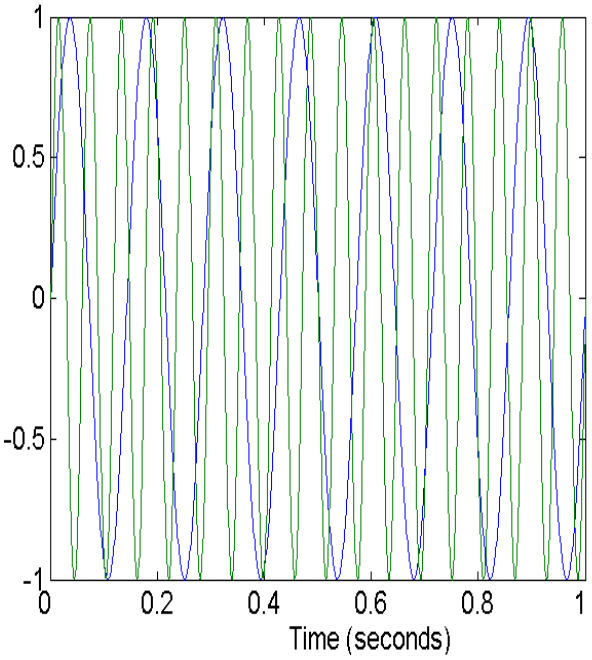
- Smoothing: removes noise from data

Noisy Data

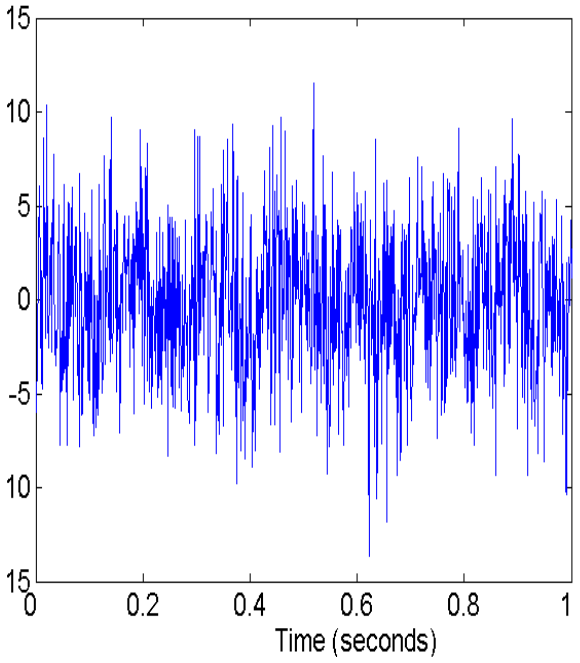
- Noise: due to random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone or "snow" on television screen



Two Sine Waves



Two Sine Waves + Noise

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Smoother summarises **the trend** of a response measurement as a function of predictors

Simple Discretization Methods: Binning

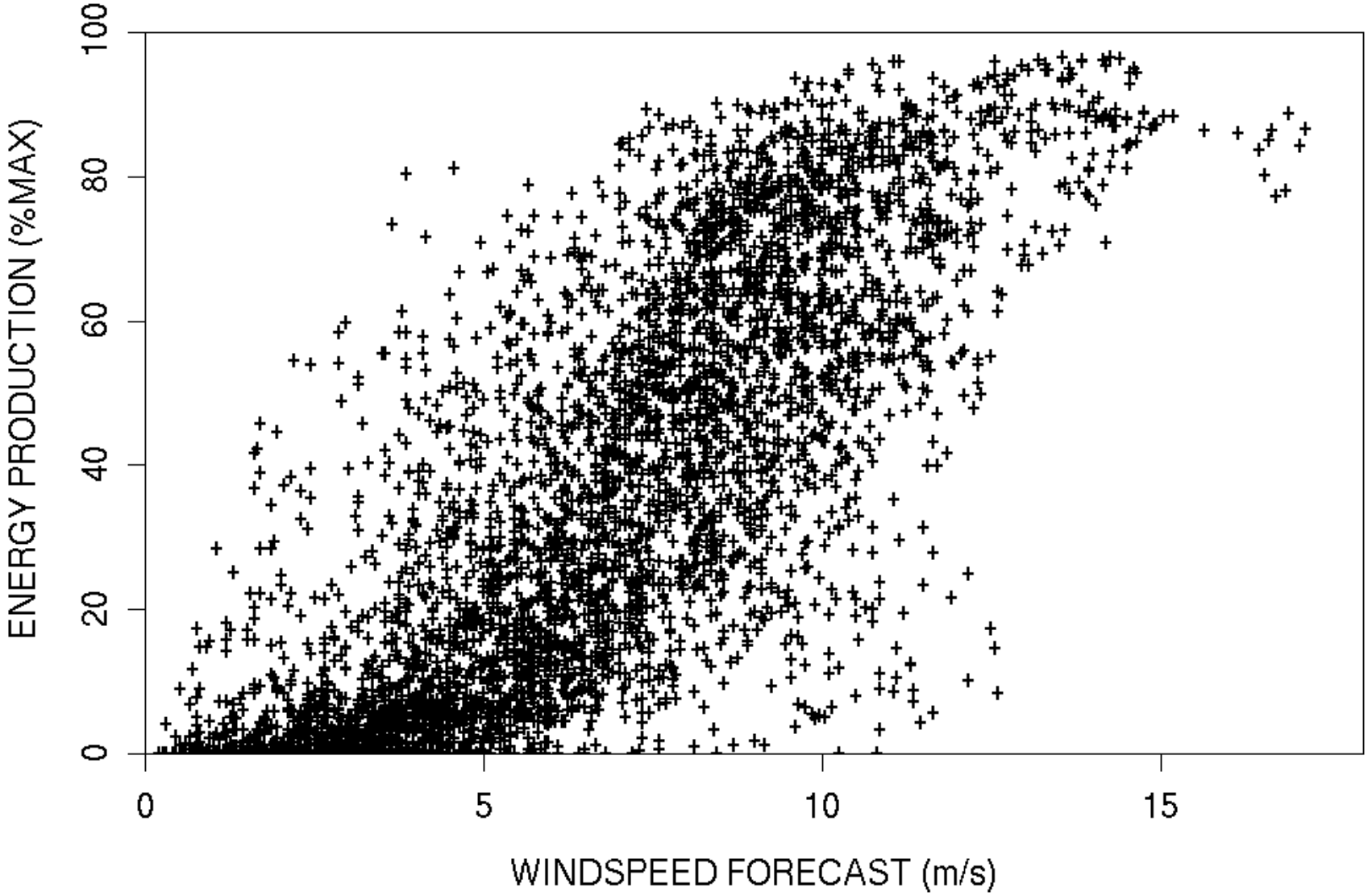
- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

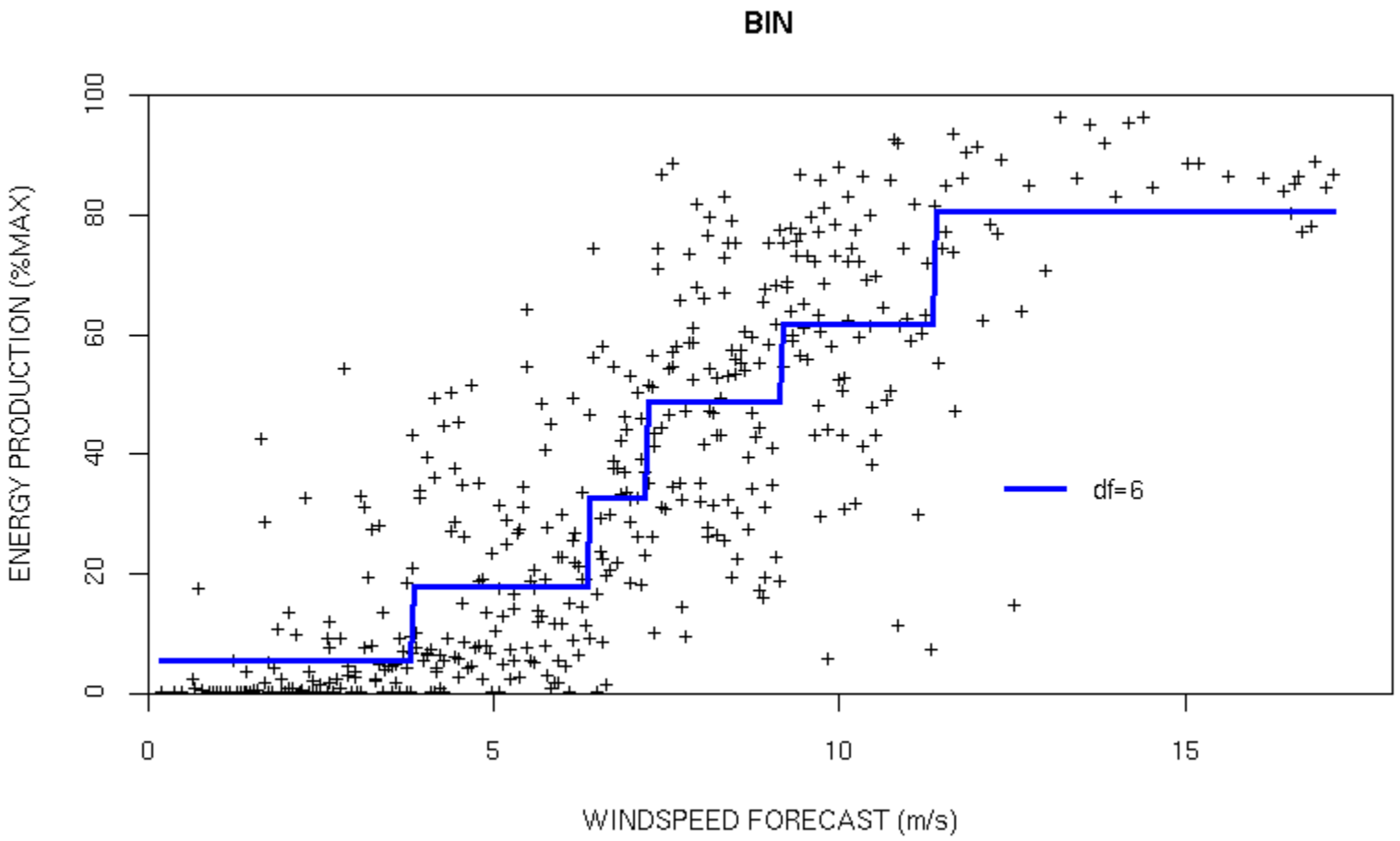
- Sorted data: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

SCATTERPLOT SMOOTHING: EXAMPLE

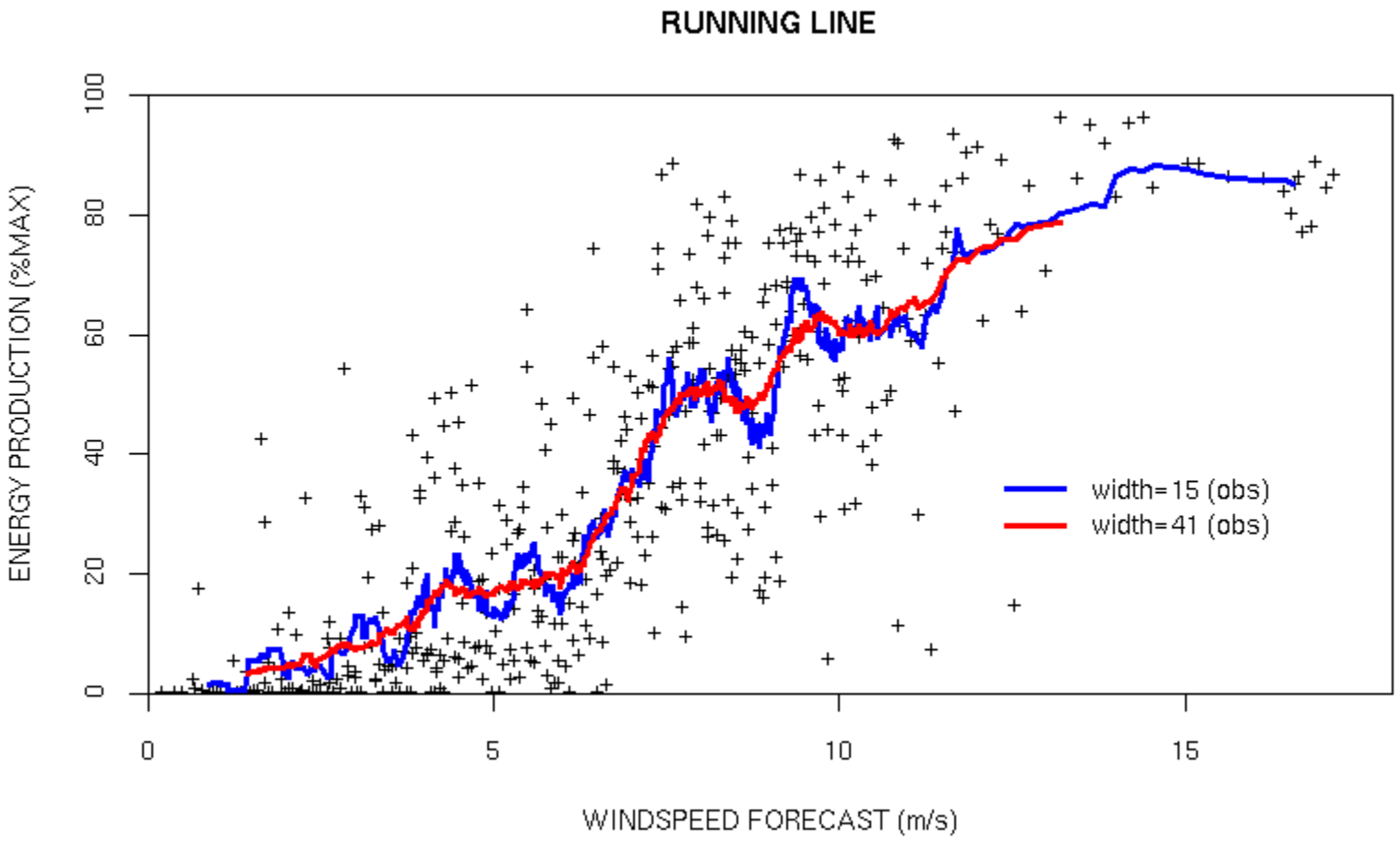


SMOOTHING: BIN SMOOTHER



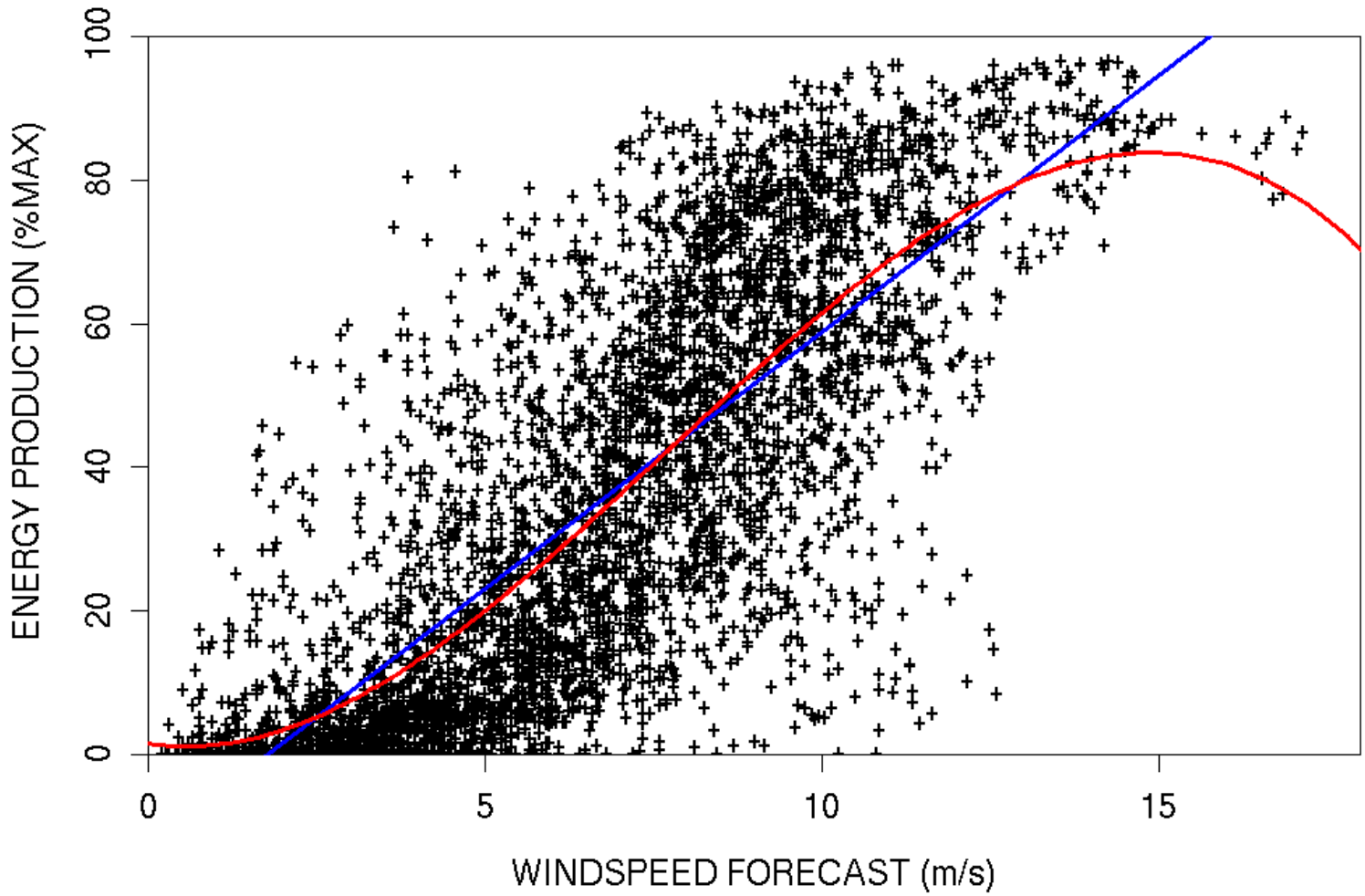
SMOOTHING: RUNNING LINE

- Running line (local regression)



SMOOTHING: POLYNOMIAL

- Linear and cubic parametric least squares fits:



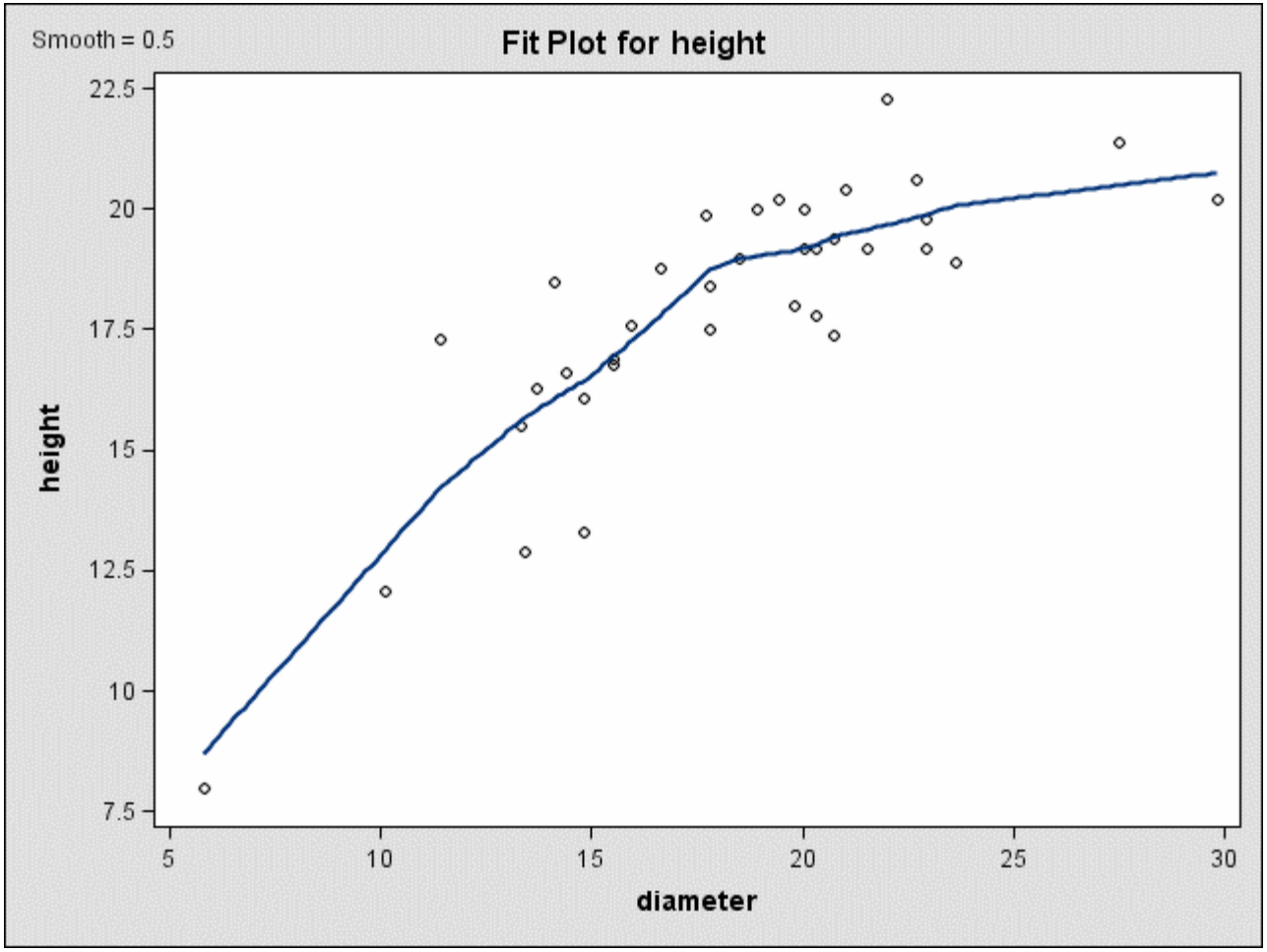
Non-parametric Smoothing: the Loess Method

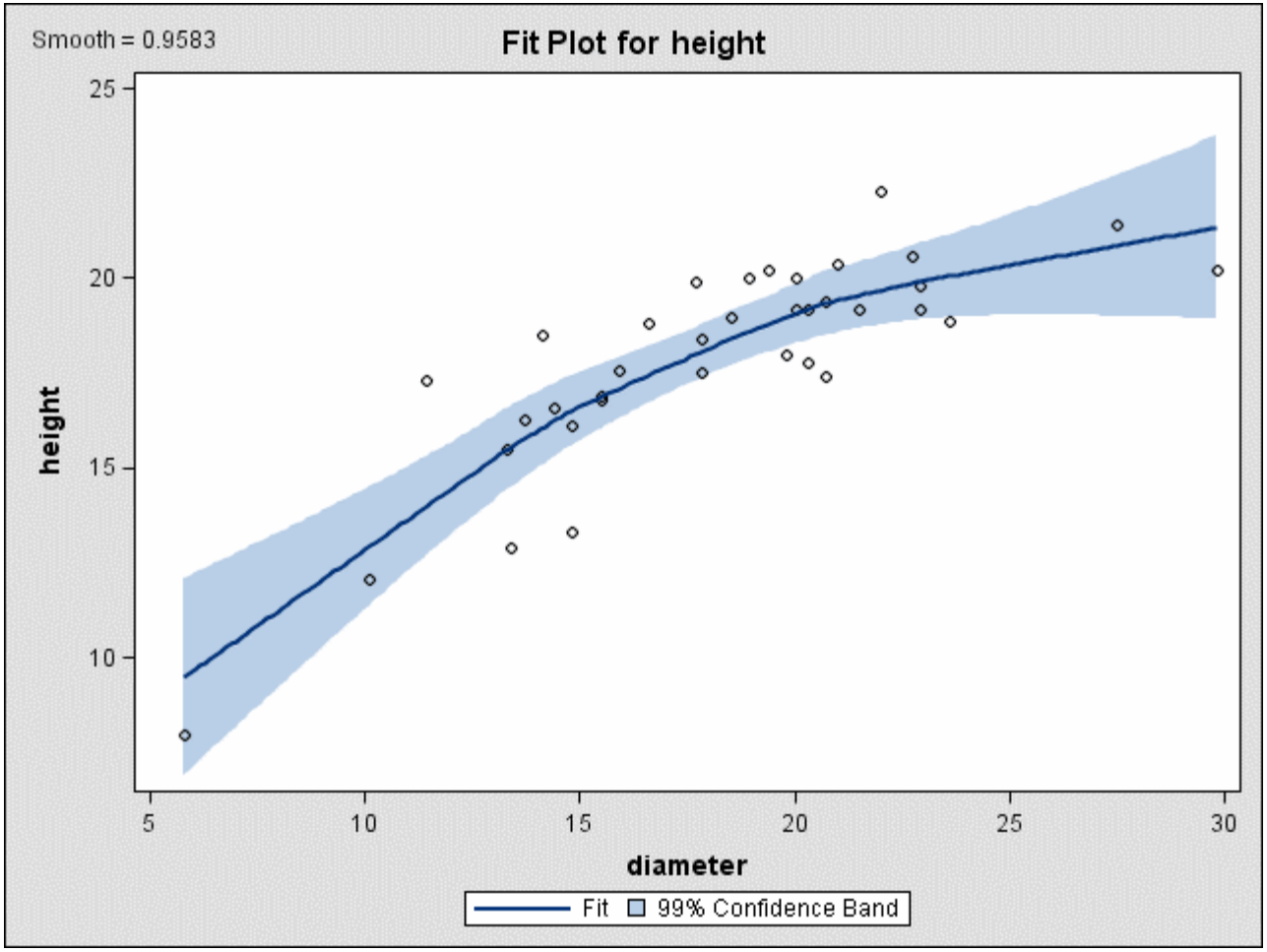
- **LOWESS= LOESS** is an Acronym for **LO**cally re**WE**ighte**d S**catterPlot **S**moothing
- In the LOESS (LOWESS) method, weighted least squares is used to fit **linear** or **quadratic** functions of the predictors at the centers of neighborhoods.
- The **radius** of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the *smoothing parameter,* in each local neighborhood controls the smoothness of the estimated surface.
- Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

The Loess Method: continued

Pseudo code:

- For i=1 to n, the ith measurement y_i of the response y and the corresponding measurement x_i of the vector x of p predictors are related by
  - $Y_i = g(x_i) + e_I$
- where g is the regression function and e_i is a random error.
- Idea: $g(x)$ can be locally approximated by a parametric function.
- Obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x.


Smooth = 0.5
Fit Plot for height
height
diameter
22.5
20
17.5
15
12.5
10
7.5
5
10
15
20
25
30


Smooth = 0.9583
Fit Plot for height
height
25
20
15
10
5
10
15
20
25
30
diameter
Fit
99% Confidence Band

Comments on LOESS

- fitting is done at each point at which the regression surface is to be estimated
- faster computational procedure is to perform such local fitting at a selected sample of points and then to blend local polynomials to obtain a regression surface
- can use the LOESS procedure to perform statistical inference provided the error distribution are i.i.d. normal random variables with mean 0.
- using the iterative reweighting, LOESS can also provide statistical inference when the error distribution is symmetric but not necessarily normal.
- by doing iterative reweighting, you can use the LOESS procedure to perform robust fitting in the presence of outliers in the data.

SMOOTHING: LOESS

- The smooth at the target point is the fit of a locally-weighted linear fit (tricube weight)