Lecture 2: Flexible Regression Models

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

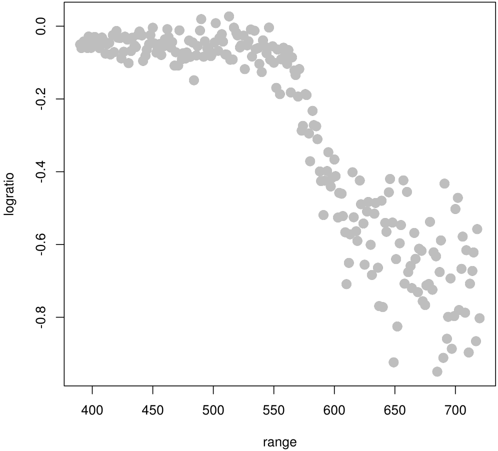
STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Linear Versus Nonlinear Regression
2. Nonlinear Regression
3. Polynomial Regression
4. Regression Splines

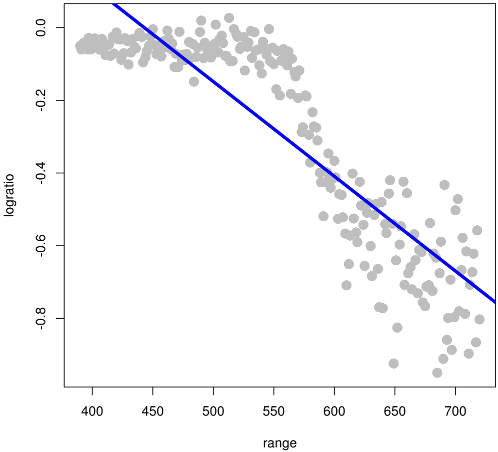
Lidar Data (Nonlinear effect of *x*)



Lidar Data

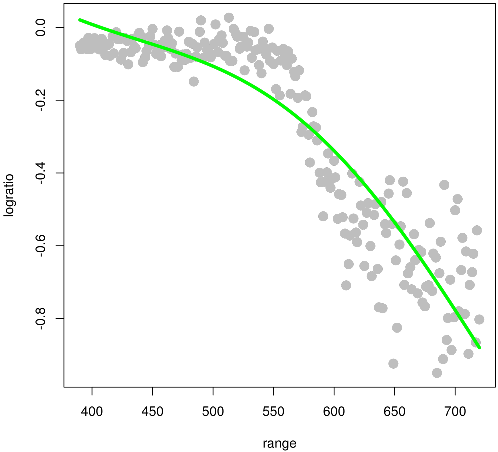
- Lidar is “light detection and ranging” or from “light” and "radar"
- Can monitor atmospheric pollutants
- *x* variable is `range`: distance travelled by light before reflection
- *y* variable is `logratio`: log of the ratio of two light sources, one tuned to mercury here
- See “Semiparametric Regression” by Ruppert, Wand, and Carroll, Chapter 2.7 for more details

Lidar Data (Simple Linear Regression)



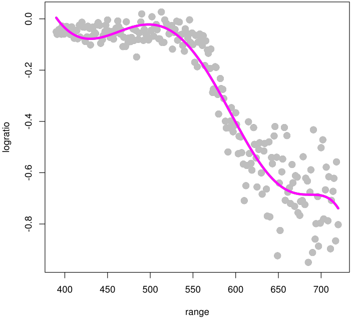
$$\text{logratio} \;=\; \beta_0 + \beta_1\,\text{range} + \varepsilon$$

Lidar Data (Linear or Nonlinear Regression?)



$$\text{logratio} \;=\; \beta_0 + \beta_1\,\text{range} + \beta_2\,\text{range}\,\sin\left((\text{range} - 400)\,\pi/300\right) + \varepsilon$$

Lidar Data (Linear or Nonlinear Regression?)



$$
\begin{aligned}
\text{logratio} &= \beta_0 + \beta_1 \text{range} \\
&+ \beta_2 \text{range} \sin\left((\text{range} - 400)\,\pi/300\right) \exp\left(\beta_3\,(\text{range} - 400)\,/500\right) \\
&+ \beta_4 \text{range} \cos\left((\text{range} - 400)\,\pi/300\right) \exp\left(\beta_5\,(\text{range} - 400)\,/500\right) \\
&+ \varepsilon
\end{aligned}
$$

Nonlinear Regression

- Model: $E\left[\,Y \middle|\, x_1, x_2, \ldots, x_p\right] = f\left(x_1, x_2, \ldots, x_p; \beta_1, \beta_2, \ldots, \beta_k\right)$
- Estimation:

$$\hat{\beta}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - f(x_{i1}, x_{i2}, \ldots, x_{ip}; \boldsymbol{\beta})\right)^2$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$

- Inference?

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - f(x_{i1}, x_{i2}, \ldots, x_{ip}; \boldsymbol{\beta})\right)^2 = \mathbf{0}$$

$$\sum_{i=1}^{n} r_i \, \frac{\partial f}{\partial \boldsymbol{\beta}}(x_{i1}, x_{i2}, \ldots, x_{ip}; \boldsymbol{\beta}) = \mathbf{0}$$

Nonlinear Regression

- Model: $E\left[\,Y \mid x_1, x_2, \ldots, x_p\right] = f\left(x_1, x_2, \ldots, x_p; \beta_1, \beta_2, \ldots, \beta_k\right)$
- This is typically a nonlinear model
- But it is fully parametric
- The parameters are $\beta_1, \beta_2, \ldots, \beta_k$
- Using MLE (or LS) we can obtain estimates $\hat{\beta}_1$, . . . , $\hat{\beta}_k$
- ... and associated standard errors!

Nonlinear Regression

- Consider the trade union data
- Compare a linear model

$$E\,[\text{wage}|\,\text{age}] \;=\; \alpha \,+\, \beta\,\text{age}$$

with the following nonlinear one

$$E\,[\text{wage}|\,\text{age}] \;=\; \alpha \,+\, \beta\,\exp\left(-(\text{age}-20)/\gamma\right)$$

- Which one is to be preferred? **Clicker question 3.**

Polynomial Regression Models

- Sometimes it's difficult to find an appropriate family of functions
- Polynomials are a natural choice. For one-dimensional x, expand $E[Y|x] = f(x)$ around x_0 by a Taylor series

$$f(x) = f(x_0) + \frac{1}{2} f'(x_0)(x - x_0) + \cdots$$

$$+ \frac{1}{k!} f^{(k-1)}(x_0)(x - x_0)^{k-1} + R_k,$$

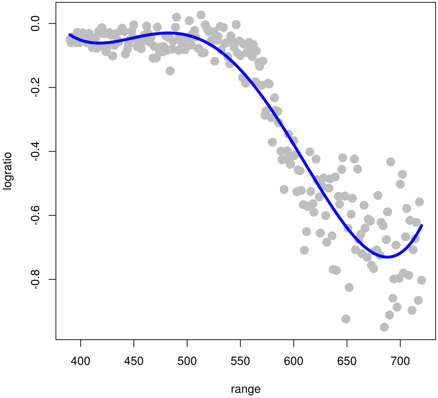
  i.e., a constant, plus a term linear in x, plus ...

- Hence, we can try

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$$

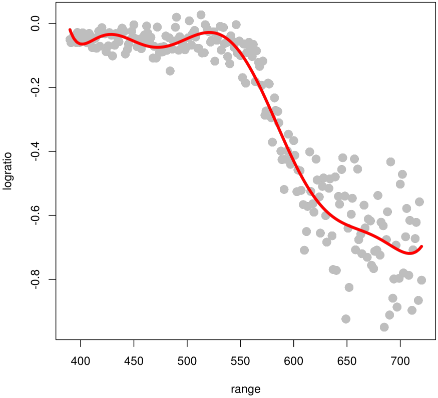
- This is a linear model! (**WHY?**)
- But...

Lidar Dataset: Polynomial Regression



$$\text{logratio} = \beta_0 + \beta_1 \,\text{range} + \beta_2 \,\text{range}^2 + \beta_3 \,\text{range}^3 + \beta_4 \,\text{range}^4$$

Lidar Dataset: Polynomial Regression



$$\text{logratio} = \beta_0 + \beta_1 \, \text{range} + \beta_2 \, \text{range}^2 + \ldots + \beta_4 \, \text{range}^{10}$$

Regression Splines

- Consider the (family of) functions for one x variable:

$$f_k(x) = (x - \kappa_k)_+ = \begin{cases} x - \kappa_k & \text{if } x - \kappa_k > 0 \\ 0 & \text{otherwise} \end{cases}$$

where κ_k, $1 \leq k \leq K$ are *knots* (to be chosen)

- Model

$$E[Y|x] = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{k+1} f_k(x)$$

(Similar terms for all variables x_j)

- Is this a linear model? **Clicker question 4.**

Regression Splines

- The knots can be chosen arbitrarily
- It is customary to select them based on the sample

$$\kappa_k = \left(\frac{k}{K+1} \right) 100\% \text{ quantile of the observed } x$$

- For example, with $K = 4$:

$$\kappa_1 = 20\% , \qquad \kappa_2 = 40\% , \qquad \text{etc.}$$

Lidar Data: Regression Splines, 5 Knots