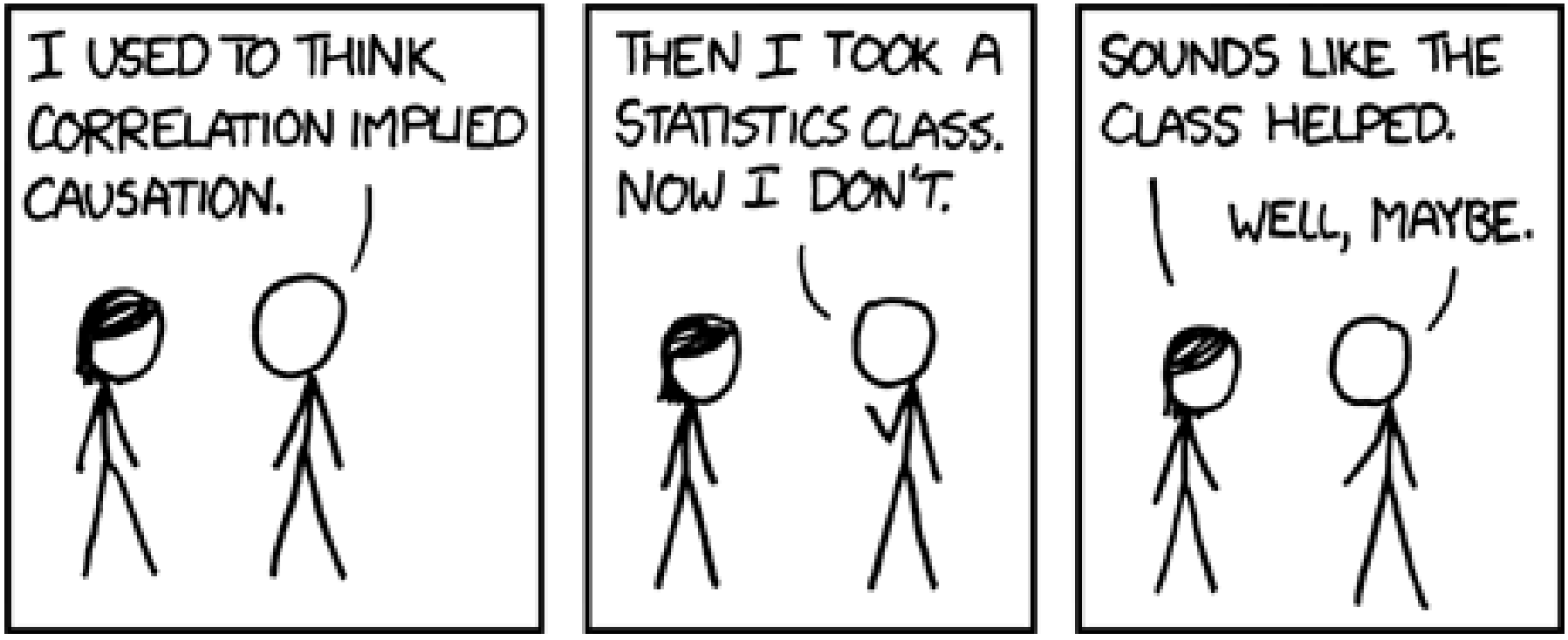
STAT406- Methods of Statistical Learning Lecture 8

Matias Salibian-Barrera

UBC - Sep / Dec 2019



https://xkcd.com/552/

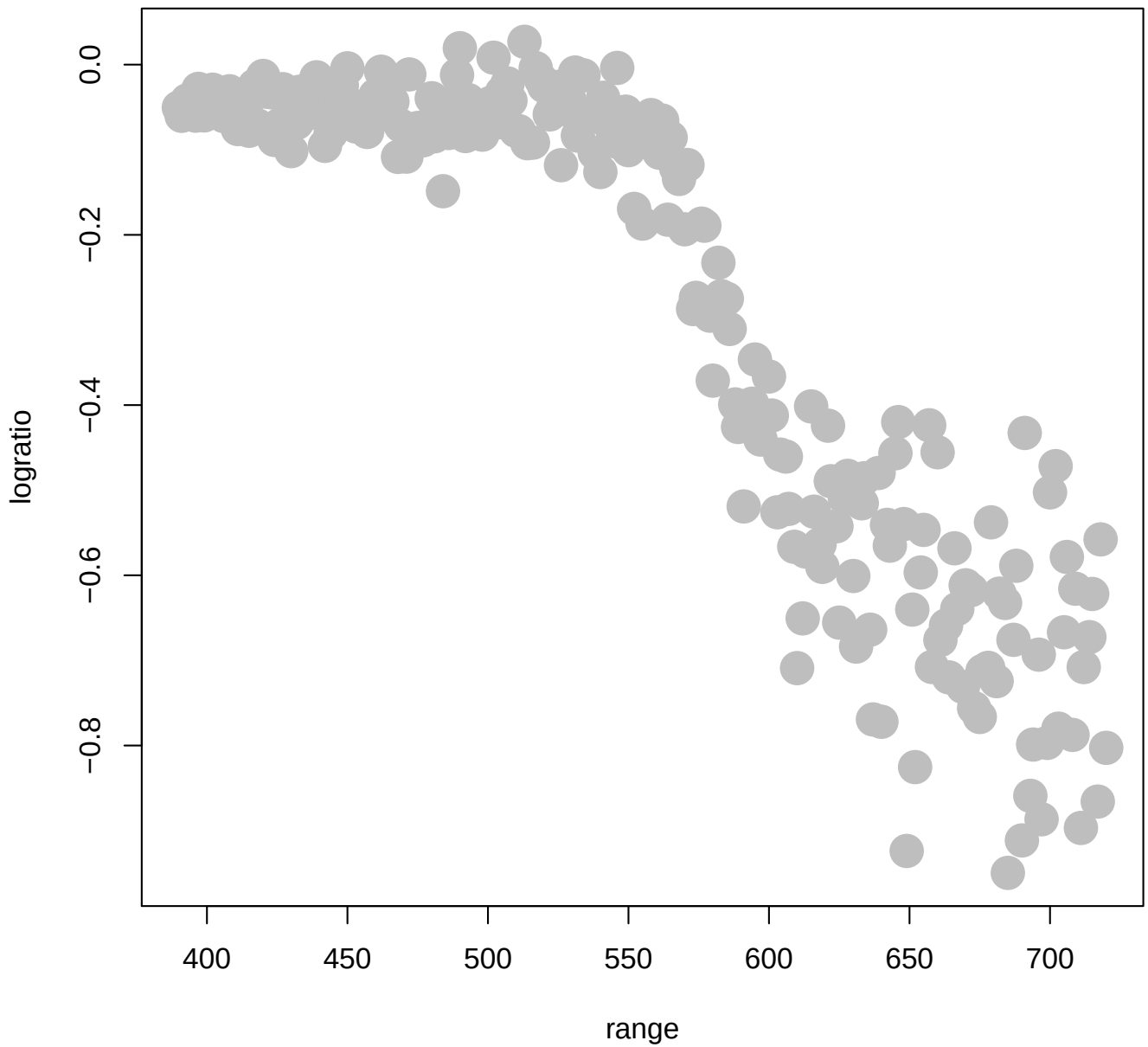
More flexible regression

- What if the regression function

$$E[Y|\mathbf{X}] = f(\mathbf{X})$$

  is not linear?
- Example LIDAR

LIDAR



Non-linear regression

- Model: $E[Y|X_1, X_2, \ldots, X_p] = f(X_1, X_2, \ldots, X_p; \theta_1, \theta_2, \ldots, \theta_k)$
- This is typically a non-linear model
- But it is fully parametric
- The parameters are $\theta_1, \theta_2, \ldots, \theta_k$
- Using MLE (or LS) we can obtain estimates $\hat{\theta}_1, \ldots, \hat{\theta}_k$
- ... and associated standard errors!

Non-linear regression

- Sometimes it's difficult to find an appropriate family of functions
- Polynomials are a natural choice

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

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

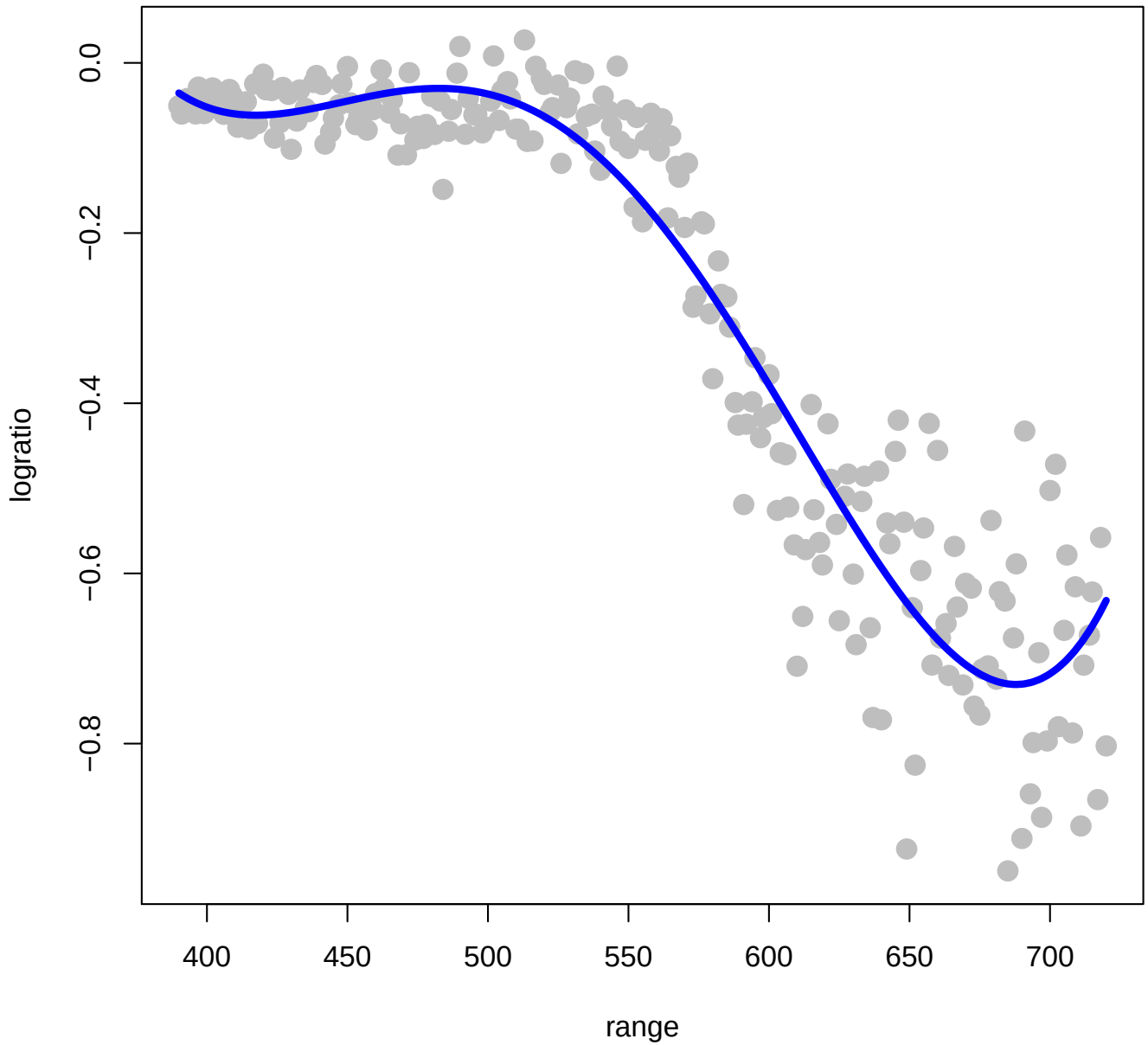
Non-linear regression

- Hence, we can try

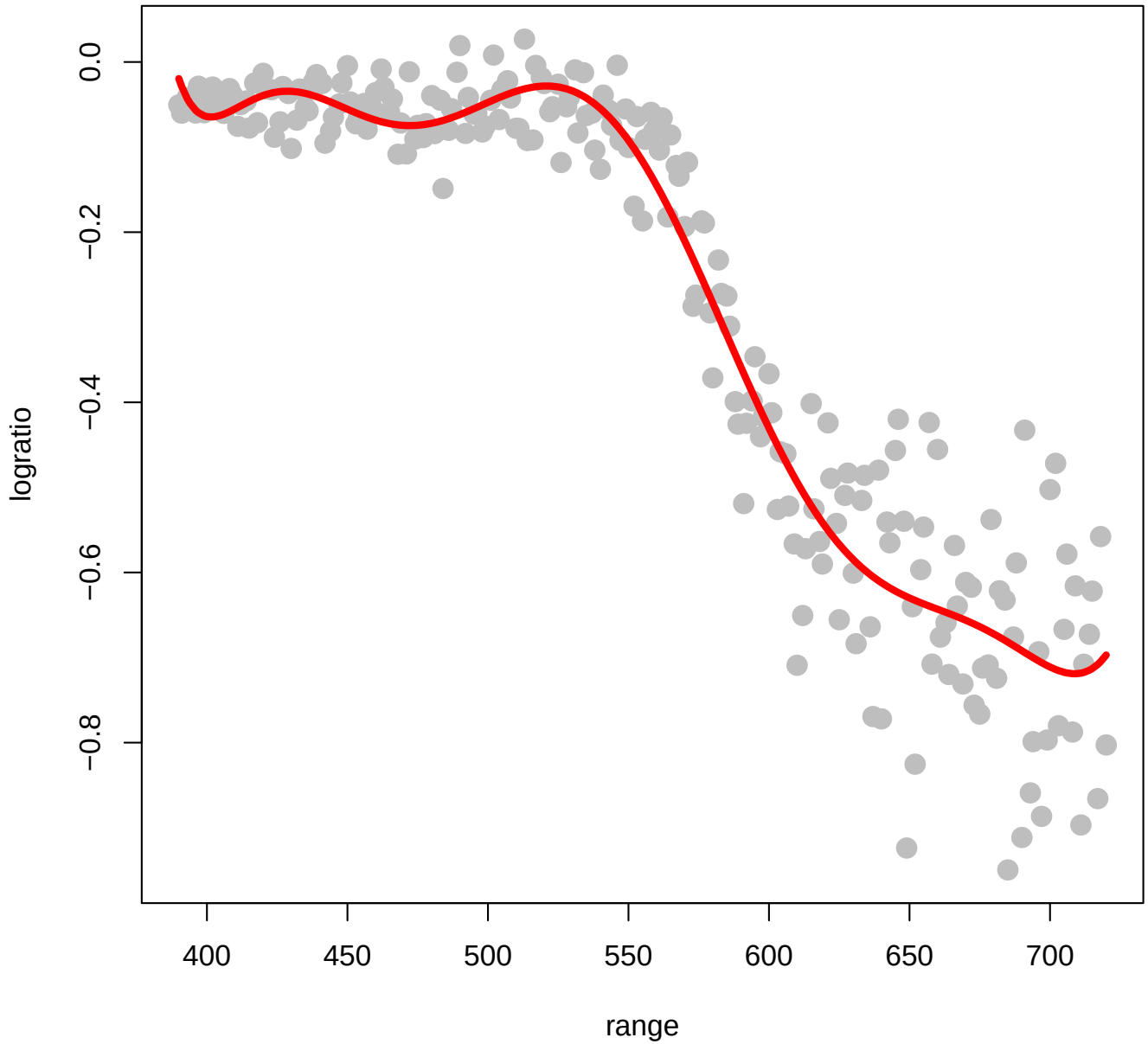
$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

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

LIDAR - 4th deg. polynomial



LIDAR - 10th deg. polynomial



More flexible bases

- Consider the (family) of function(s)

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

  where κ_j are *knots*
- Model

$$E[Y|X] = \beta_0 + \beta_1 X + \sum_{j=1}^{K} \beta_{j+1} f_j(X)$$

- This is a linear model

More flexible bases

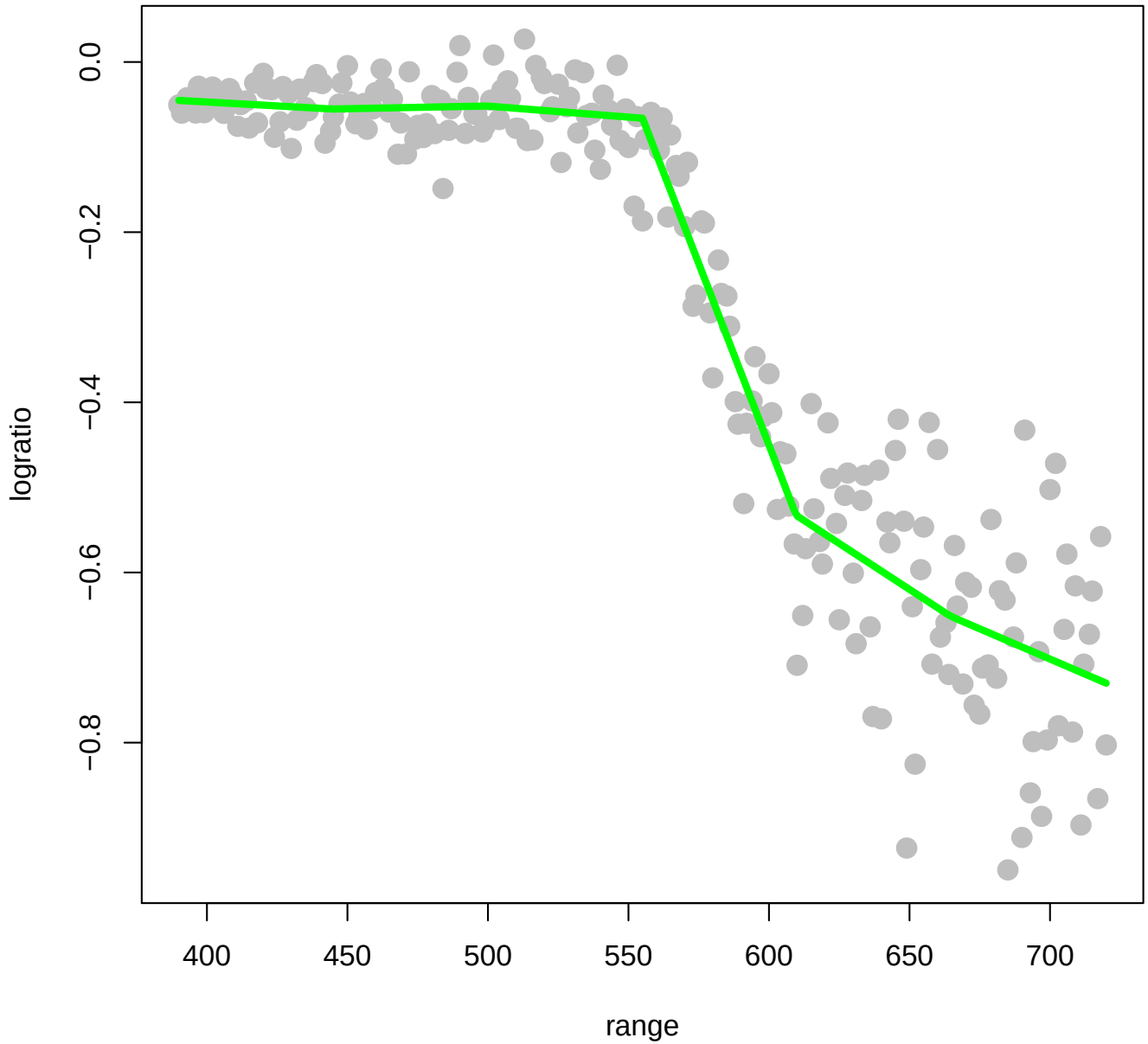
- The knots can be chosen arbitrarily
- It is customary to select them based on the sample

$$\kappa_j = \frac{j}{K+1} \text{ 100\% quantile of } x$$

- For example, with $K = 4$:

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

Regression splines, 5 knots



More flexible bases

- Consider a smoother basis

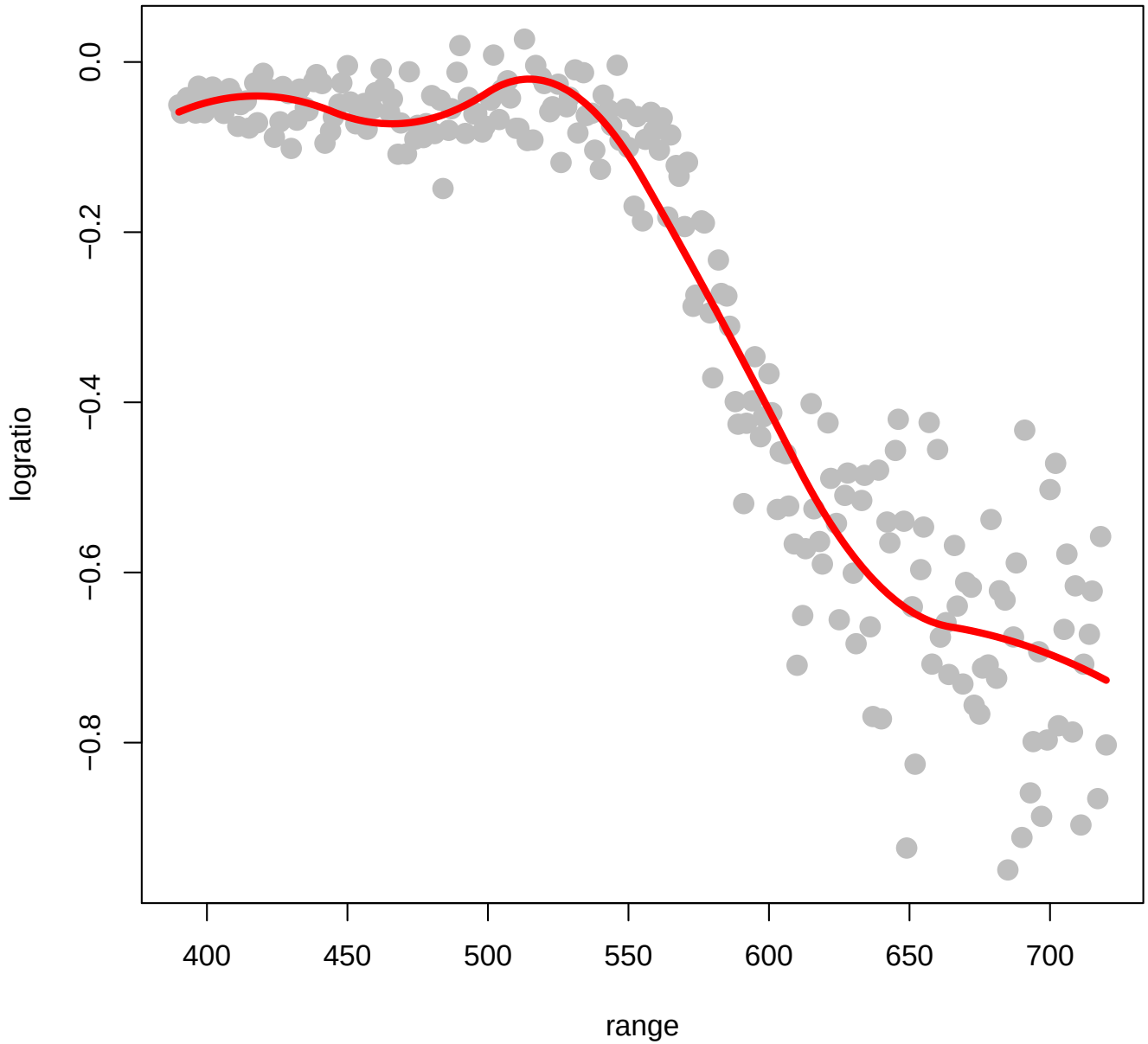
$$f_j(x) = (x - \kappa_j)_+^2 = \begin{cases} (x - \kappa_j)^2 & \text{if } x - \kappa_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where κ_j, $1 \leq j \leq K$ are *knots*

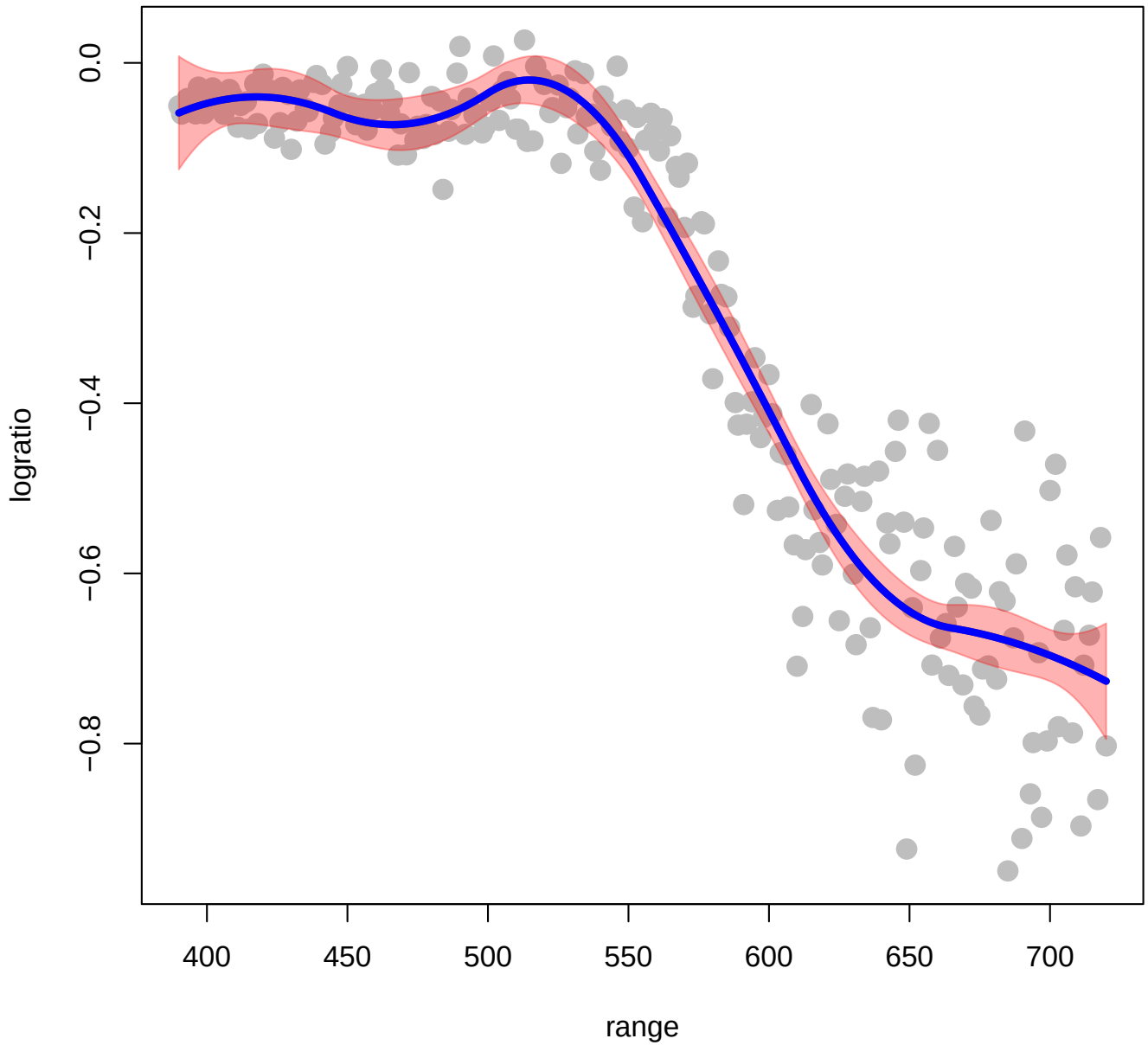
- Model

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \sum_{j=1}^{K} \beta_{j+2} f_j(X)$$

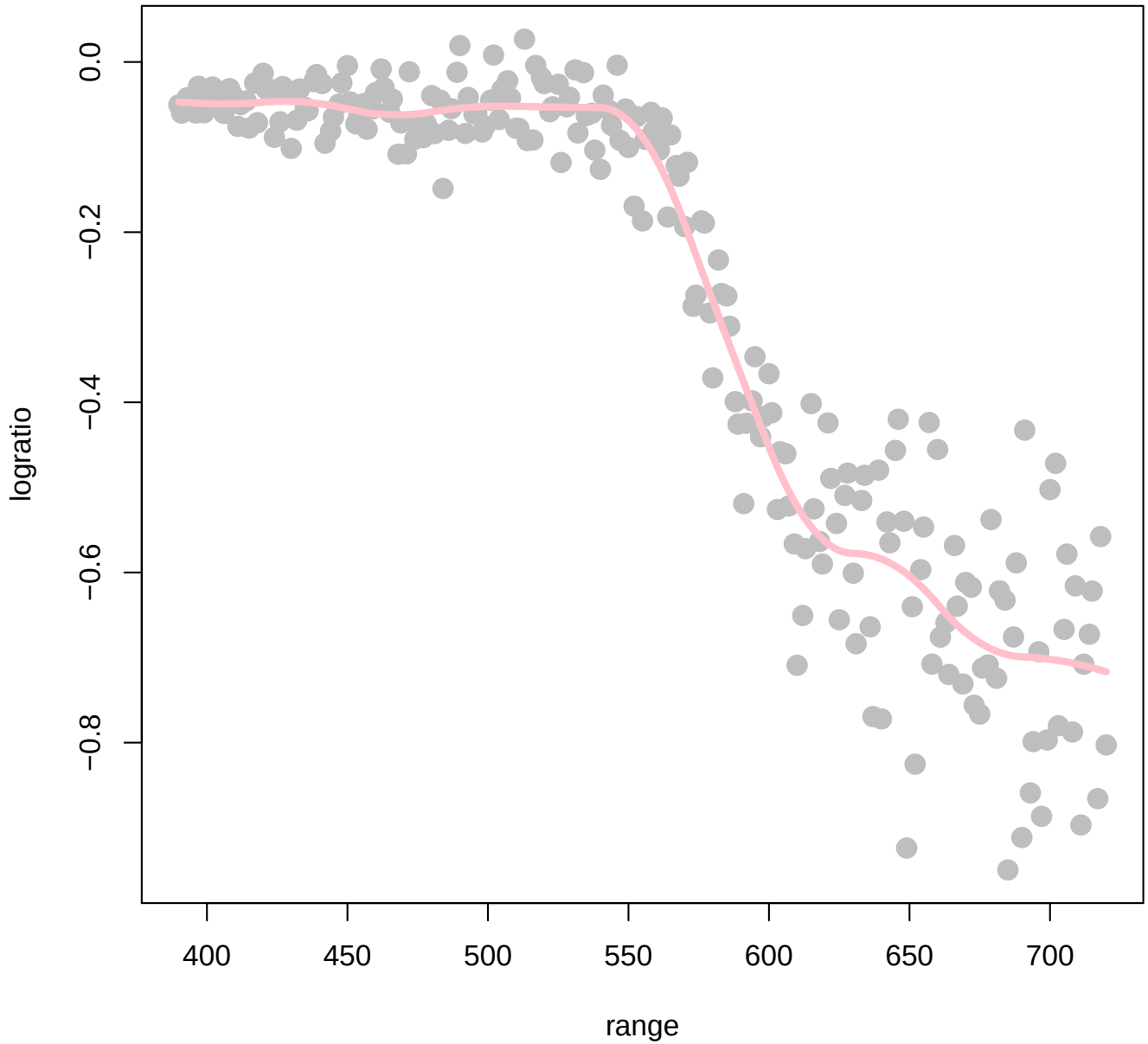
Quadratic splines, 5 knots



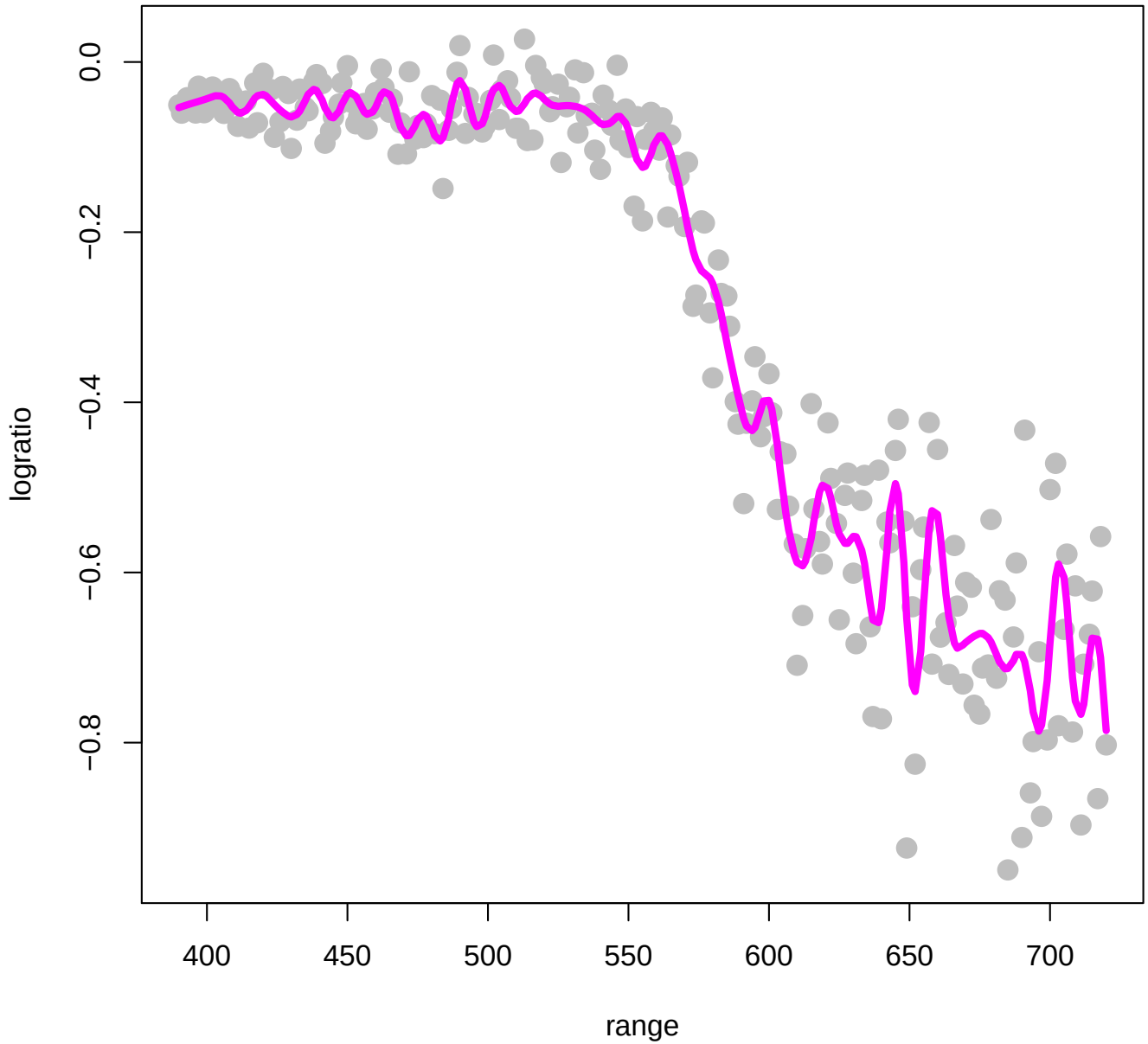
Quadratic splines, 5 knots + SEs



Quadratic splines, 10 knots



Quadratic splines, 50 knots



More flexible bases

- Cubic splines will be useful

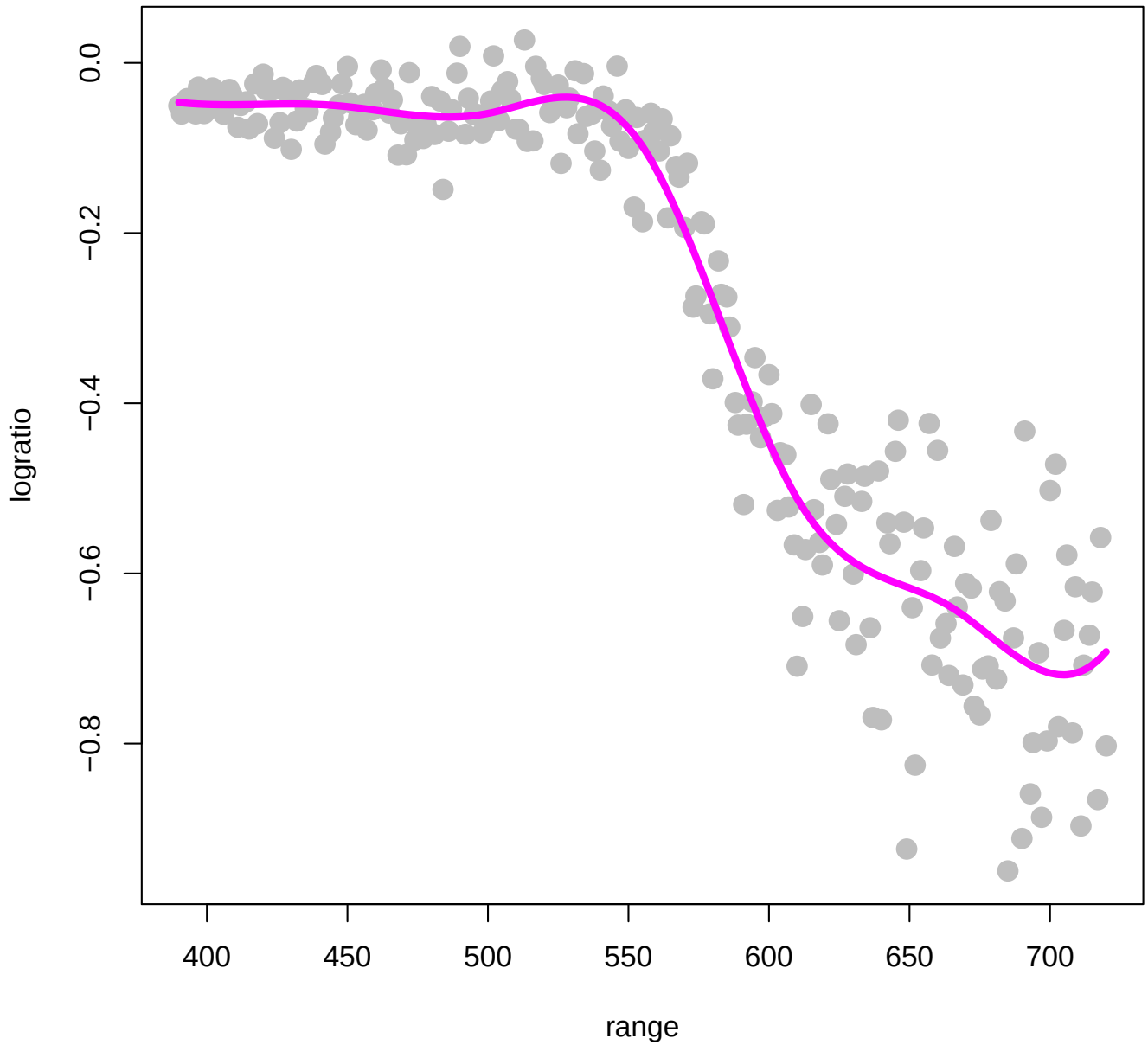
$$f_j(x) = (x - \kappa_j)_+^3 = \begin{cases} (x - \kappa_j)^3 & \text{if } x - \kappa_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

where κ_j, $1 \leq j \leq K$ are *knots*

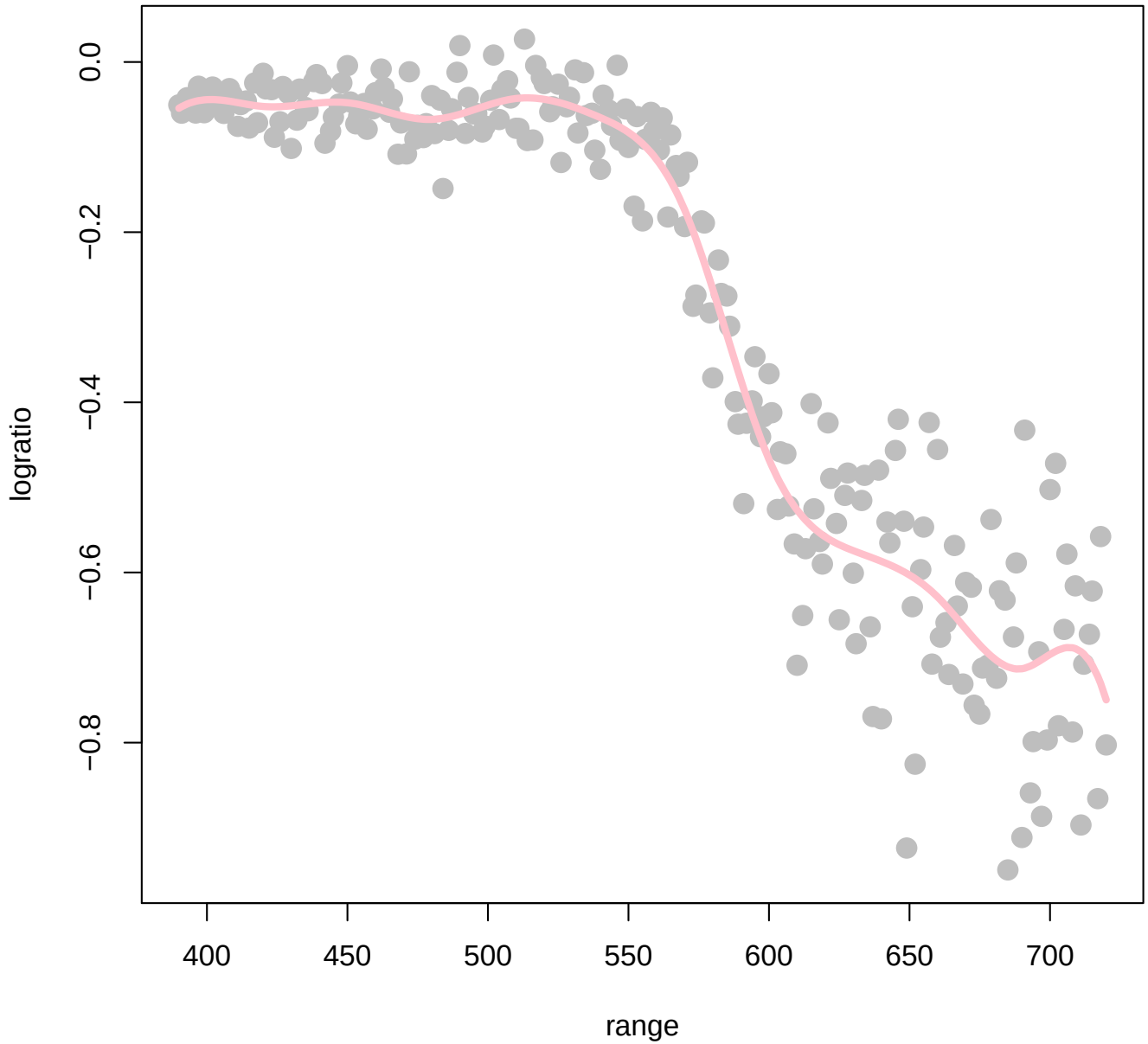
- Model

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \sum_{j=1}^{K} \beta_{j+3} f_j(X)$$

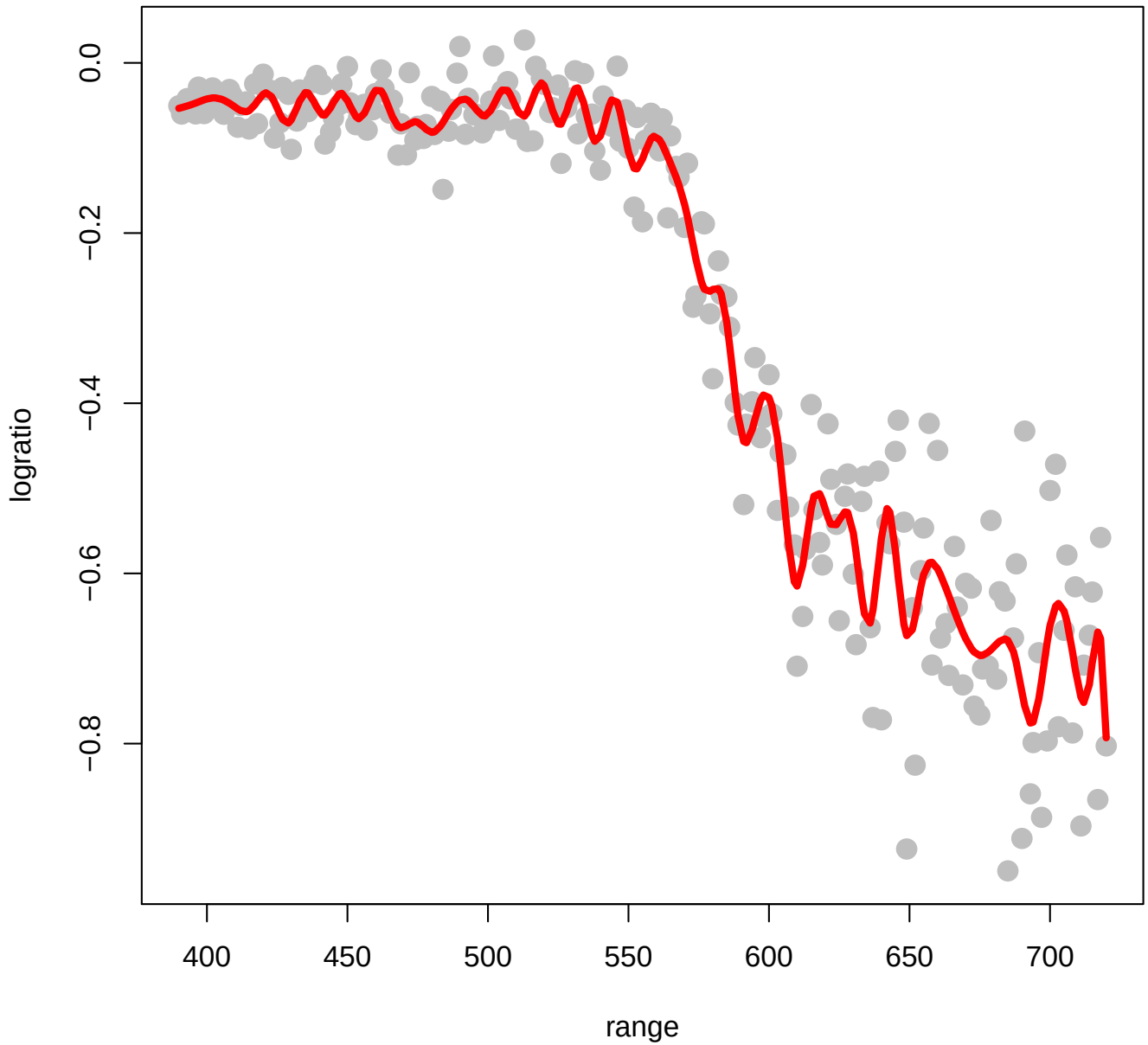
Cubic splines, 5 knots



Cubic splines, 10 knots



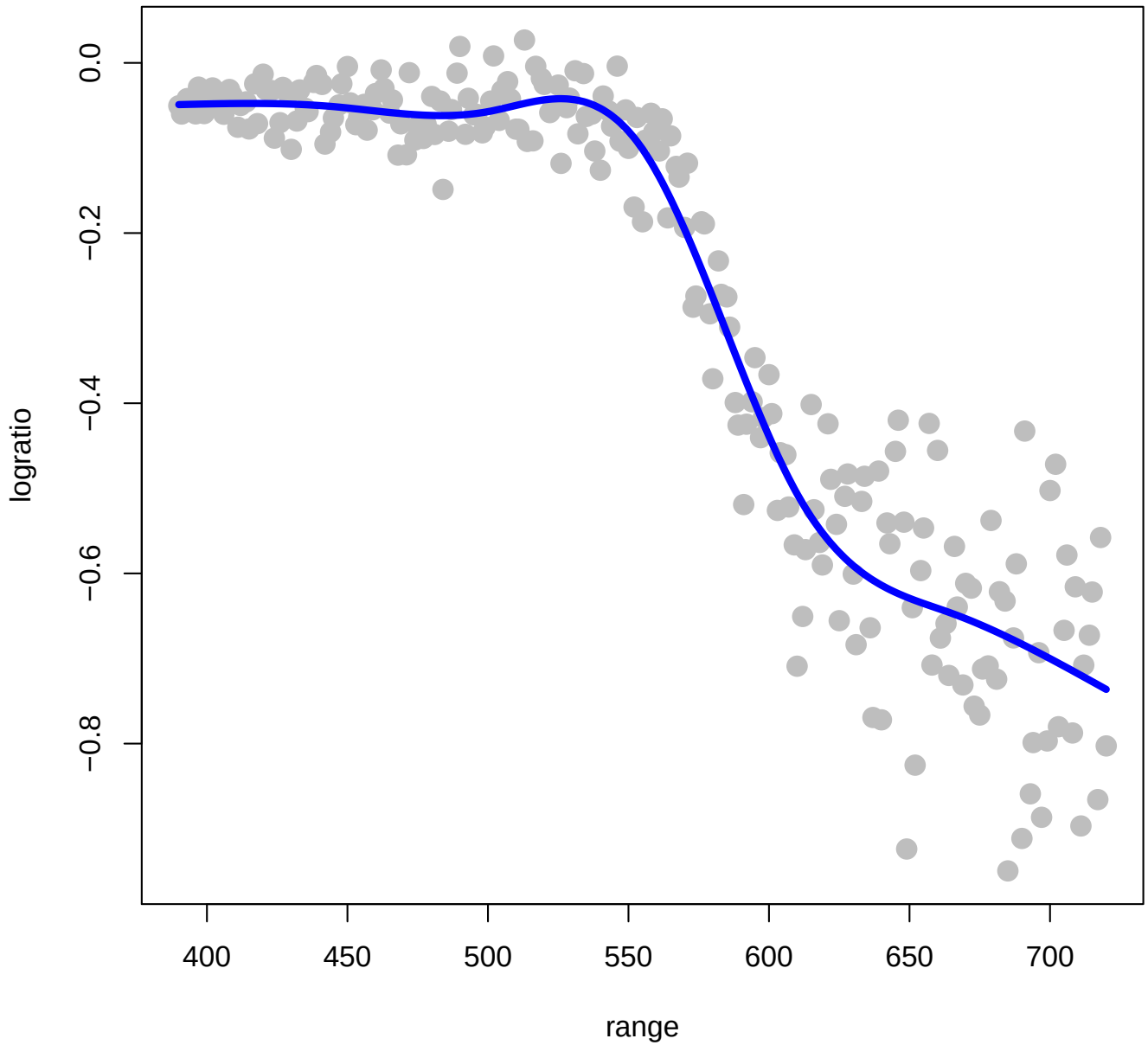
Cubic splines, 50 knots



More flexible bases

- Need to choose number and location of knots
- Need to make them less wiggly at the ends (Natural cubic splines)

Natural cubic spline, 5 knots



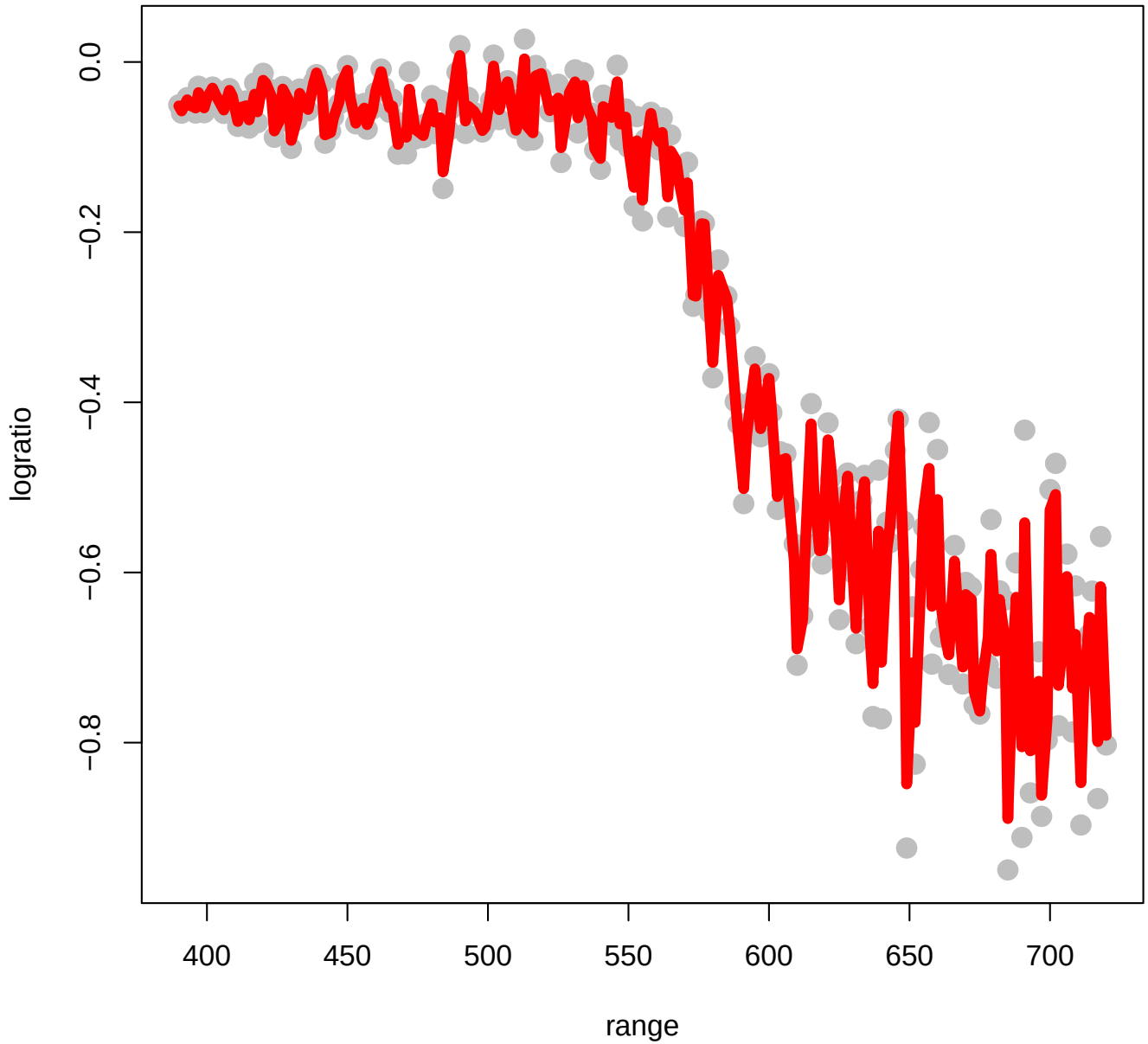
Smoothing splines

- Consider the following problem

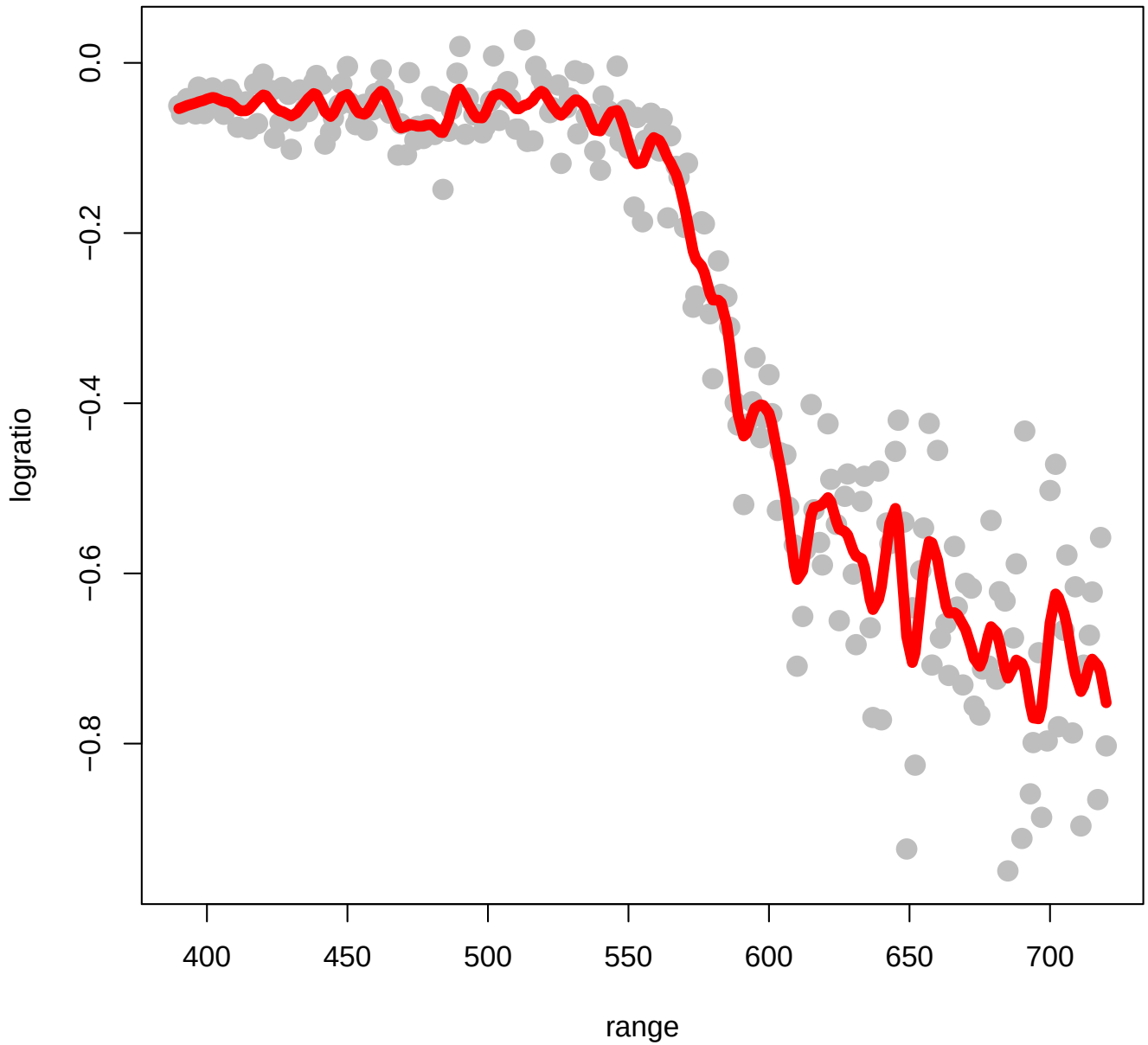
$$\min_f \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int \left(f^{(2)}(t)\right)^2 dt$$

- The solution is a *natural* cubic spline with n knots at $X_1, X_2, \ldots, X_n$.
- *Natural* cubic splines are cubic splines with the restriction that they are linear beyond the boundary knots.

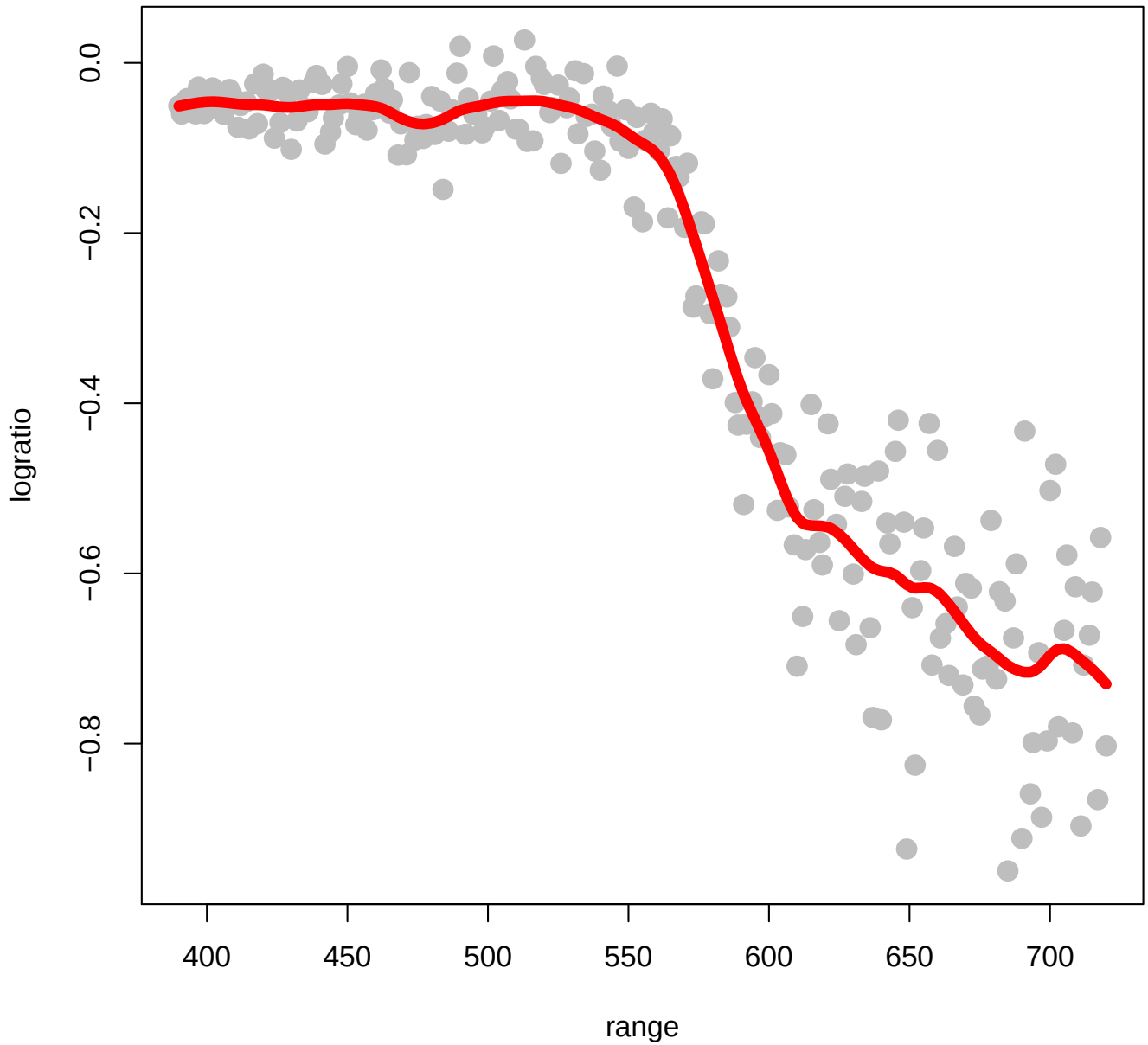
Smoothing spline, $\lambda = 0.20$



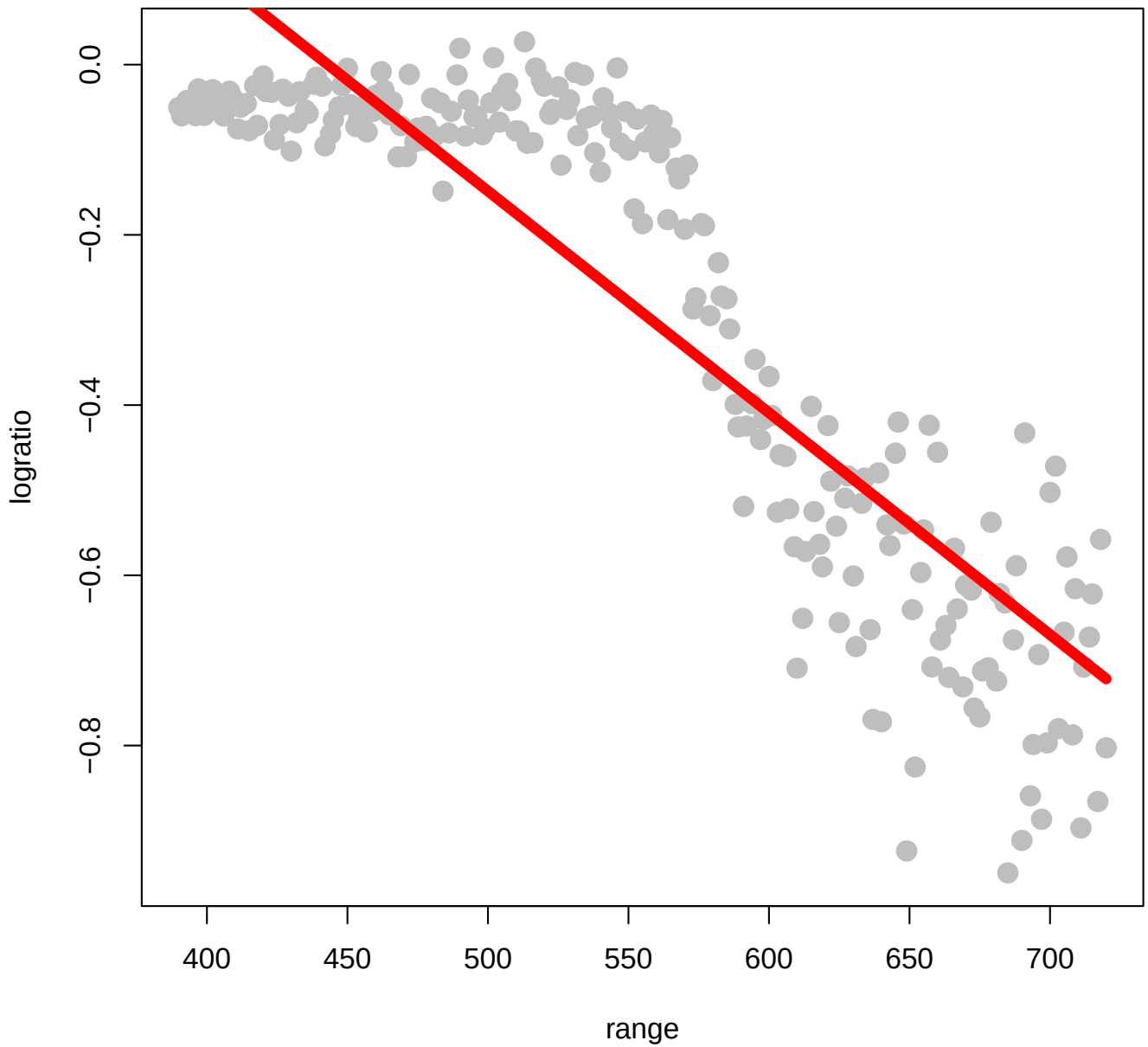
Smoothing spline, $\lambda = 0.50$



Smoothing spline, $\lambda = 0.75$



Smoothing spline, $\lambda = 2.00$



Selecting the penalty parameter

- How do we select λ?
- Minimizing

$$RSS(\lambda) = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' \beta_\lambda)^2$$

is not a good idea...

Selecting the penalty parameter

- Cross-validation: consider

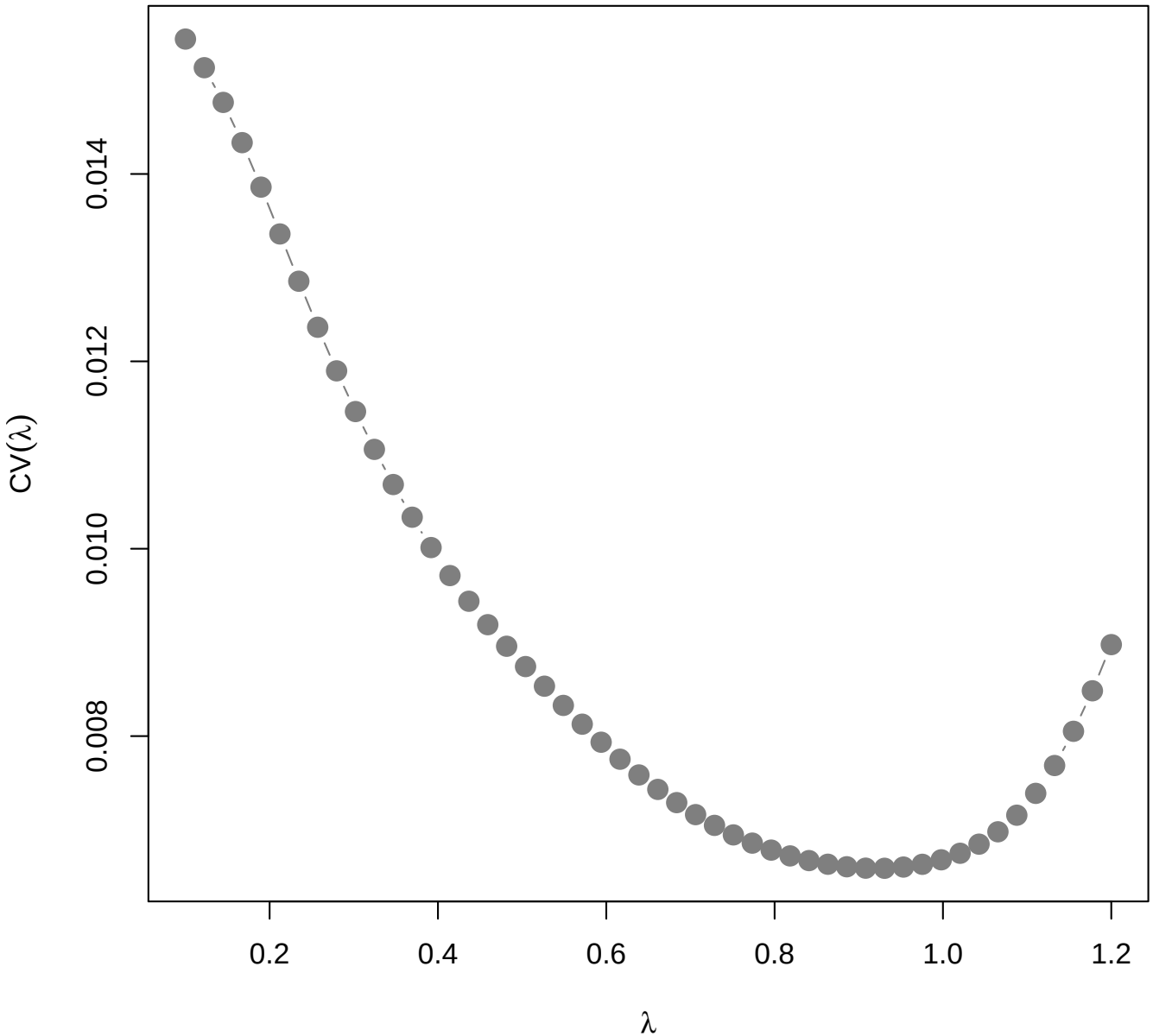
$$CV(\lambda) = \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' \beta_\lambda^{(-i)} \right)^2$$

where $\beta_\lambda^{(-i)}$ is the fit without using the point (Y_i, X_i)

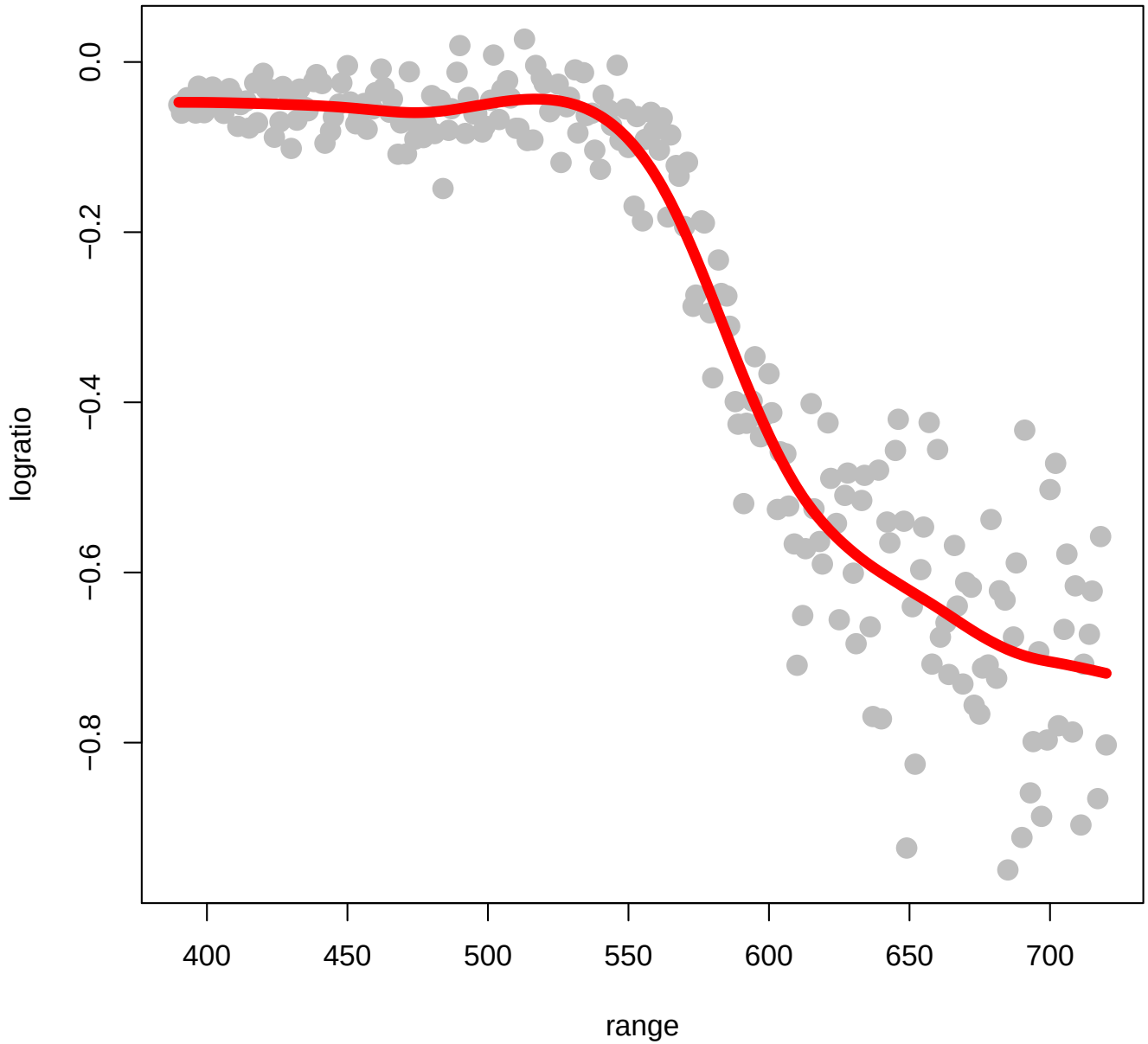
and choose a value λ_0 such that

$$CV(\lambda_0) \leq CV(\lambda) \quad \forall \ \lambda \geq 0$$

CV, smoothing spline



Optimal fit via leave-1-out CV



Selecting the penalty parameter

- Computing leave-one-out CV

$$CV(\lambda) = \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' \beta_\lambda^{(-i)} \right)^2$$

We might need to re-fit the model n times

Selecting the penalty parameter

- For some smoothers and models this is not necessary. For many linear smoothers $\hat{\mathbf{Y}} = \mathbf{S}_\lambda \mathbf{Y}$ we have

$$\hat{\mathbf{Y}}_r = \sum_{i=1}^{n} \mathbf{S}_{\lambda,r,i} Y_i \qquad r = 1, \ldots, n$$

and then

$$\hat{\mathbf{Y}}_r^{(-r)} = \frac{\sum_{i \neq r} \mathbf{S}_{\lambda,r,i} Y_i}{\sum_{i \neq r} \mathbf{S}_{\lambda,r,i}}$$

Selecting the penalty parameter

- Furthermore

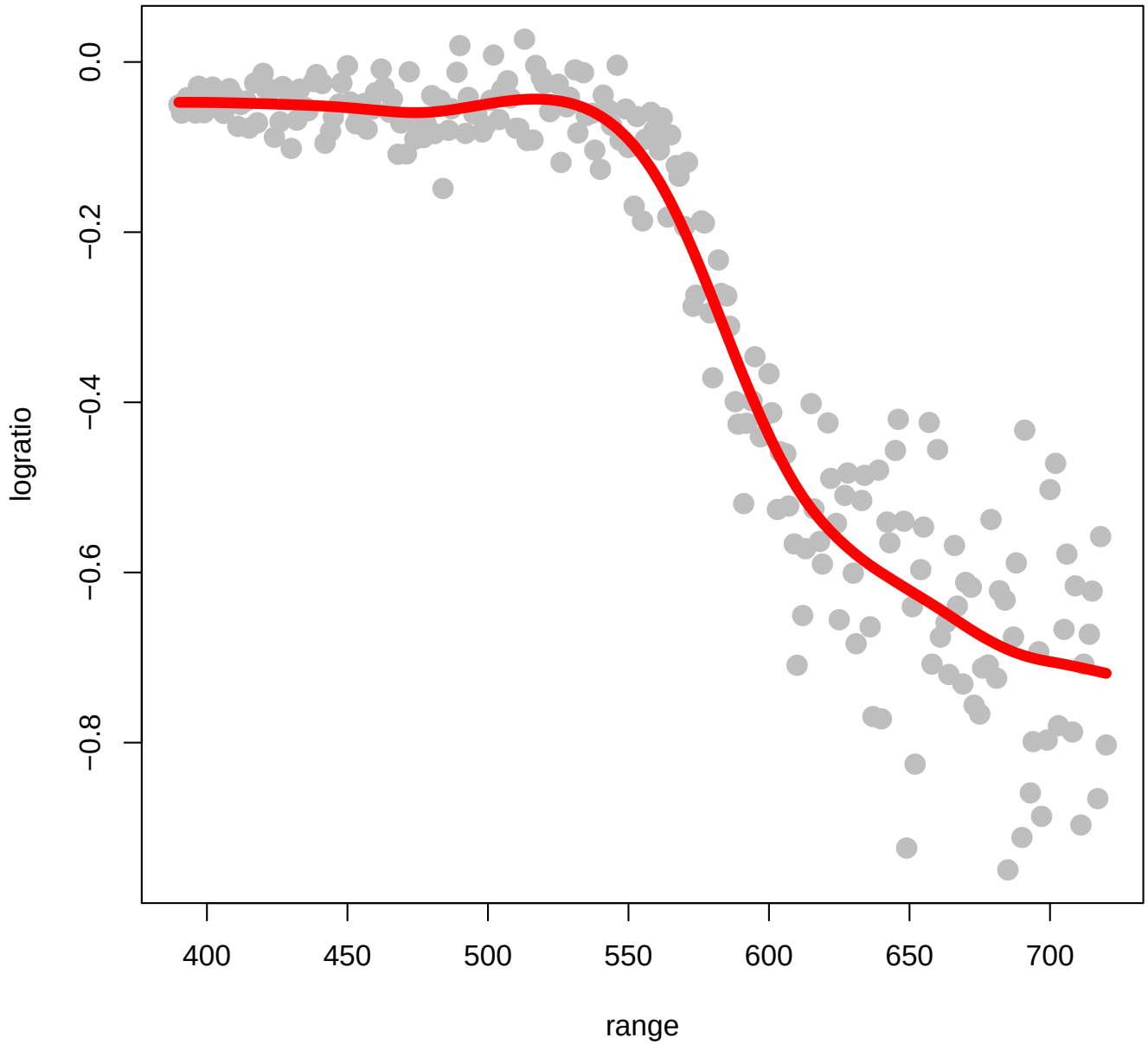
$$\mathbf{S}_\lambda \mathbf{1} = \mathbf{1}$$

thus

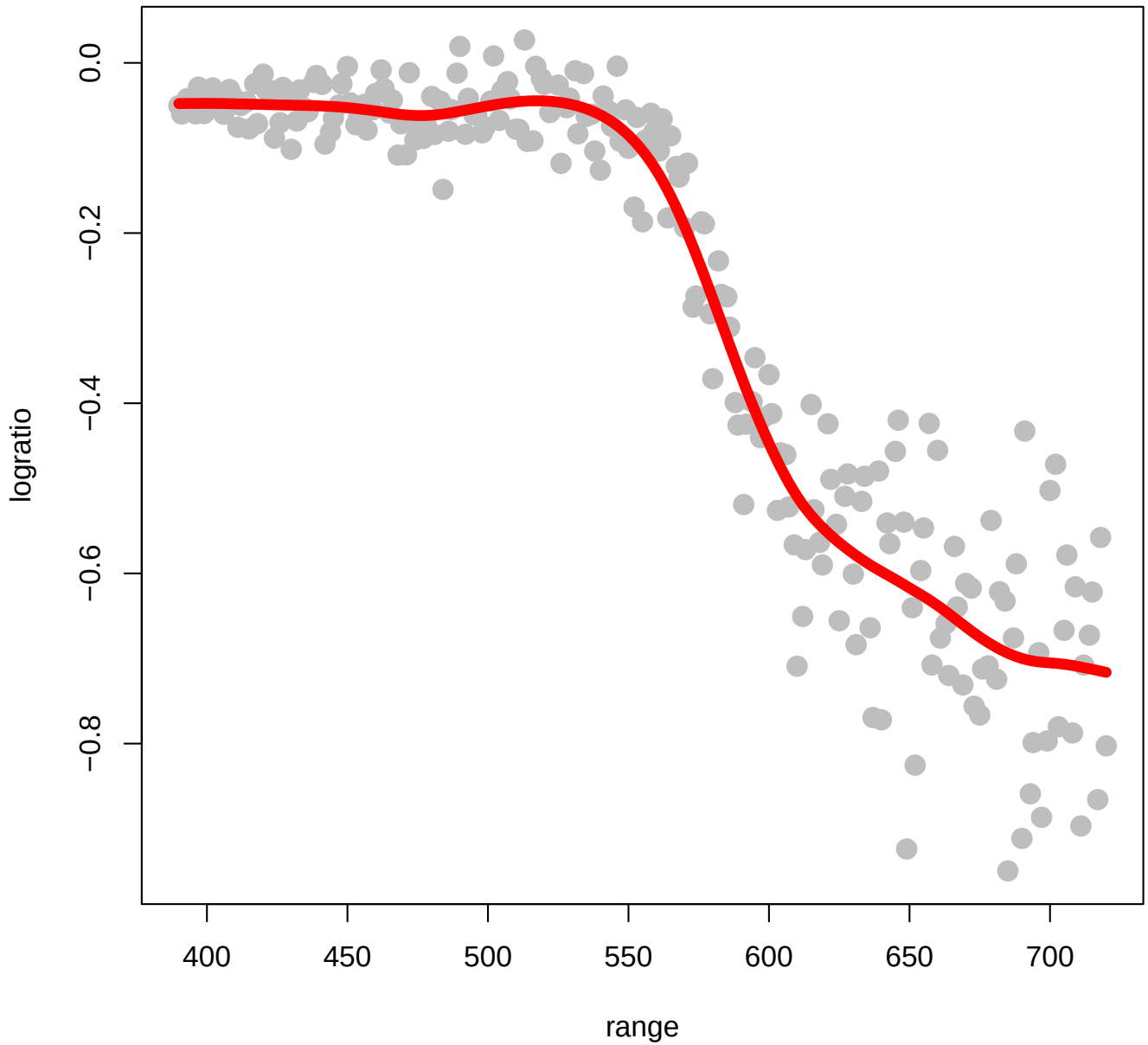
$$\hat{\mathbf{Y}}_r^{(-r)} = \frac{\sum_{i \neq r} \mathbf{S}_{\lambda,r,i} Y_i}{1 - \mathbf{S}_{\lambda,r,r}}$$

$$CV(\lambda) = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\mathbf{Y}}_i}{1 - \mathbf{S}_{\lambda,i,i}} \right)^2$$

Optimal fit via leave-1-out CV



Compare with 5-fold CV optimal



Selecting the penalty parameter

- Computing $\mathbf{S}_{\lambda,i,i}$, $i = 1, \ldots, n$ can be demanding

$$GCV(\lambda) = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\mathbf{Y}}_i}{1 - \text{tr}(\mathbf{S}_\lambda)/n} \right)^2 =$$

$$= \frac{\sum_{i=1}^{n} \left(Y_i - \hat{\mathbf{Y}}_i \right)^2}{(1 - \text{tr}(\mathbf{S}_\lambda)/n)^2}$$