Lecture 4: Cross-Validation and Quadratic and Cubic Splines

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. Review
2. Cross Validation
3. Quadratic Splines
4. Cubic Splines

Comparing Fits: Cross Validation (Last Slide of Last Class)

- Cross-validation: remove some observations from the data set and pretend that they are “new” ones.
- More formally: split the data into a training set and a test set
- Fit the model using the training set, use it to predict the test set
- Compare the quality of the different fits using their prediction errors on the test set

Cross-Validation (CV): Forming *v* Folds

- Split the data into v folds (or groups), e.g., $v = 2$ gives 2-fold CV

```
> v <- 2
> # So the (random) grouping can be reproduced
> # You should use a DIFFERENT seed
> set.seed(100)
> fold = sample(rep(1:v, length=nrow(lidar)))
> # Look at first few values of fold
> fold[1:10]
 [1] 1 1 1 1 2 1 1 2 1 1
```

- Observations in the first fold: $1, 2, 3, 4, 6, 7, 9, 10, \ldots$
- Observations in the second fold: $5, 8, \ldots$
- The groups are roughly equal-sized

```
> # Number of observations in each fold (group)
> sum(fold == 1)
[1] 111
> sum(fold == 2)
[1] 110
```

Cross-Validation

- Training data: all observations except those in fold i, denoted by $\mathcal{D}_{-i}$ ($\mathcal{D}$ for x, y "data")
- Test data: all observations in fold i, denoted by $\mathcal{D}_i$
- Fit the model using the training data $\mathcal{D}_{-i}$ only to give the predictive model $\hat{y}^{(-i)}(x)$
- Predict y for the test data $\mathcal{D}_i$
  - For each row index j of one observation in fold i
  - Error $= y_j - \hat{y}^{(-i)}(x_j)$
  - Observation j was used for testing but not for training here

v-Fold Cross-Validation

- Repeat for $i = 1, \ldots, v$
  - Every fold and hence every observation is once in the test set
  - Get errors $y_j - \hat{y}^{(-i)}(x_j)$ for every observation
  - Every observation is used for training and test but never at the same time
- CV root mean squared error

$$\text{CV RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}^{(-i)}(x_j)\right)^2}$$

Notation here: fold i is the fold containing observation j

Lidar Data: Linear Splines, 1 Knot Versus 5 Knots Revisited

- Comparing SS(Residual) is equivalent to comparing R^2
- MS(Residual) penalize models for the number of parameters but uses same data for training and testing

$$\text{MS(Residual)} = \frac{1}{n - \#\text{ parameters}} \sum_{j=1}^{n} \left(y_j - \hat{y}_{(}x_j)\right)^2$$

- CV here based on $v = 10$

# knots	R^2	CV RMSE
5	0.9242	0.0803
10	0.9247	0.0826

Clicker questions 1–4.

Cross-Validation: v?

- $v = 2$
  - Only needs 2 model fits: good if fitting is computationally expensive
  - But training data are only half the data set: poor estimate of error from model trained with all the data?

$v = n$ is n-fold or leave-one-out CV

  - Needs n model fits: bad if n is large and/or fitting is computationally expensive (can be avoided for `lm` fits using some math)
  - Uses $n - 1$ observations for fitting each time: good estimate of error from model trained with all the data

- For computationally challenging problems, $v = 10$ is a compromise

Quadratic Splines

- Consider a smoother set of basis functions

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

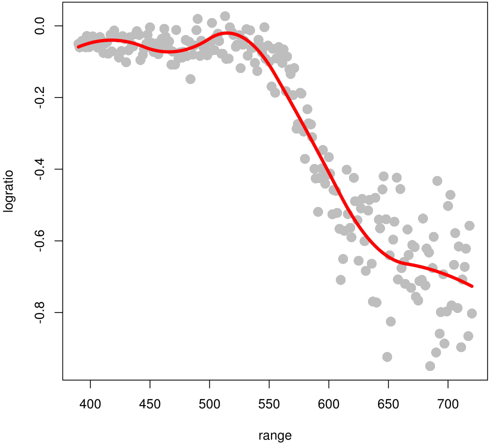
where κ_k, $1 \le k \le K$ are *knots* (to be chosen)

- Model

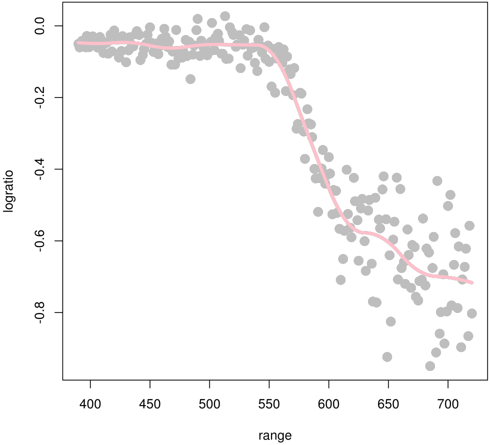
$$E\,[\,Y|\,x] \;=\; \beta_0 + \beta_1\,x + \beta_2\,x^2 + \sum_{k=1}^{K} \beta_{k+2}\,f_k(x)$$

- Is this still a linear model? **Clicker question 5.**

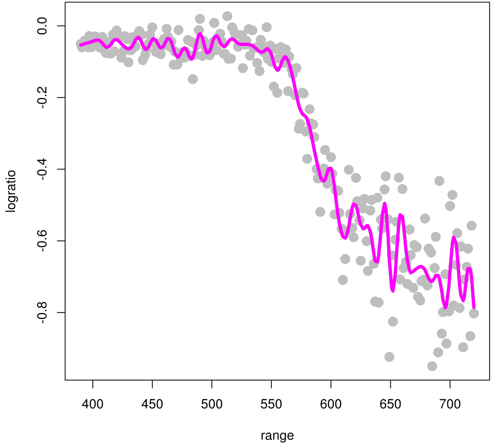
Lidar Data: Quadratic Splines, 5 Knots



Quadratic Splines, 10 knots

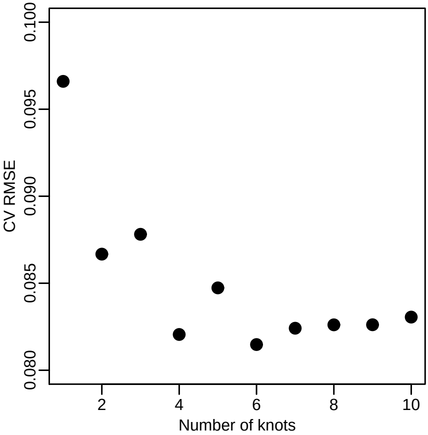


Lidar Data: Quadratic Splines, 50 knots



Lidar Data: Quadratic Splines, Cross Validation

Choose number of knots via 10-fold CV. **Clicker question 6.**



Cubic Splines

- Cubic splines are often used

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

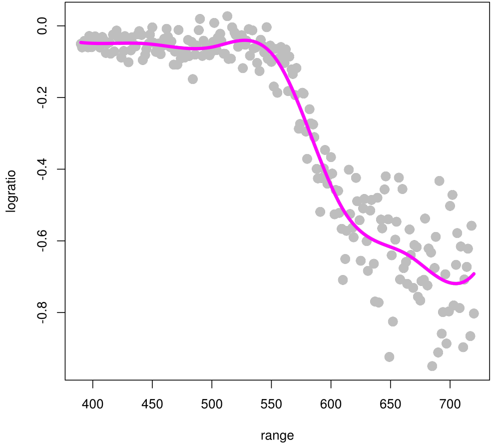
  where κ_k, $1 \leq k \leq K$ are *knots* (to be chosen)

- Model

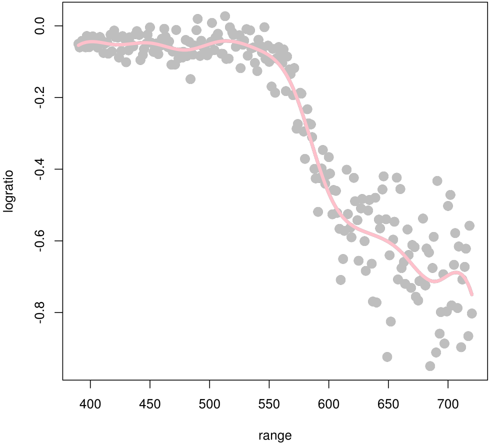
$$E\,[\,Y|\,x] = \beta_0 + \beta_1\,x + \beta_2\,x^2 + \beta_3\,x^3 + \sum_{k=1}^{K} \beta_{k+3}\,f_k(x)$$

- We can generalize it to higher orders as well, but...

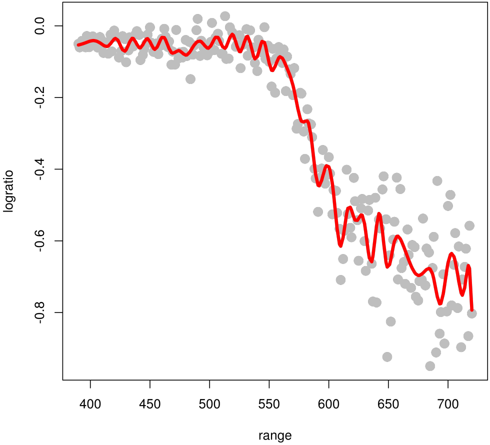
Cubic regression splines, 5 knots



Cubic regression splines, 10 knots



Cubic regression splines, 50 knots



Lidar Data: Cubic Splines, Cross Validation

Choose number of knots via 10-fold CV