The Practicalities of Scoring with Continuous Predictors

Ross Gayler

Motivation

- Interested in the craft/artisan aspects of credit scoring
- What the junior practitioner should know but might not be told and can't find
- Aspects of professional practice that are not published in the academic literature
  - Probably also true of statistical practice
- A collection of bits not worth publishing?

Importance of mundane innovation

- Small technical advances as tools
- Innovation mind-set:
  - Look at every situation as an opportunity to develop tools
  - Even if you don't develop a tool in every situation you will probably understand the situation better

Betty (New Caledonian Crow)

Tool user

Tool builder

Use of continuous predictors

- Statistics texts on regression primarily deal with continuous predictors as a single df

$$\hat{y} = w_1 x + w_0$$

- Scorecards generally partition continuous predictors into discrete indicators

$$\hat{y} = w_3 I(x > b_2) + w_2 I(b_2 \geq x > b_1) + w_1 I(b_1 \geq x_1) + w_0$$

Why treat continuous predictors as discrete?

- Allow nonlinear effects (commonly needed)
- Historical implementation constraints
  - Manual implementation (avoid multiplication)
  - Regularity of form (consistency with discrete predictors)
- Software & methodological inertia

Example of Application Scoring Table

Years on Job	Less than 6 Months **5**	Six Mos to 1 Yr 6 Mos **14**	1 Yr 7 Mo to 6 Yr 8 Mo. **20**	6 Yrs 9 Mo to 10 Yr 5 Mo. **27**	10 Yrs 6 Mos or More **39**	
Own or Rent	Own or Buying **40**	Rent **19**	All Other **26**			
Banking	Checking Account **22**	Savings Account **17**	Checking and Savings **31**	None **0**		
Major Credit Card	Yes **27**	No **11**				
Occupation	Retired **41**	Professional **36**	Clerical **27**	Sales **18**	Service **12**	All Other **27**
Age of Applicant	18 to 25 **19**	26 to 31 **14**	32 to 34 **22**	35 to 51 **26**	52 to 61 **34**	62 and Over **40**
Worst Credit Reference	Major Derogatory **-15**	Minor Derogatory **-4**	No Record **-2**	One Satisfactory **9**	Two or More Satis-factory **18**	No Investig. **0**

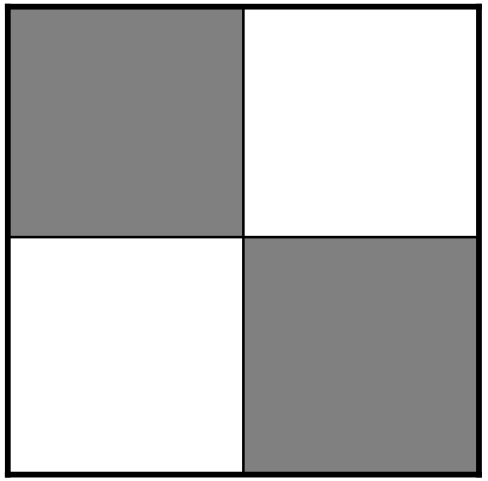
From E.M. Lewis (1992) “An introduction to credit scoring”

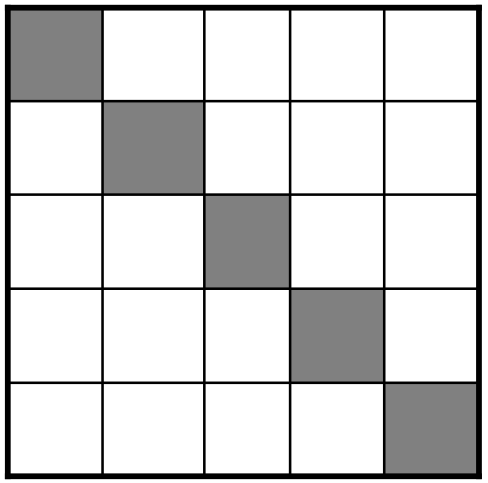
Advantages of continuous treatment

- (Slightly) increased predictive power
- Capture nonlinearity (if you can)
- Finer-grained score distributions
- Lower variance estimates of effects
- Better visual diagnostics
- Postponement of effort

Predictive power

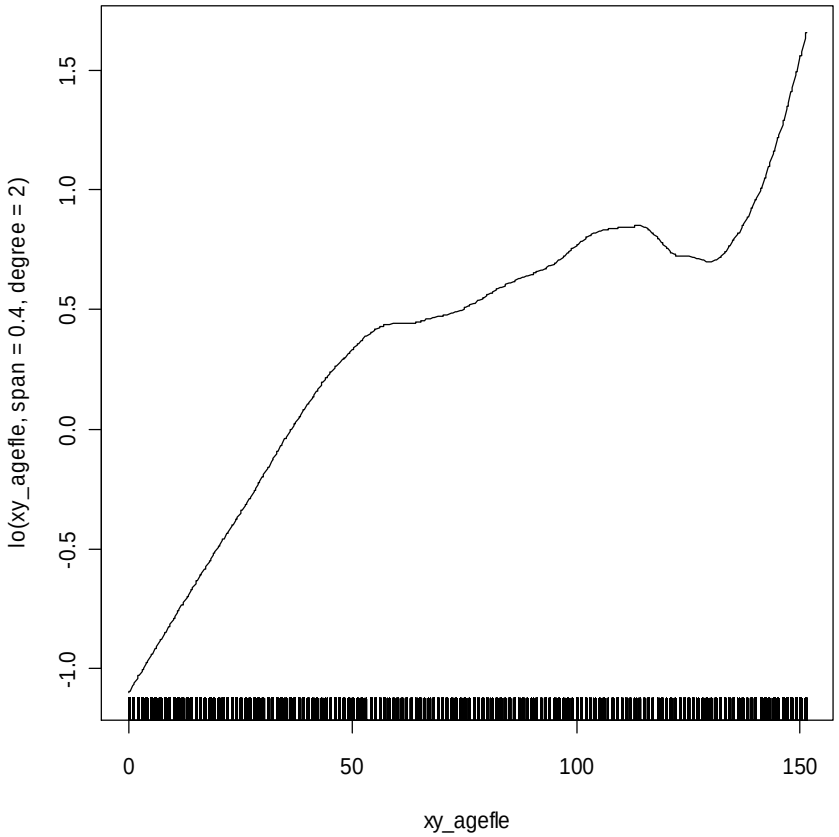
- Categorisation loses information
- How much is lost?
- Gini/AUC based on pairwise comparisons
- Diagonal blocks lose information
- Proportion of tied pairs scales as 1/k



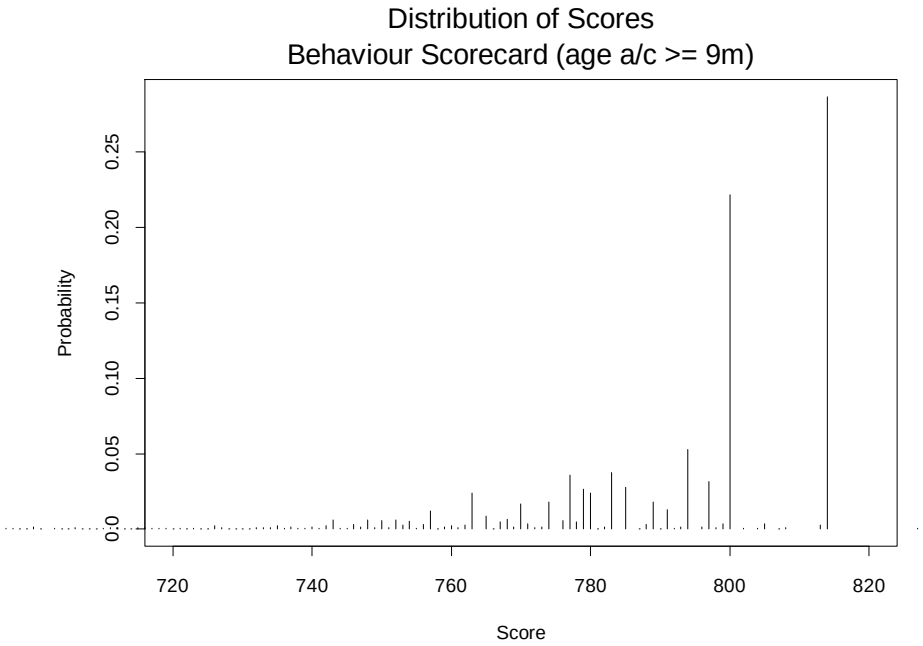


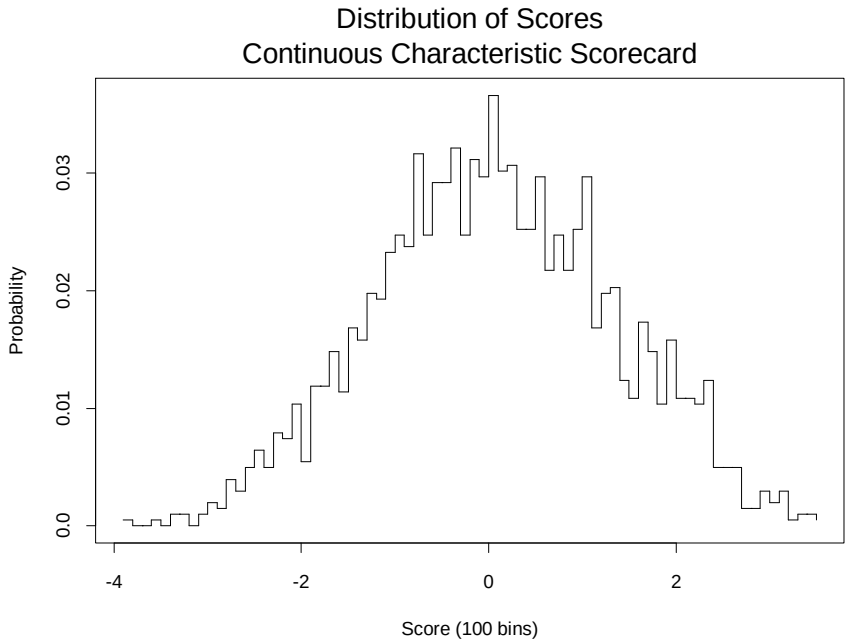
Capture nonlinearities

- Nonlinear effects exist
- The trick is to model them



Granularity of score distributions

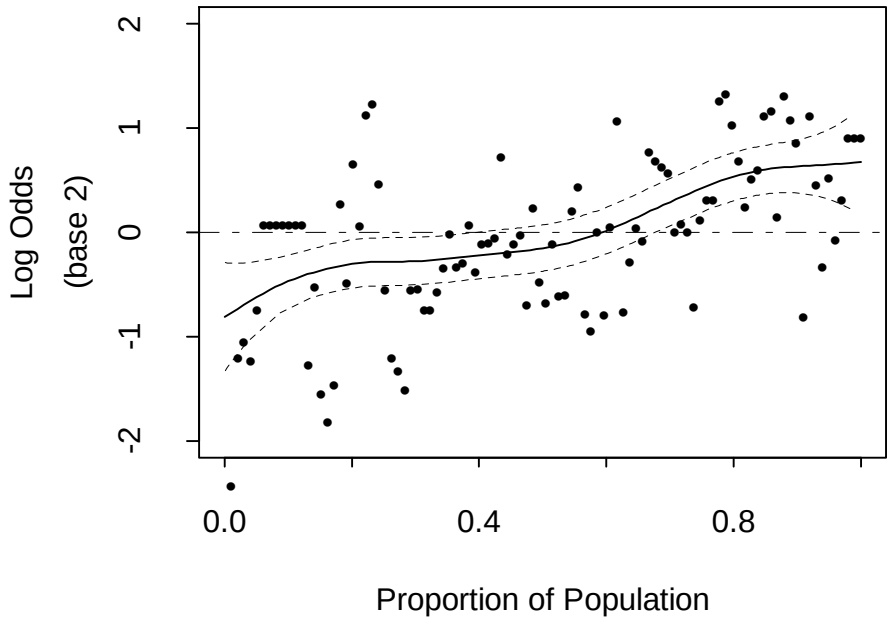




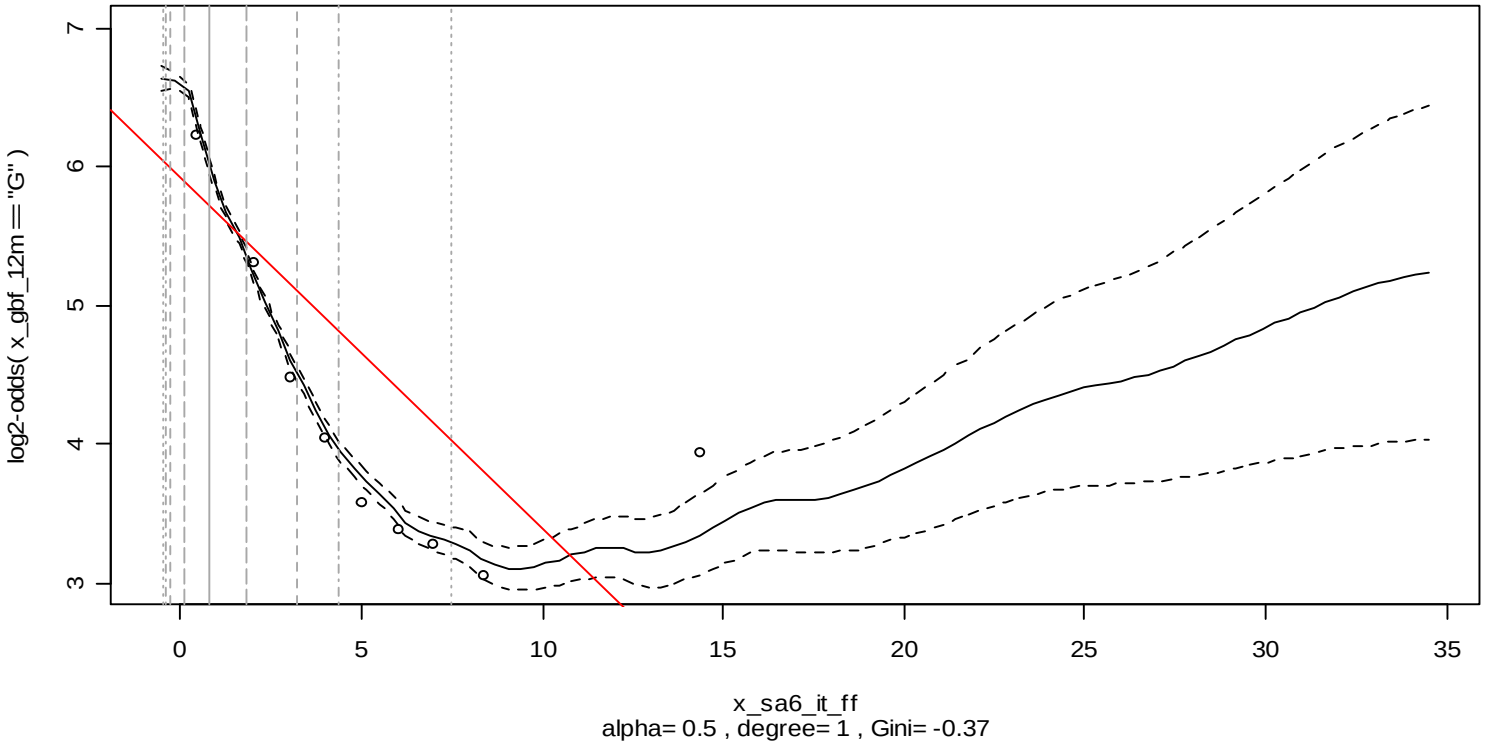
- More fine-grained scores are easier to use

Lower variance estimates

- Standard error of the outcome estimate
- Discrete estimates each category in isolation
- Continuous “borrows strength” from its neighbours

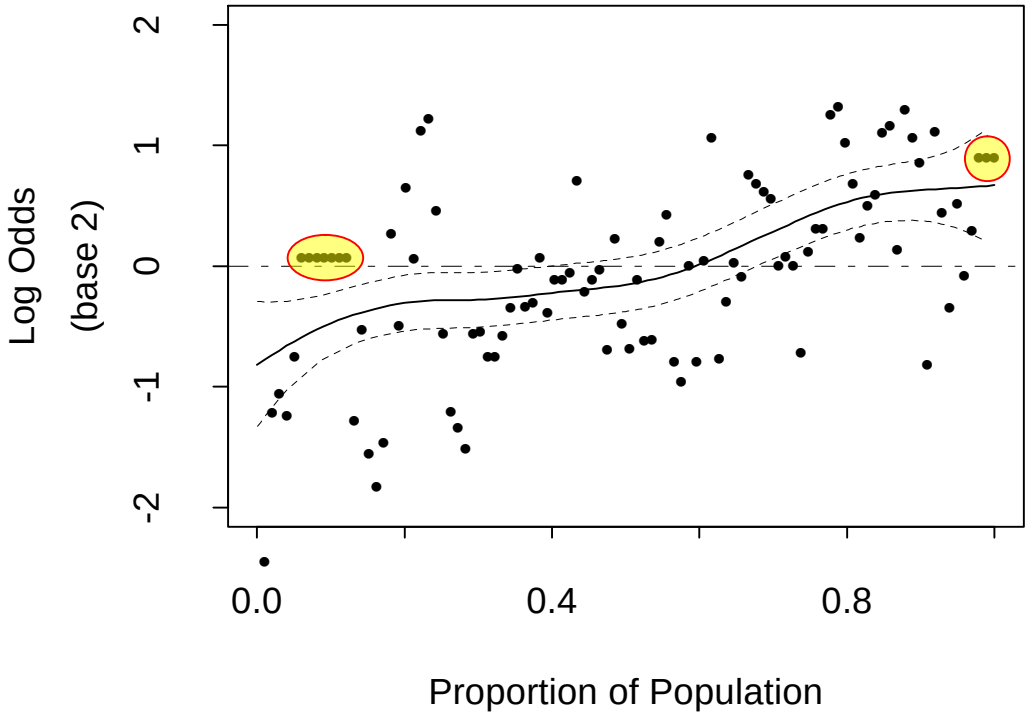


Visual diagnostics



- Subtle effects may be invisible in a discrete analysis unless the boundaries are well placed
  - Slope changes sign above 99th percentile

Visual diagnostics



- Engages the capabilities of visual perception
  - Small outlier groups (at 5th & 97th percentiles) would be lost in the noise without the smoothed relationship

Postponement of effort

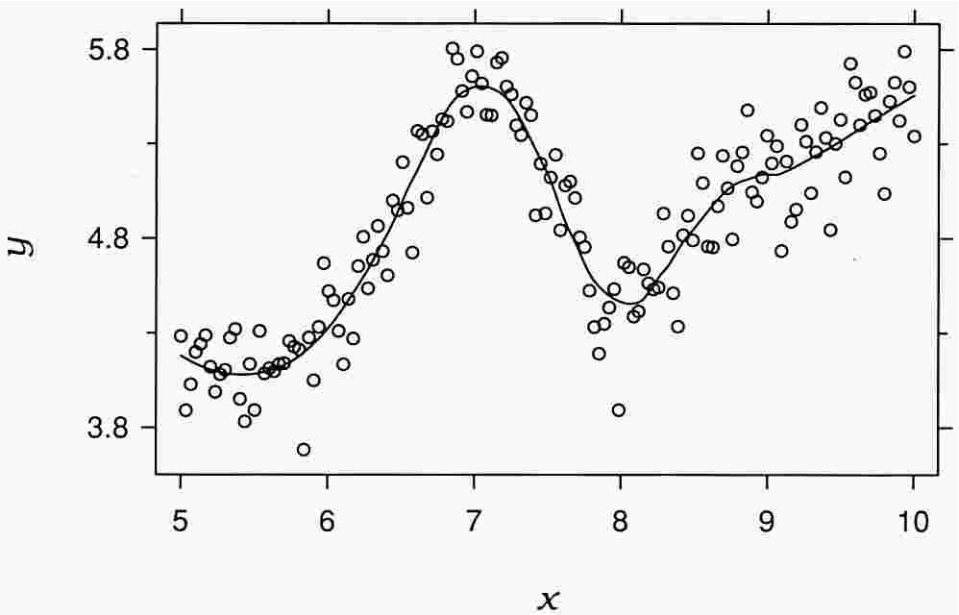
- Avoid the time taken to do fine and coarse classing
- Treat variables as continuous during exploration
- Only convert to discrete form if:
  - The variable will be used as a predictor AND
  - The implementation requires discrete form

Characteristic analysis

- Fit an arbitrary smooth function from the predictor to the outcome

$$\hat{y} = f(x)$$

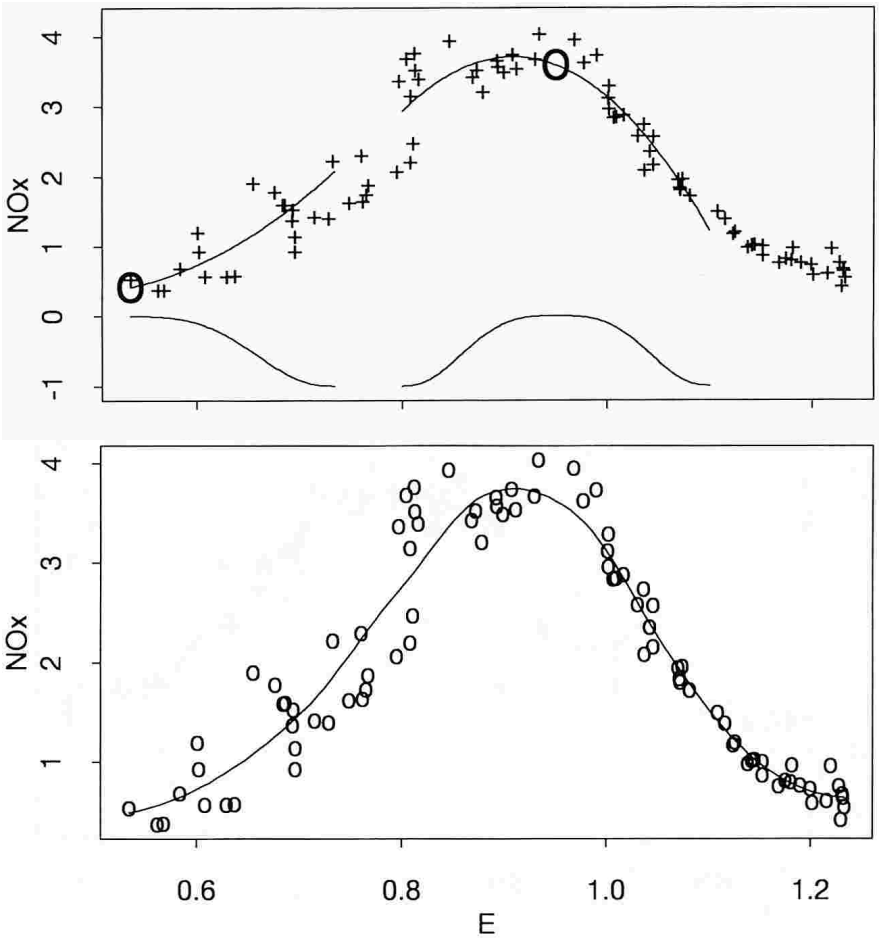
- Use a scatterplot smoother
  - (The example shows a continuous outcome)



From W.S. Cleveland (1993) “Visualizing data”

Scatterplot smoothing

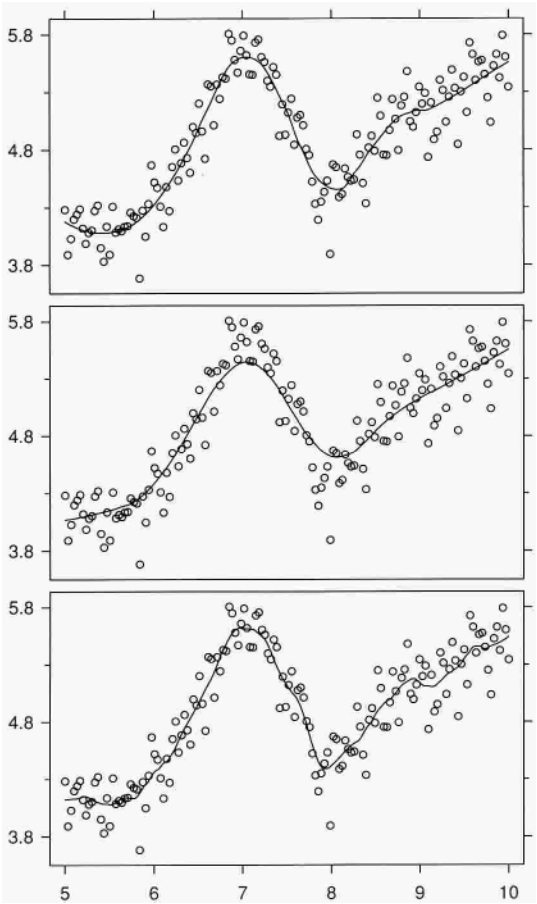
- Generalisation from moving averages to moving regressions
- Local low-order polynomial
- Can be further generalised to be a local GLM
  - Local logistic regression



From C. Loader (1992) “Local regression and likelihood”

Understand how to use it

- Smoothing parameters
- Treatment of discontinuities
- Boundary cases
- Skewed predictors



From C. Loader (1992) “Local regression and likelihood”

Outcome scaling: WOE

- Local logistic regression predicts log-odds
- WOE = log-odds – population log-odds
- Use smoothing function as WOE transform

$$\hat{y} = \cdots + w_i f_i(x_i) + \cdots + w_0$$

- May want to modify the smoothing function before using it as a WOE function
  - e.g. points nondecreasing with age
  - Parameterisation should allow tweaking

Outcome scaling: Score

- Score usually a linear scaling of log-odds
- Rescale outcome axis to score units
  - Comparability of CA with scorecard points
  - Plot smoothed score on the CA to show how well the scorecard approximates the effect
- If score is scaled to log-odds
  - Offset the score to show the partial effect

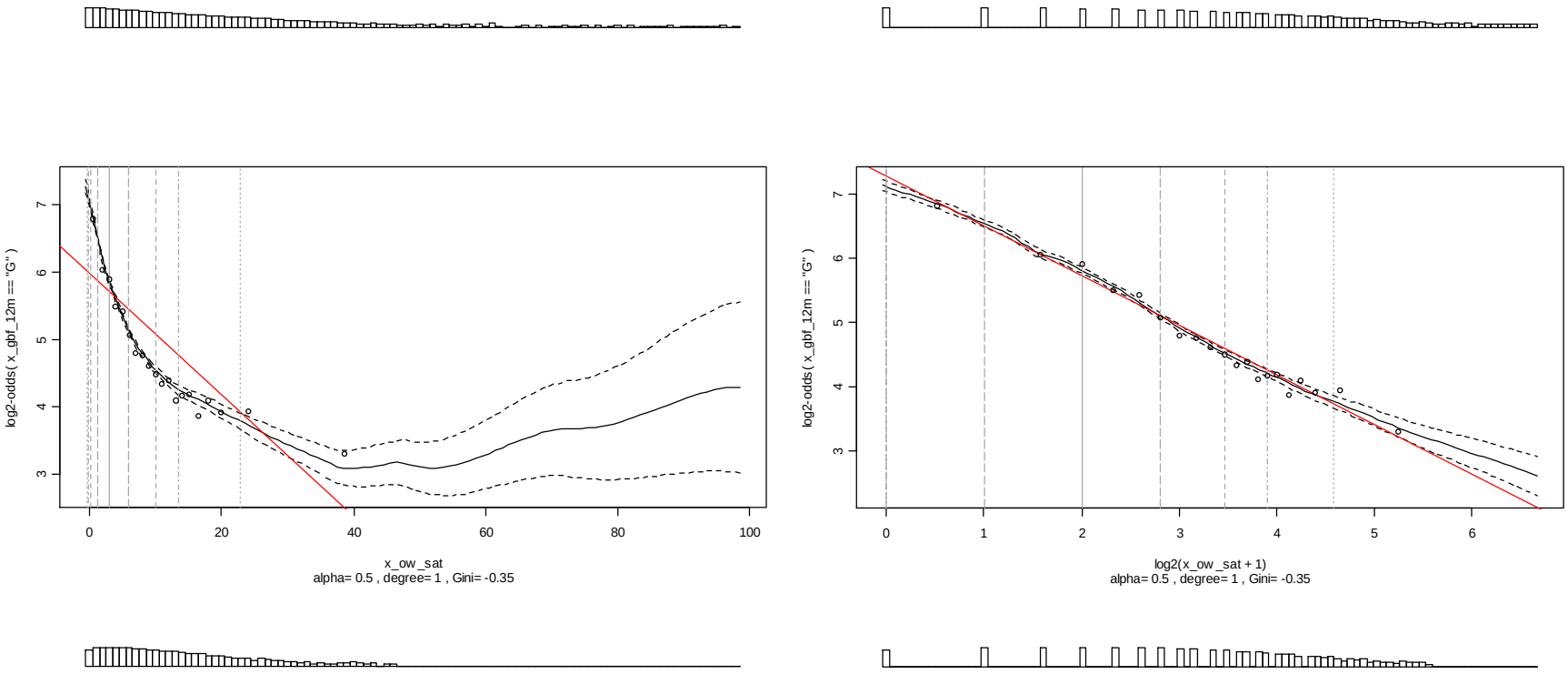
$$\hat{y} = f(x) + offset(score)$$

$$\hat{y} - offset(score) = f(x)$$

Transforming the predictor

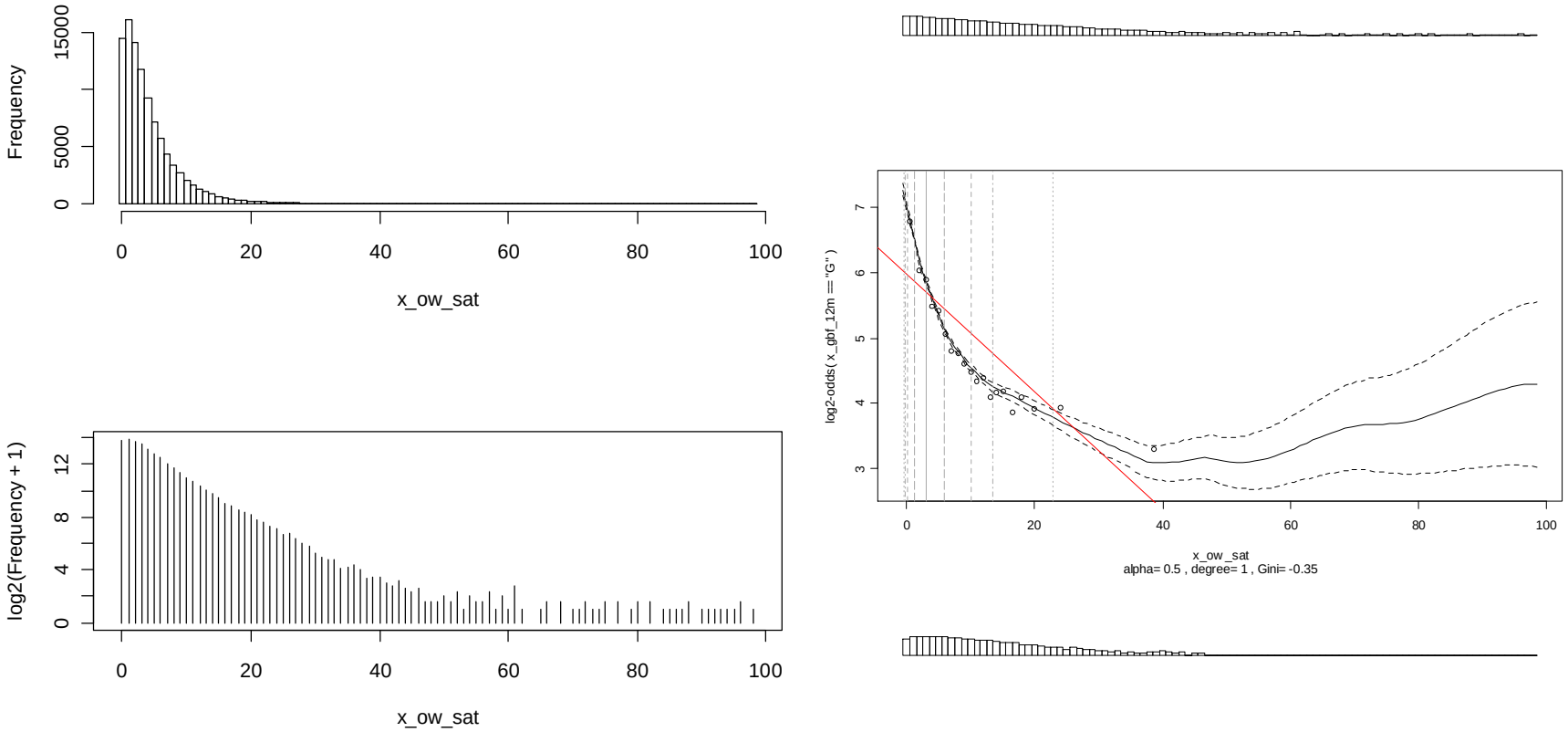
- Apply the smoothing transform as a check
  - Should linearise the CA
- Apply an analytical transform (e.g. log)
  - May linearise the effect
  - May improve very skewed predictors (smooths may be dominated by long tails)
- Transform to percentiles
  - Makes skewed predictors easier to see
  - Able to interpret population fractions
  - Integral of |WOE| indicates predictive power
  - Consider smoothing a discontinuous ECDF

Linearisation



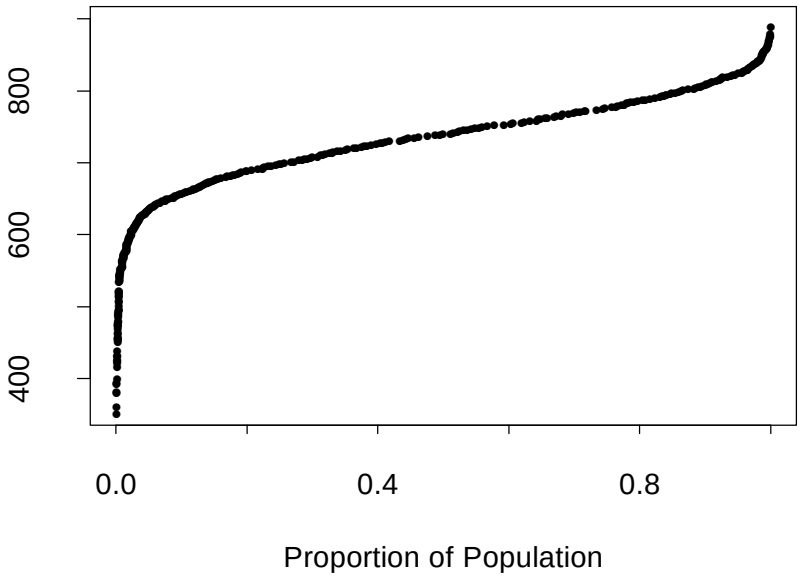
Take the log of the predictor

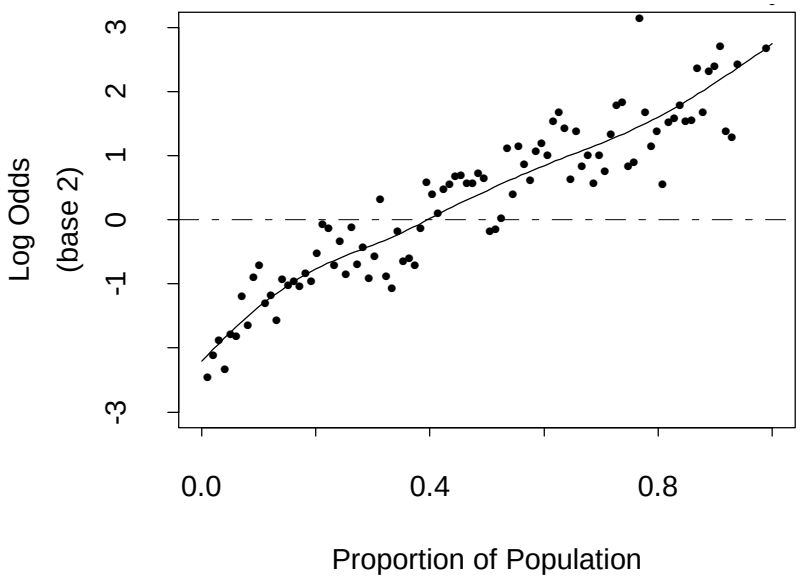
A dominating tail



Leverage in the fitting window

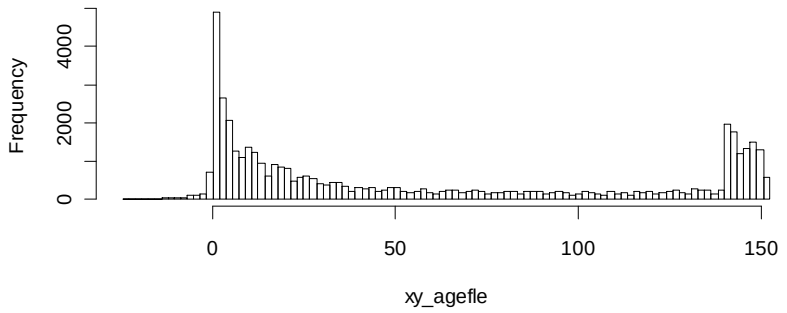
Transform predictor to percentiles

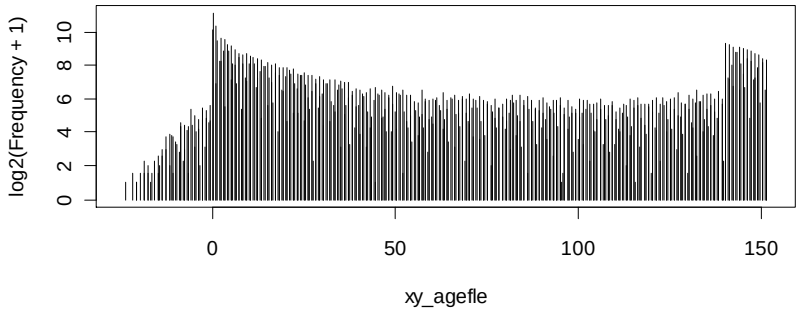




Visual diagnostics (again)

- Treat everything as continuous (avoid binning)
- Converting histograms to spike plots may show odd patterns
- E.g. Age coded as months.days (5.31)





Implementation practicalities

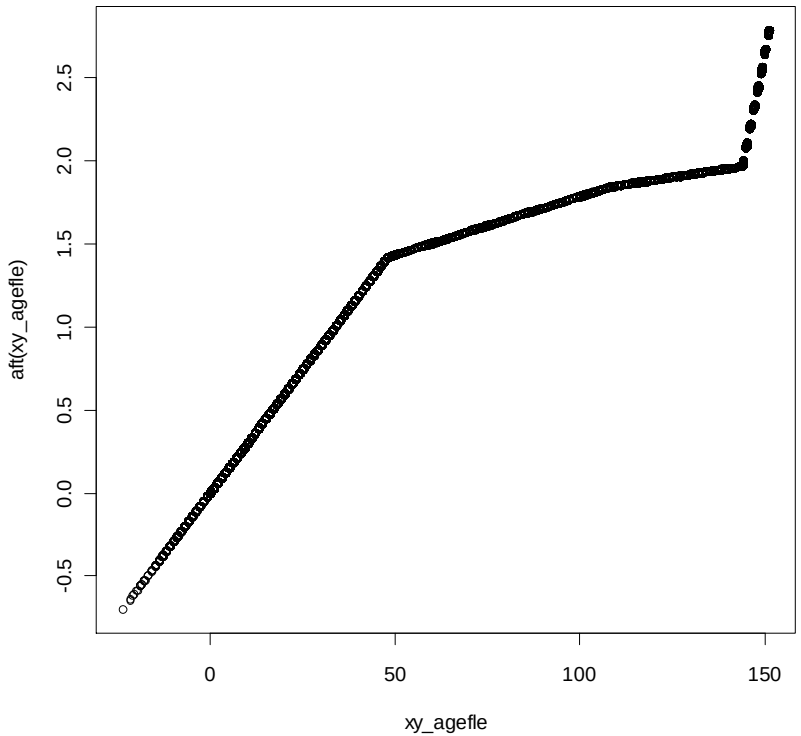
- Think of smoothing function as a look-up table
- Conversion to discrete
  - Arbitrarily partition the smoothing function and calculate the mean value in each category
- Conversion to piecewise linear

$$\hat{y} = w_0 + w_1 I(x < b_1)x + w_2 I(x \geq b_1)(x - b_1)$$

- Direct use (linear or analytical functions)
  - Bound open-ended predictors
- Don’t use polynomial unless bounded

Piece-wise linear effect

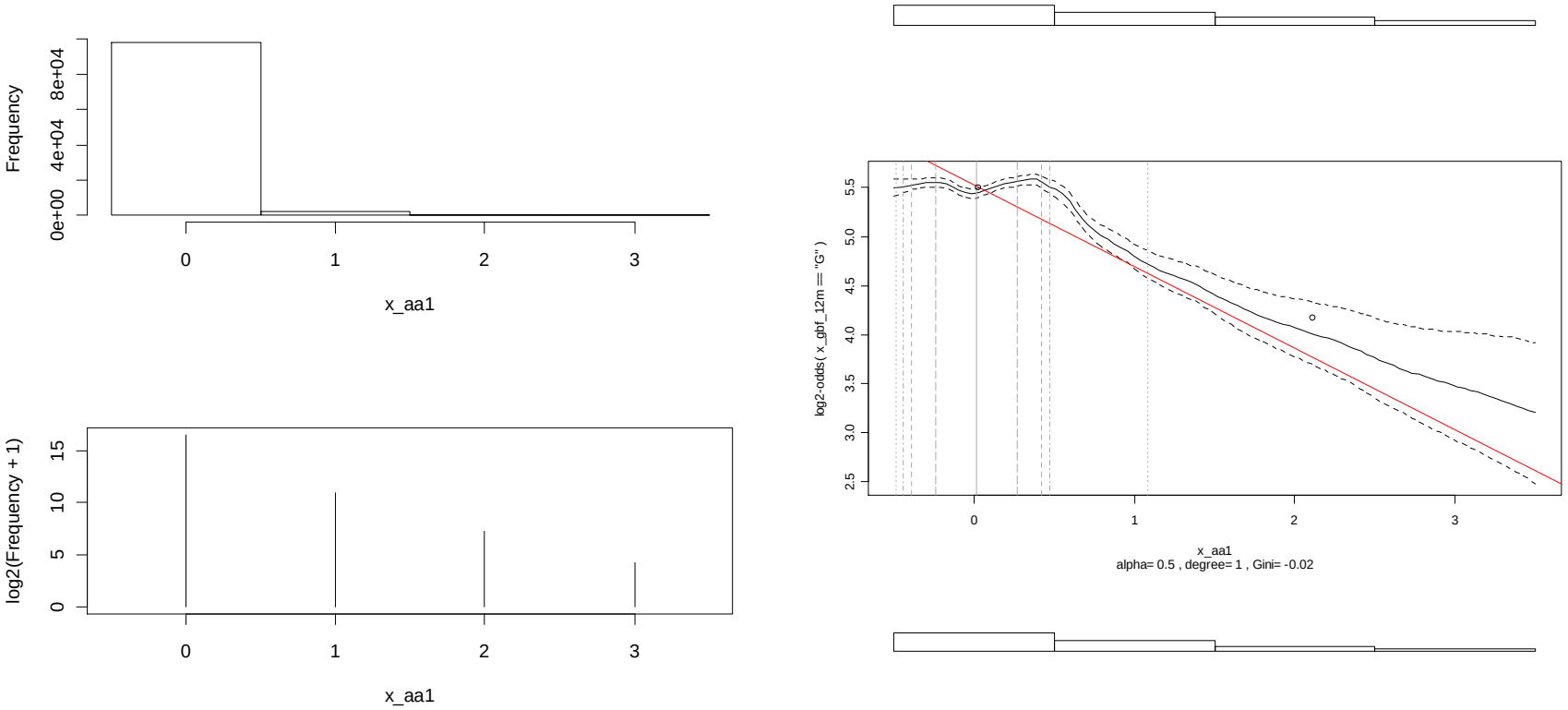
```
aft <- function(x)
{
    a0 <- 0.03 * x
    a48 <- -0.02 * (x >= 48) *
          (x - 48)
    a108 <- -0.004 * (x >=
          108) * (x - 108)
    a144 <- 0.1 * (x >= 144) *
          (x - 144)
    a0 + a48 + a108 + a144
}
```



Handling the tricky bits

- Many scoring predictors are semi-discrete
  - Exceeding the smoothing window
  - Jittering
- Handling special values
  - Flags & Discontinuous values
  - Indicators for special values
    - Flags are typically exclusive
    - Discontinuous values may be parameterised exclusively or additively depending on “conceptual continuity”

Jittering



95% of observations on one value