Fitting curves using non-parametric approaches

Keith R. Shockley, NIEHS

October 24, 2017
OUTLINE:

Parametric modeling

Non-parametric modeling

Estimating point of departure from fitted curves
Parametric modeling
Parametric models

- Pre-specified model form
  - Linear model: $f(x) = mx + b$
  - Hill model: $f(x) = f_0 + f_{\text{max}} * x^h / (AC_{50}^h + x^h))$
  - Cubic polynomial: $f(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0$

- Contain parameters, some of which might be useful
  - Slope and y-intercept
  - $AC_{50}, f_{\text{max}}, f_0, h$
  - $a_3, a_2, a_1, a_0$
Parametric models (pros and cons)

Pros
- Reduce unknown (and possibly complicated) function $f(x)$ to a simple form with few parameters
- Can produce consistent results when the curve fits the data well
- May have familiar and useful parameters

Cons
- A pre-specified parametric model may not fit the data well
- Carry distributional assumptions (e.g., Normality)
- Different parametric models may produce different BMD estimates, reflecting model uncertainty
- Model averaging can be helpful when true function is not on edge of model averaging space
Hill model

Hill equation

\[ R = R_0 + \frac{R_{\text{max}} C^h}{AC_{50}^h + C^h} \]

Response (\( R_i \))

Log Concentration (\( C_i \))

Data Set 1

Data Set 2
Non-parametric modeling
Non-parametric models

- Flexible model form
  - Interpolation
  - LOESS (nonparametric local regression)
  - Splines (continuous piece-wise polynomials between knots)

- Parameters may not be readily interpretable
  - Interpolation – estimates values that lie between data points
  - LOESS – fits segments of the data at each point in the range of the data set by calculating many polynomial coefficients using weighted least squares, but doesn’t estimate a single “coefficient” for a global model
  - Splines – calculates many “polynomial coefficients” between intervals of data
Non-parametric models (pros and cons)

Pros
- Makes fewer assumptions about $f(x)$
- Uses the data to learn about the potential shape of $f(x)$
- Should fit the data very well

Cons
- Parameters may not be readily interpretable
- Carry distributional assumptions too
- May be computationally intensive
- May not be as familiar as parametric approaches
**B-spline**

A basis spline (B-spline) is a piecewise polynomial function, where the pieces meet at the knots. Any spline can be expressed as a linear combination of B-splines.

For each interval $[x_i, x_{i+1}]$:

$$B(x) = a_3x^3 + a_2x^2 + a_1x + a_0$$
Estimating point of departure from fitted curves
Curve Fitting and Potency Estimation (Case 1)

- Case 1: Hill equation model and $AC_{10}$ parameter

1. Fit the Hill model to the data.
2. $AC_{10}$ is the point of departure.
Case 1: Hill equation model and AC\textsubscript{10} parameter

Case 2: B-spline and concentration that curve crosses a response threshold

Curve Fitting and Potency Estimation (Case 2)

1. Fit a B-spline to the data.
2. The point of departure is the concentration at which the curve crosses the detection band.
Curve Fitting and Potency Estimation (Case 3)

- Case 1: Hill equation model and $AC_{10}$ parameter
- Case 2: B-spline and concentration that curve crosses a response threshold
- Case 3: Polynomial interpolation and entropy-based point of departure

1. Fit an interpolation curve to the data.
2. Calculate a “weighted entropy” along the curve (Shockley, 2014).
3. The $POD$ is the concentration at which the change in entropy is maximal (Shockley, 2016).
Simulation Study

<table>
<thead>
<tr>
<th>Data Set Parameters</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $</td>
<td>\text{RMAX}</td>
</tr>
<tr>
<td>5% error (15% Detection Limit)</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.0001</td>
</tr>
<tr>
<td>25</td>
<td>0.01</td>
</tr>
<tr>
<td>25</td>
<td>1.1</td>
</tr>
<tr>
<td>50</td>
<td>0.0001</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>1.1</td>
</tr>
<tr>
<td>100</td>
<td>0.0001</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>100</td>
<td>1.1</td>
</tr>
<tr>
<td>10% error (30% Detection Limit)</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.0001</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>1.1</td>
</tr>
<tr>
<td>100</td>
<td>0.0001</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
</tr>
<tr>
<td>100</td>
<td>1.1</td>
</tr>
</tbody>
</table>

$\text{Log}_{10}$Precision ($\text{log}_{10}$Bias) in 15-point concentration response data simulated from 10,000 Hill model curves, with $R_0 = 0$ and $h = 1$. Adapted from Shockley (2016).
Repeatability of Potency Estimates
(Tox21 Phase II BG1 estrogen receptor agonist)

adapted from Shockley, 2016
Examples: Uncertainty in $POD_{WES}$ (bootstrapping) (Tox21 Phase II BG1 estrogen receptor agonist)

Oxymetholone: CASRN 434-07-1
Examples: Uncertainty in $POD_{WES}$ (bootstrapping)
(Dunnick et al., *Arch. Toxicol.*, 2017)

$Ptgr1$: Affy Probe ID 1388102_at
Summary

- Parametric modeling requires pre-specifying the model, but is more familiar and may have interpretable parameters.

- Nonparametric modeling is more flexible, but may be less familiar and may not have readily interpretable parameters.

- Simulation studies and repeatability of experimental results can be used to evaluate the performance of proposed modeling approaches.