Calibrating Natural History Models

Ben Youngman

University of Sheffield, UK
http://b-youngman.postgrad.shef.ac.uk/
b.youngman@sheffield.ac.uk

MUCM Team Meeting
21 September 2011
Durham University

This talk

- Essence of a Natural History Model (NHM)
- Refining the NHM's input region
- Emulation

Natural History Models

- Focusing on a model for Bowel Cancer
- Developed by School of Health and Related Research (ScHaRR) at Sheffield
- The NHM's concept
  - patient-level model
  - person has an underlying cancer, pre- or non-cancer state
  - person progresses through states in order of severity
  - at any point person may present (eg. visit doctor), progress to next state, or die
  - if they present, enter the Health System
- NHM offers thorough description of how patients move through Health system—eg. incurred costs can be tracked

Calibrating NHMs

- Parameters
  - control inputs—up to 100 (?)—eg. probability of being allocated to a given treatment
  - variable inputs—25—eg. rate of progression between states 2 and 3
- Cannot calibrate against underlying states as in general unknown
- Target data
  - cancer cases by age, cases by type
  - comparable model output from NHM available

Refining the input region

- Allows emulator to be built
- Bounds provided by ScHaRR, but model can give wildly different output from target data
    - eg. progression times assumed Weibull, but scale and shape parameters not orthogonal
- Use posterior predictive density of target data given model parameters to identify plausible region

Refining the input region

Cases by age

- Age group j

$$Y_{obs,age,j} \sim Bi(n_{obs,age,j}, \theta_{age,j})$$

- $f_{age}()$ assumed to determine $\theta_{age,j}$
- For run i, $i = 1, \ldots, n$, inputs x_i,

$$Y^{(i)}_{mod,age,j} \sim Bi(n^{(i)}_{mod,age,j}, \theta^{(i)}_{age,j} = f_{age}(x_i))$$

- Want to maximise $pr(Y_{obs,age,j} = y_{obs,age,j} \mid x_i)$
- Use $U[0, 1]$ priors for all θs
- Repeating for all j calibrates total number of cases and distribution of cases by age

Refining the input region

Cases by type

- Types Duke's A, B and C and Stage D

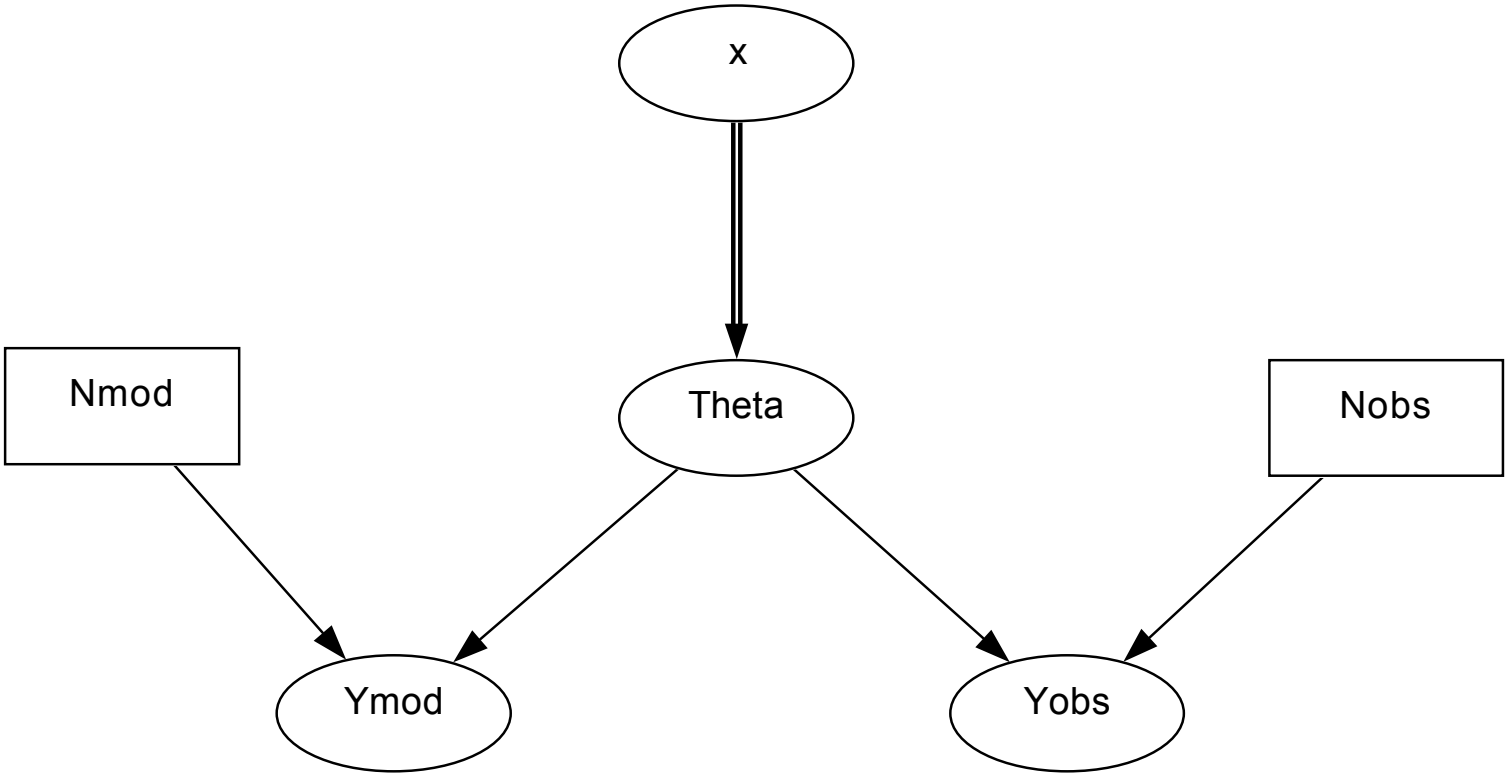
$$Y_{obs,type} \sim Mn(n_{obs,type}, \theta_{type})$$

- For run i, $i = 1, \ldots, n$, inputs x_i,

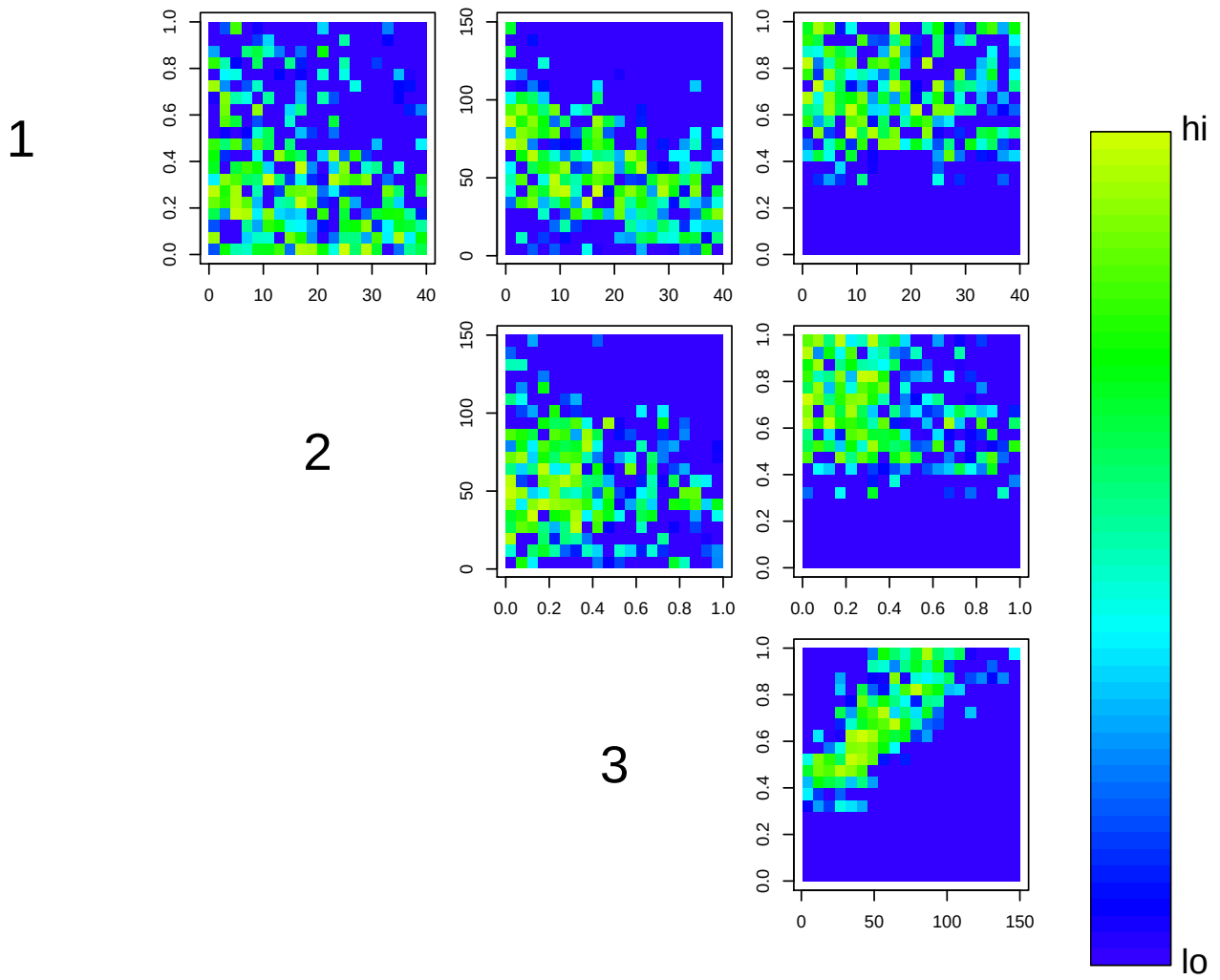
$$Y^{(i)}_{mod,type} \sim Bi(n^{(i)}_{mod,type}, \theta^{(i)}_{type} = f_{type}(x_i))$$

- Also maximise $pr(Y_{obs,type} = y_{obs,type} \mid x_i)$
- Only calibrates distribution of cases by type: total number cases already dealt with
- A couple of other data that we calibrate against similarly

Input plausibility



Input plausibility



25

Refining the input region

Technicalities

- Model begins with n_p patients of age zero then tracks age and movements
  - so to get number cases in five-year age band, use

$$\sum_{\text{age band}} \frac{\#\text{cases for each year}}{\#\text{people alive in that year}}$$

    ie. count people on multiple occasions
- Model run time approx. $\propto n_p$
  - so use three sets of $10k$ runs with $n_p = 1k$, $n_p = 10k$ and $n_p = 100k$ to refine region in three stages
- Posterior predictive ratios used to determine plausible region

Emulator

Specification

- Assume

$$Y_{age,j} \,|\, x \sim Bi\big(n_{age,j}(x), p_{age,j}(x)\big)$$

  and let $p_{age}(x) = \big(p_{age,1}(x), p_{age,2}(x), \dots\big)$

- Assume

$$Y_{type} \,|\, x \sim Mn\big(n_{type}(x), p^*_{type}(x)\big)$$

  and let $p^*_{type}(x) = \big(p_{type}(x), 1 - \sum_{type} p_{type}(x)\big)$

- Let $p() = \big(p_{age}(), p_{type}(), \dots\big)^T$

- Assume

$$\log\left\{\frac{p()}{1-p()}\right\} = f() \,|\, B, \Sigma, r \sim GP\big(m(), c(,)\Sigma\big)$$

  for $m() = H^T()B$

Emulator

Statistics

- Consider inputs x, corresponding model output y; calibration data z corresponding unknown x_z. First want

$$\pi(x_z \mid x, y, z)$$

- Given model formulation, parameters θ, random probabilities p,

$$\begin{aligned}\pi(x_z \mid x, y, z) &= \int\int \pi(x_z, p, \theta \mid x, y, z) dp d\theta \\ &\propto \int\int \pi(y, z \mid p)\, \pi(p \mid \theta, x, x_z)\, \pi(\theta, x_z, p) dp d\theta\end{aligned}$$

- Simulate from $\pi(x_z \mid x, y, z)$

Emulator

Further considerations

- May benefit from block diagonal form for Σ, partly based on intuition, partly to ease computational burden
- Overdispersion in eg. $Y \mid X = x \sim Bi(n(x), p(x))$
  - handled with a further level of hierarchy

$$\log\left\{\frac{p()}{1-p()}\right\} \mid q() \sim N\left(q(), \sigma^2\right)$$

$$q() \mid \ldots \sim GP(m(), c(,)\Sigma)$$

- What if we're not interested in the outputs used for calibration?
  - eg. total cost of all treatments
  - extend emulator

Present state of work

- Parameter region refinement almost finished—model runs take a while though
- Emulator coded and tested on toy data
- Precise emulator specification—eg. mean function form, whether Σ block diagonal—may require tinkering
- Draft paper (target Sept. 2011) part-written, awaiting results from emulator; then ready