Dealing with Data Gradients: “Backing Out” & Calibration

Nathaniel Osgood
CMPT 858
Sources for Parameter Estimates

- Surveillance data
- Controlled trials
- Outbreak data
- Clinical reports data
- Intervention outcomes studies
- Calibration to historic data
- Expert judgement
- Systematic reviews

Content redacted for copyright compliance
Sensitivity Analyses

• Same relative or absolute uncertainty in different parameters may have hugely different effect on outcomes or decisions

• Help identify parameters that strongly affect
  – Key model results
  – Choice between policies

• We place more emphasis in parameter estimation into parameters exhibiting high sensitivity
Dealing with Data Gradients

• Often we don’t have reliable information on some parameters, but do have other data
  – Some parameters may not be observable, but some closely related observable data is available
  – Sometimes the data doesn’t have the detailed breakdown needed to specifically address one parameter
    • Available data could specify sum of a bunch of flows or stocks
    • Available data could specify some function of several quantities in the model (e.g. prevalence)

• Some parameters may implicitly capture a large set of factors not explicitly represented in model

• There are two big ways of dealing with this: manually “backing out”, and automated calibration
“Backing Out”

- Sometimes we can manually take several aggregate pieces of data, and use them to collectively figure out what more detailed data might be.
- Frequently this process involves imposing some (sometimes quite strong) assumptions:
  - Combining data from different epidemiological contexts (national data used for provincial study).
  - Equilibrium assumptions (e.g. assumes stock is in equilibrium. Cf deriving prevalence from incidence).
  - Independence of factors (e.g. two different risk factors convey independent risks).
Example

- Suppose we seek to find out the sex-specific prevalence of diabetes in some population
- Suppose we know from published sources
  - The breakdown of the population by sex ($c_M$, $c_F$)
  - The population-wide prevalence of diabetes ($p_T$)
  - The prevalence rate ratio of diabetes in women when compared to men ($rr_F$)
- We can “back out” the sex-specific prevalence from these aggregate data ($p_F$, $p_M$)
- Here we can do this “backing out” without imposing assumptions
Backin Out

\# male diabetics + \# female diabetics = \# diabetics

\[(p_M \times c_M) + (p_F \times c_F) = p_T \times (c_M + c_F)\]

• Further, we know that \(p_F / p_M = rr_F \Rightarrow p_F = p_M \times rr_F\)

• Thus

\[(p_M \times c_M) + ((p_M \times rr_F) \times c_F) = p_T \times (c_M + c_F)\]

\[p_M \times (c_M + rr_F \times c_F) = p_T \times (c_M + c_F)\]

• Thus

\[-p_M = p_T \times (c_M + c_F) / (c_M + rr_F \times c_F)\]
\[-p_F = p_M \times rr_F = rr_F \times p_T \times (c_M + c_F) / (c_M + rr_F \times c_F)\]
Disadvantages of “Backing Out”

• Backing out often involves questionable assumptions (independence, equilibrium, etc.)

• Sometimes a model is complex, with several related known pieces
  – Even thought we may know a lot of pieces of information, it would be extremely complex (or involve too many assumptions) to try to back out several pieces simultaneously
Another Example: Joint & Marginal Prevalence

<table>
<thead>
<tr>
<th></th>
<th>Rural</th>
<th>Urban</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>$p_{MR}$</td>
<td>$p_{MU}$</td>
<td>$p_{M}$</td>
</tr>
<tr>
<td>Female</td>
<td>$p_{FR}$</td>
<td>$p_{MU}$</td>
<td>$p_{F}$</td>
</tr>
<tr>
<td></td>
<td>$p_{R}$</td>
<td>$p_{U}$</td>
<td></td>
</tr>
</tbody>
</table>

Perhaps we know

• The count of people in each \{Sex, Geographic\} category
• The marginal prevalences ($p_R$, $p_U$, $p_M$, $p_F$)

We need at least one more constraint

• One possibility: assume $p_{MR} / p_{MU} = p_R / p_U$

We can then derive the prevalences in each \{Sex, Geographic\} category
Calibration: “Triangulating” from Diverse Data Sources

- Calibration involves “tuning” values of less well known parameters to best match observed data
  - Often try to match against *many* time series or pieces of data at once
  - Idea is trying to get the software to answer the question: “What must these (less known) parameters be in order to explain all these different sources of data I see”

- Sometimes we learn from this that our model structure just can’t produce the patterns!
Calibration

• Calibration helps us find a reasonable (specifics for) “dynamic hypothesis” that explains the observed data
  – Not necessarily the truth, but probably a reasonably good guess – at the least, a consistent guess

• Calibration helps us leverage the large amounts of diffuse information we may have at our disposal, but which cannot be used to directly parameterize the model

• Calibration helps us falsify models
Calibration: A Bit of the How

• Calibration uses a (global) optimization algorithm to try to adjust unknown parameters so that it automatically matches an arbitrarily large set of data

• The data (often in the form of time series) forms constraints on the calibration

• The optimization algorithm will run the model many (minimally, thousands, typically 100K or more) times to find the “best” match for all of the data
Example Global Optimization Algorithm

• Starts at random position, tries to improve match (minimize error) by
  – “Tweaking” parameters
  – Running Model
  – Recording error function

• Keeps on improving until reaches “local minimum” in error of fit
  – May add some randomness to knock out of local minima
Running Calibrations in Vensim:
(Under Model/Simulate Commands)
Optimization Control

Filename: GDMCalibration16.voc

Output Level: On
Trace: Off
Sensitivity: Off
Multiple Start: Random
Random type: Linear
Seed: 254
#Restart: 10000
Optimizer: Powell
Max Iterations: 1000
Max Sims:
Pass Limit: 2
Fractional Tolerance: 0.0003
Tolerance Multiplier: 21
Absolute Tolerance: 1
Scale Absolute: 1
Vector Points: 25

Currently active parameters (drag to reorder):
0 <= Net RI Emigration Weight Coefficient for Age Sex[Reproductive]
0 <= Net RI Emigration Weight Coefficient for Age Sex[PostReproductive]
0 <= Net RI Emigration Weight Coefficient for Age Sex[Child,Male] = 0.5
0 <= Net RI Emigration Weight Coefficient for Age Sex[Child,Female]
0 <= Net GP Emigration Weight Coefficient for Age Sex[Reproductive]
0 <= Net GP Emigration Weight Coefficient for Age Sex[PostReproductive]
0 <= Net GP Emigration Weight Coefficient for Age Sex[Child,Male] = 0.5
0 <= Net GP Emigration Weight Coefficient for Age Sex[Child,Female]

Model value of constant 1

OK
Payoff Definition

Payoff Definition. Edit the filename to save changes to a different control file

Filename: GDMCalibration.vpd

Type: Calibration

Payoff Elements

Total Weighted Absolute Value of All Discrepancy for Current Time/1-1

Variable

Compare to

Weight

The weight should be positive for calibration. For policy optimizations use a positive number when more is better and a negative number when less is better.

OK

Cancel
Assessing Model “Goodness of Fit”

• To improve the “goodness of fit” of the model to observed data, we need to provide some way of quantifying it!

• Within the model, we
  – For each historic data, calculate discrepency of model
    • Figure out absolute value of discrepency from comparing
      – Historic Data
      – The model’s calculations
    • Convert the above to a fractional value (dividing by historic data)
  – Sum up these discrepency
Required Information for Calibration

• Specification of what to match (and how much to care about each attempted match)
  – Involves an “error function” (“payoff function”, “penalty function”, “energy function”) that specifies “how far off we are” for a given run (how good the fit is)

• A statement of what parameters to vary, and over what range to vary them

• Characteristics of desired tuning algorithm
  – Single starting point of search?
Considerations

• Adding constraints helps increase identifiability (selection of realistic best fit)
• Adding parameters to tune leads to larger space to explore
• Adding too many parameters to tune can lead to underdetermined situation
• All fits are within constraints of model
The Pieces of the Elephant
Example Model of Underlying Process & Time Series It Must Match
Single Model Matches Many Data Sources
Slides Adapted from External Source
Redacted from Public PDF for Copyright Reasons
Pieces of the Elephant: STI

ObsVsPred_Cases

ObsVsPred_Incidence

ObsVsPred_Population

ObsVsPred_Prevalence

ObsVsPred_Tested

Science