Dealing with Data Gradients: “Backing Out” & Calibration

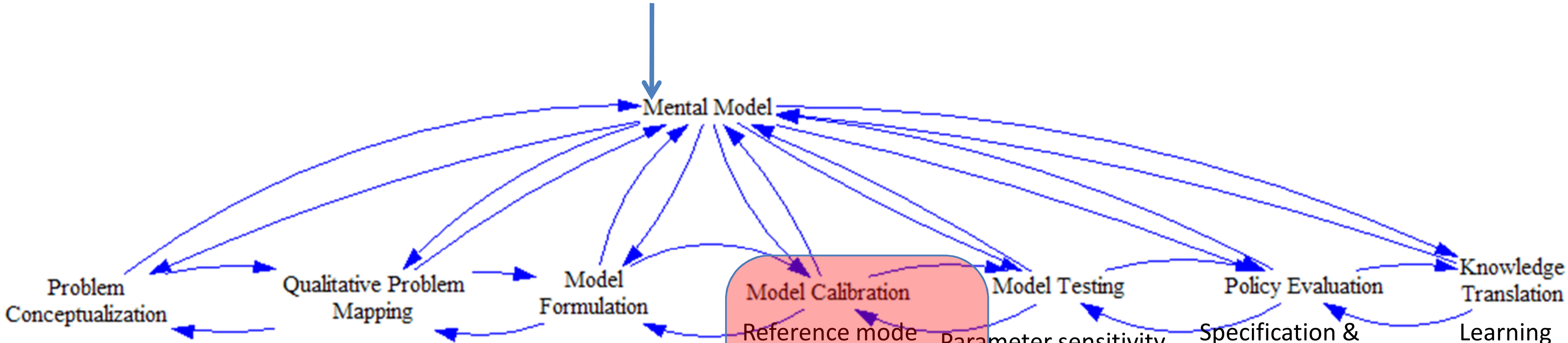
Nathaniel Osgood

CMPT 858

Term Project Updated Due Date

- Because of Holiday weekend, date is now midnight April 25

A Key Deliverable!



Model scope/boundary selection.
Model time horizon
Identification of key variables
Reference modes for explanation

Causal loop diagrams
Stock & flow diagrams
Policy structure diagrams

Group model building

Specification of
•Parameters
•Quantitative causal relations
•Decision rules
Initial conditions

Reference mode reproduction
Matching of intermediate time series
Matching of observed data points
Constrain to sensible bounds
Structural sensitivity analysis

Parameter sensitivity analysis
Cross-validation
Robustness&extreme case tests
Unit checking
Problem domain tests

Specification & investigation of intervention scenarios
Investigation of hypothetical external conditions
Cross-scenario comparisons (e.g. CEA)

Learning environments/Microworlds/flight simulators

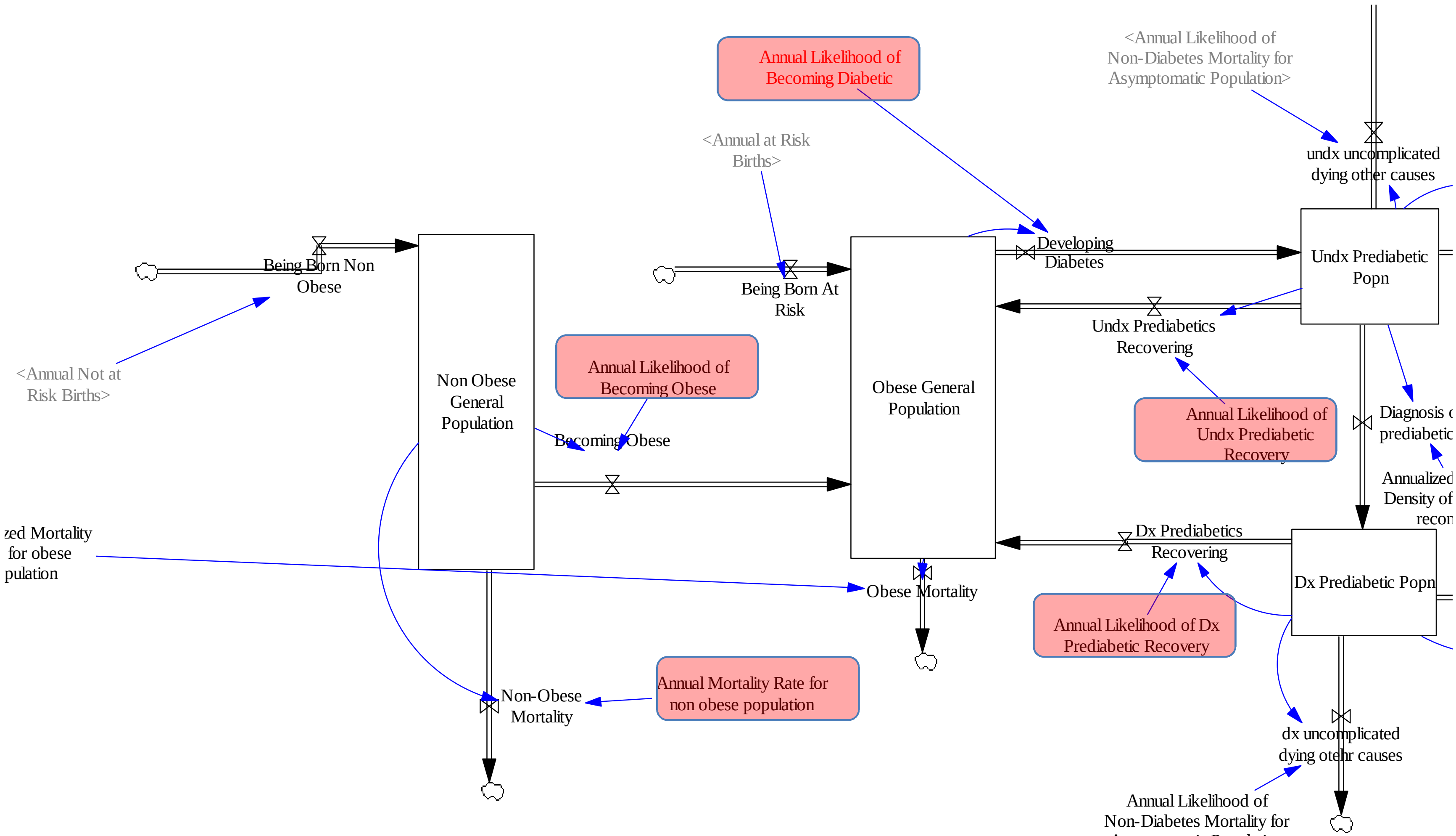
Some elements adapted from H. Taylor (2001)

Sources for Parameter Estimates

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

Parameter*	Description	Baseline value (units)	Reference
μ	Entry/exit of sexual activity	0.0056 (years^{-1})	Garnett and Bowden, 2000
c	Partner change rate per Susceptible	16.08 (years^{-1})	Approximated from Garnett and Bowden, 2000
β	Probability of infection per sexual contact	0.70	Garnett and Bowden, 2000
φ	Fraction of Infectives who are symptomatic	0.20	Garnett and Bowden, 2000
$1/\gamma$	Latent period	0.038 (years)	Brunham et. al., 2005
$1/\sigma$	Duration of infection	0.25 (years)	Brunham et. al., 2005
θ	Asymptomatic recovery coefficient	1.5	Garnett and Bowden, 2000
$1/\pi$	Duration of naturally-acquired immunity	1 (year)	Approximated from Brunham et. al., 2005

Introduction of Parameter Estimates



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

Backing Out

male diabetics + # female diabetics = # diabetics

$(p_M * c_M) \quad + \quad (p_F * c_F) \quad = p_T*(c_M+c_F)$

- Further, we know that $p_F / p_M = rr_F \Rightarrow p_F = p_M * rr_F$
- Thus

$(p_M * c_M) \quad + \quad ((p_M * rr_F) * c_F) \quad = p_T*(c_M+c_F)$

$p_M*(c_M + rr_F * c_F) = p_T*(c_M+c_F)$

- Thus
  - $p_M = p_T*(c_M+c_F) / (c_M + rr_F * c_F)$
  - $p_F = p_M * rr_F = rr_F * p_T*(c_M+c_F) / (c_M + rr_F * c_F)$

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

	Rural	Urban	
Male	p_{MR}	p_{MU}	p_M
Female	p_{FR}	p_{MU}	p_F
	p_R	p_U	

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”
- Observed data can correspond to complex combination of model variables, and exhibit “emergence”
- Frequently 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

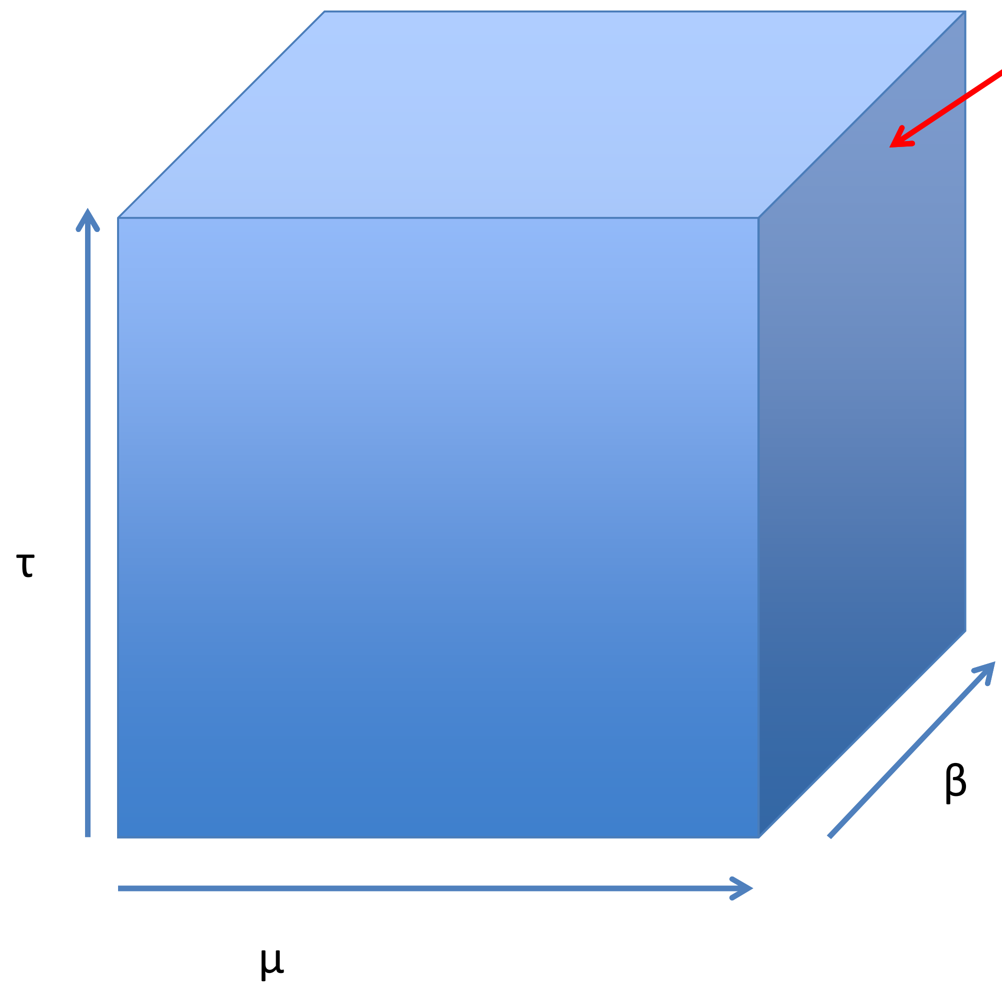
- Calibration uses a (global) optimization algorithm to try to adjust unknown parameters so that it automatically matches an arbitrarily large set of data
- The data (often in the form of time series) forms constraints on the calibration
- The optimization algorithm will run the model many (minimally, thousands, typically 100K or more) times to find the “best” match for all of the data

Required Information for Calibration

- Specification of what to match (and how much to care about each attempted match)
  - Involves an “error function” (“penalty function”, “energy function”) that specifies “how far off we are” for a given run (how good the fit is)
  - Alternative: specify “payoff function” (“objective function”)
- A statement of what parameters to vary, and over what range to vary them (the “parameter space”)
- Characteristics of desired tuning algorithm
  - Single starting point of search?

Envisioning “Parameter Space”

For each point in this space, there will be a certain “goodness of fit” of the model to the collective data



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 discrepancy of model
    - Figure out absolute value of discrepancy from comparing
      - Historic Data
      - The model's calculations
    - Convert the above to a fractional value (dividing by historic data)
  - Sum up these discrepancy

Characteristics of a Desirable Discrepancy Metric

- **Dimensionless**: We wish to be able to add discrepancies together, regardless of the domain of origin of the data
- **Weighted**: Reflecting different pedigrees of data, we'd like to be able to weigh some matches more highly than others
- **Analytic**: We should be able to differentiate the function one or more times
- **Concave**: Two small discrepancies of size *a* should be considered more desirable than having one big discrepancy of size 2*a* for one, and no discrepancy at all for the other.
- **Symmetric**: Being off by a factor of two should have the same weight regardless of whether we are 2x or ½x
- **Non-negative**: No discrepancy should cancel out others!
- **Finite**: Finite inputs should yield infinite discrepancies

A Good Discrepancy Function (Assuming non-negative h & m)

Exponent
>1 $\Rightarrow$ concave with respect to h-m

Taking average in denominator (together w/squaring of result) ensures symmetry with respect to h&m

$$w \cdot \left(\frac{h-m}{average(h,m)} \right)^2 = w \cdot \left(\frac{h-m}{\left(\frac{h+m}{2} \right)} \right)^2$$

Division $\Rightarrow$ Dimensionless
(Judging by *proportional* error, not absolute)

Only zero if h=m=0.
Denominator is only very small if numerator is as well!

Considerations for Weighting

- **Purpose of model**: If we “care” more about a match with respect to some variables, we can more heavily weight matches for those variables
- **Uncertainty in estimate**: The more uncertain the estimate of the quantity, the lower the weight
- **Whether data exists**: no data => weight should be zero