Introduction to systems Science Modeling

Roch Nianogo & Ashley Buchanan

Workshop outline

- Overview of systems science modeling
- Aggregate-level models
  - Systems dynamic model (SDM)
  - Markov state transition model (STM)
- Individual-level models
  - Microsimulation models (MSM)
  - Agent-based models (ABM)
- Q&A

Outline

- Background
- Simulation fundamentals
- Modeling steps
- Good modeling practices
- Lab

Background

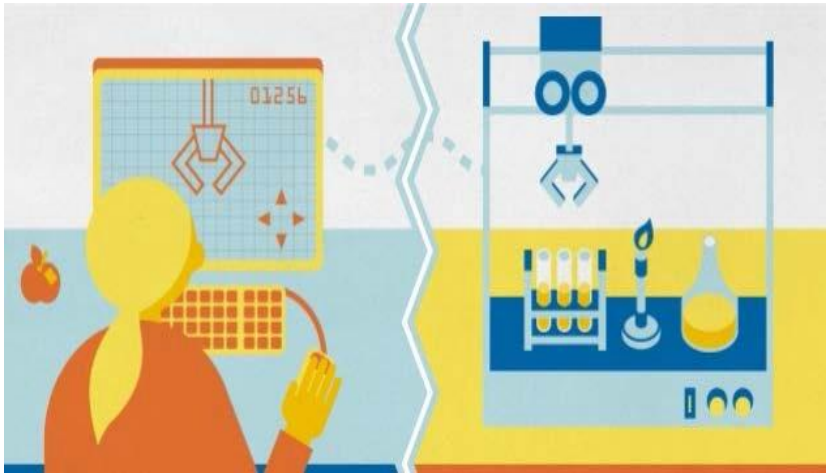
What is it?

- Systems science is a "*conceptual framework that emphasizes the relationships that connect constituent parts of a system rather than the parts themselves"* whose goal is to help understand the interactions and influences between the parts of a system

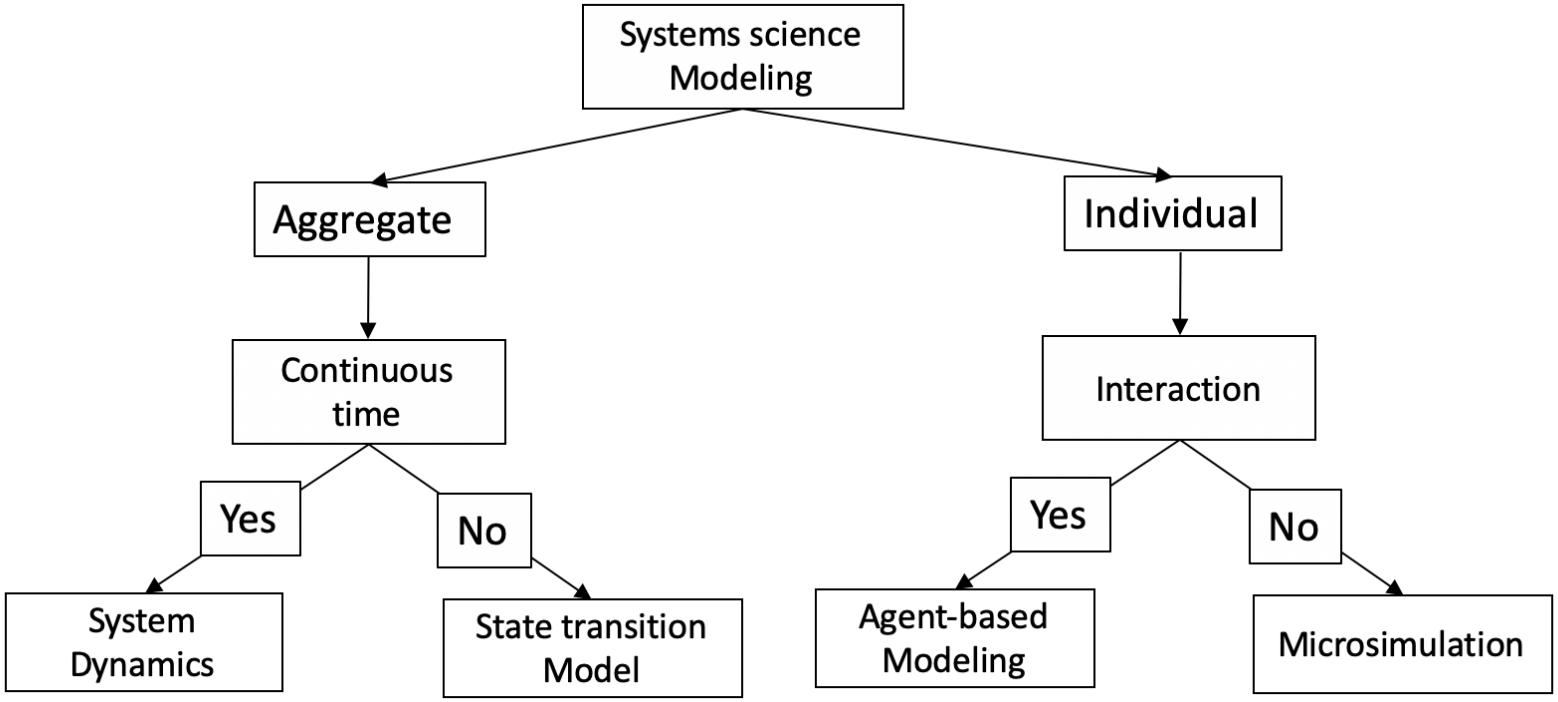
El-Sayed & Galea, Systems Science and Population Health. Oxford; New York: Oxford University Press, 2016

What is it?

- Systems science at its core, uses "simulation modeling", i.e., replication/representation of reality using computer models

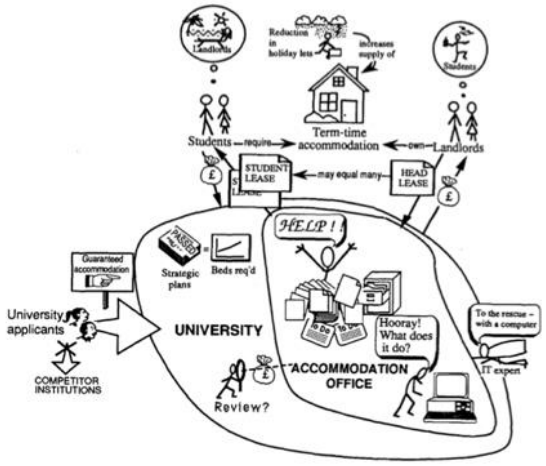


Different types of modeling approaches



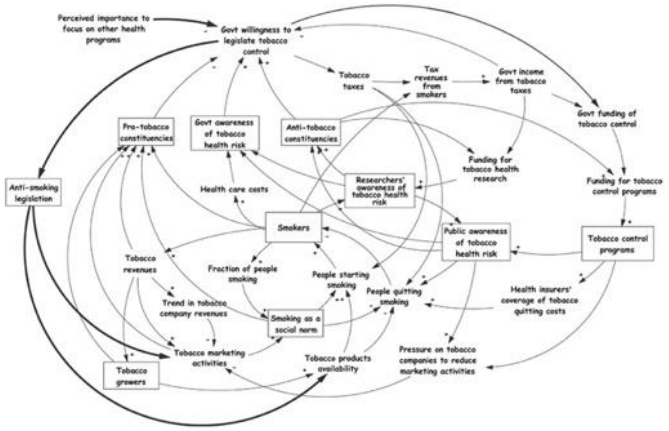
Diagrammatic representation

Rich Picture Diagram (RPD)



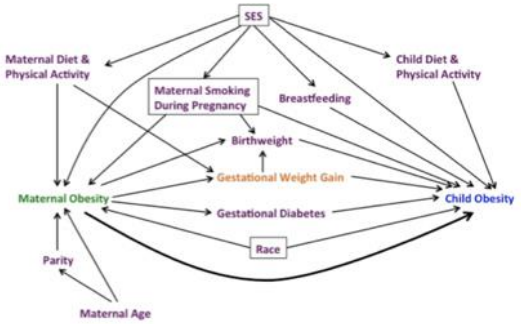
P.J. Lewis. Rich picture building in the soft systems methodology. 1992

Causal Loop diagram (CLD)



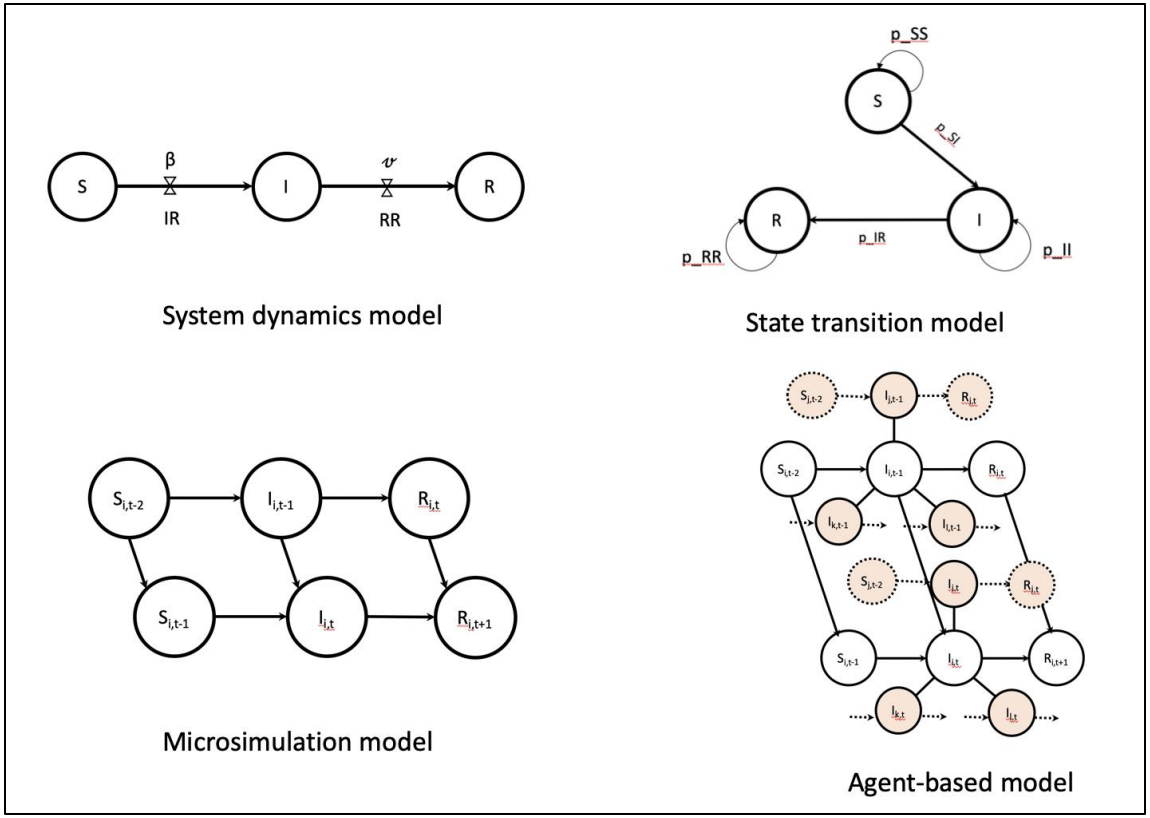
System Dynamics Model for Tobacco Control, Luke et al, ARPH, 2010

Directed Acyclic Diagram (DAG)



Josey et al. Overall gestational weight gain mediates the relationship between maternal and child obesity. BMC, 2019

Diagrammatic representation



Dia grammatic representation of the four analytical simulation models
S: Susceptible, I: Infected, R: recovered; RR= recovery rate, IR = infection rate, beta = force of infection, v = force of recovery, p_SI: transition probability from the susceptible state (S) to the infected state (I), For this figure, it would be good to explain and define what the nodes and edges in each subfigure

Simulation fundamentals

Generating hypothetical data

- General features of simulation models
  - Stochastic vs deterministic
  - Dynamic vs static
  - Continuous vs discrete time scale
  - Empirical vs stylized
  - Individuals vs aggregate
- Seed and random number generator
- Distributions and their properties
  - Four essential distribution functions (PDF, CDF, random generator, etc…)
  - Common discrete and continuous distributions (Binomial, Gaussian)

Stochastic vs deterministic

- *A model is said to be stochastic* if it involves some element of randomness
  - $y \sim 2 + 3x + \varepsilon$; $\varepsilon \sim N(0,1)$
- Otherwise, it is said to be *deterministic* and its behavior is entirely predictable
  - $y \sim 2 + 3x$

Dynamic vs static

- A model is said to be *dynamic* if it represents a system as it evolves over time,
- Otherwise, it is *static* and represents the system at a particular point in time (e.g. it only “jumps” from, say, the baseline to the end of the simulation without intermediate steps)

Discrete vs continuous

- A dynamic model is discrete when for instance, the features/variables in the system change discretely over time (e.g. discrete event simulation),
- Otherwise, it is continuous and the features/variables in the system changes continuously over time (e.g. most systems dynamics model)

Stylized vs empirical

- A *stylized* model is a model in which parameters used in the model (e.g. annual rate of BMI change among adolescents) are fake estimates or guesstimates—generally done in theory building and hypothesis testing;
- Otherwise, the model is said to be *empirical* (e.g. where most estimates come from the literature).
- Of note, a model can have a mixture of guestimates and empirical estimates and sometimes, a model can start stylized and end up empirical—this can generally done in the model specification step

Individual-based vs aggregate

- A model is said to be *individual-based* (e.g. microsimulation, agent-based models) when individual characteristics such as personal attributes (e.g. age, sex) are incorporated in the model,
- otherwise, it is said to be *aggregated* (e.g. systems dynamic models, Markov/cohort state transition model, etc…), when it is at the population level.
- In the individual-based models, the model incorporates individual heterogeneity while in the latter case, the population is *often* assumed to be homogenous—i.e., everyone in the population have the same attribute.

Seed and random number generator

- Every simulation model begins with a random number generator (RNG).
- The RNG produces a stream of pseudo-random numbers using some underlying (usually recursive) algorithms but which appear random and satisfy measures of randomness
- Invoking any RNG function entails pointing at any number (seed) of this chain and, starting from there, taking a chunk of numbers (of a length = our desired sample size).

Seed and random number generator

- Setting the seed is essential for reproducing results as the simulation conducted using the same seed should yield the same results.
- A corollary of this is that since different seeds will generate different results, we can use this phenomenon to introduce randomness in the runs.
  - Resampling (bootstrap)
  - Simulation variation

Distributions and their properties

- When considering simulation modeling, it is important to understand the following
  - 1) the four essential distribution functions,
  - 2) their defining parameters and
  - 3) the common discrete and continuous distributions and their properties.

Four essential distribution functions

Functions	Definition and notations	Example with the Normal distribution in R
Probability density function (PDF)	f(x) Returns the probability density at a given point for a variety of distributions (it is akin the likelihood function): P(X=x)	dnorm(x, mean, sd) returns the density or the value on the y-axis of a probability distribution for a discrete value of x
Cumulative distribution function (CDF)	F(X) = P(X≤x) Returns the probability that an observation from the specified distribution ≤ to a particular value	pnorm(q, mean, sd) returns the cumulative density function (CDF) or the area under the curve to the left of an x value on a probability distribution curve

Four essential distribution functions

Functions	Definition and notations	Example with the Normal distribution in R
Quantile function (Q)	$Q(p) = F^{-1}(p)$ Closely related to the CDF function, but solves an inverse problem. It is the inverse of CDF. For a given probability, p, it returns the smallest value, q, for which CDF(q) is >= to P	qnorm(p, mean, sd) returns the quantile value, i.e. the standardized z value for x
Random generation function	Generates a random sample from a distribution	rnorm(n, mean, sd) returns a random simulation of size n

Common distribution parameters

- a location parameter (e.g. mean),
- a scale or dispersion parameter (e.g. standard deviation) and
- a shape parameter which is neither the location nor the scale and which determines the shape of the distribution (e.g. shape). Distribution that have a shape parameter include the beta and gamma distribution.
- Other parameters include the mode, median, skew, etc…

Some continuous distributions

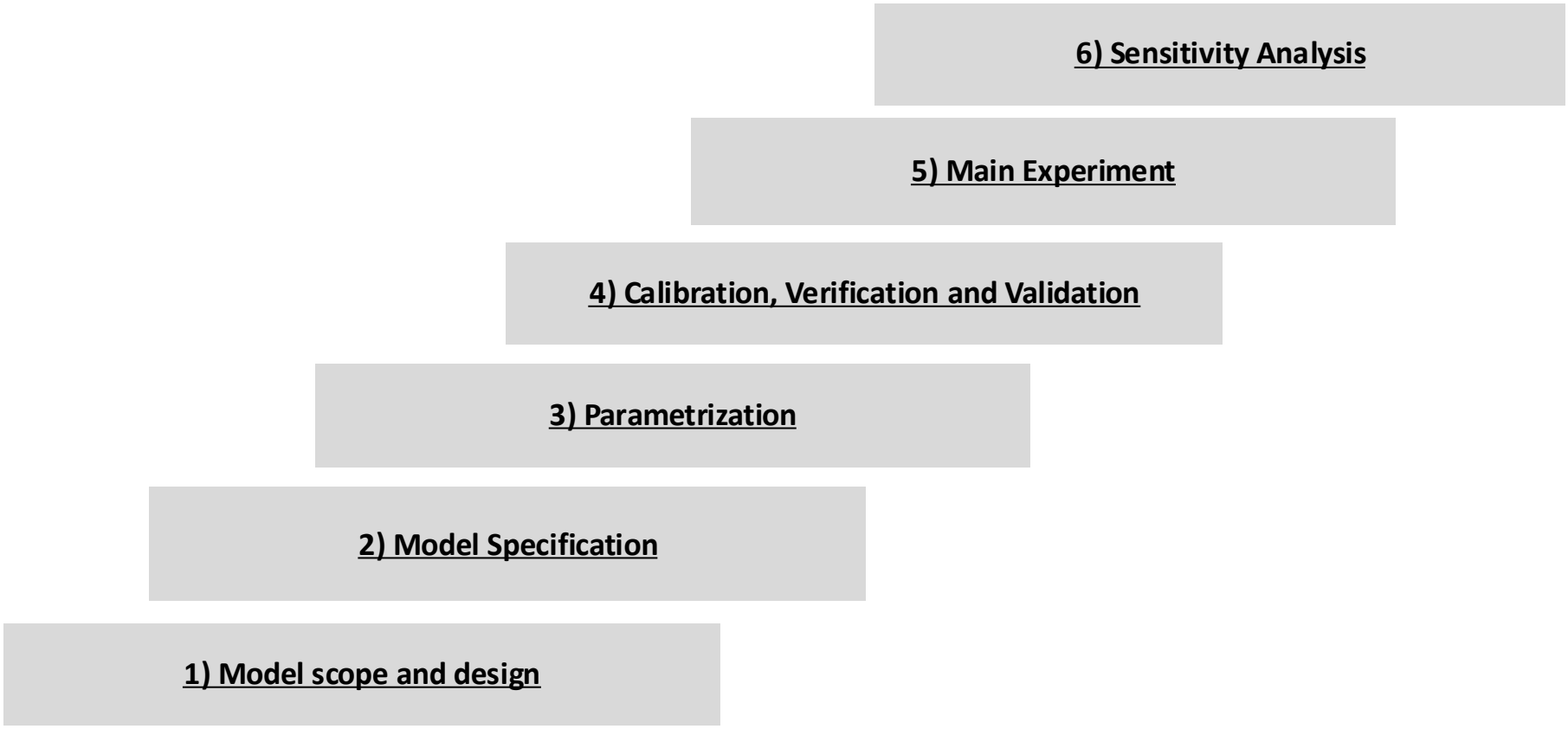
Distributions	R	Parameters	Examples
Beta	*rbeta*	Shape: α and β	Typically used for percentages and proportion and for random variables limited to interval [0,1]
Expo-nential	*rexp*	Scale: σ = 1/λ Where λ is the rate	Survival time (i.e., time between events)
Gamma	*rgamma*	Shape: k Scale: θ = 1/λ where λ is the rate	Cost, survival time Insurance claims Age distribution of cancer incidence When k=1, the distribution reduces to an exponential distribution
Normal	*rnorm*	Location (mean): μ is Scale (standard deviation): σ	Most continuous processes (e.g. age)
Weibull	*rweibull*	Shape: k Scale: σ = 1/λ where λ is the rate	Survival time When k=1, the distribution reduces to an exponential distribution

Some discrete distributions

Distributions	R	Parameters	Examples
Binomial	*rbinom*	Probability of success: p Number of trials: n	Number of successes in a sequence of n independent experiments (Bernoulli trial) A special case is the Bernoulli distribution with number of trials = 1
Poisson	*rpois*	Location: $\mu = \lambda T$; where λ is the rate of events over the period of time T $E(X) = \lambda T = Var(X) = \lambda T = \mu$	Event counts
Negative binomial	*rnbinom*	Location: $\mu = \lambda T$; $E(X) = \mu$ $Var(X) = \mu + p\mu^2$ where p is the dispersion parameter	Event counts (overdispersion) When p=0, the distribution reduces to a Poisson distribution
Uniform	*runif*	Minimum (min): a Maximum (max): b	Typically used when no assumption is made about the distribution of a process (arbitrary outcome that lies between certain bounds). Can be discrete or continuous.

Simulation steps

Simulation model building steps



Step 1: Model scope and design

- Define the purpose of the model
  - What is the goal of the model?
  - What interventions would be explored?
  - What factors are important and relevant?
- Determine the type of modeling
  - Observational vs Interventional
  - Stylized vs empirical
- Review with stakeholders the important pieces that need to be included in the model (system mapping)
  - Focus groups
  - Panel discussions
  - Review of the literature
- Draw a causal systems map
  - Directed Acyclic Diagram (DAG)
  - Causal Loop diagram

Step 2: Model Specification

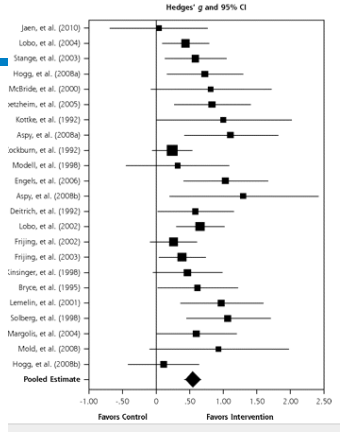
$$Y = \alpha + \beta \cdot x$$

$$f(x) = \frac{dy}{dx}$$

```
if (condition == TRUE){
  #do something
} else {
  #do something else
}
```

- Definition
  - Conceptualizing and developing the structure of the model
- Methods
  1. Attributes (e.g. population size, socio-demographics, behaviors, outcomes,…)
  2. Environment (e.g. hospital settings, states,…)
  3. Decision rule
     - Regression-based simulation (e.g. Microsimulation)
     - Differential equation-based simulation (e.g. systems dynamics)
     - Conditional statements (e.g. if then, else), (e.g. agent-based)
     - Matrix multiplication (e.g. state transition model)

Step 3: Model Parametrization



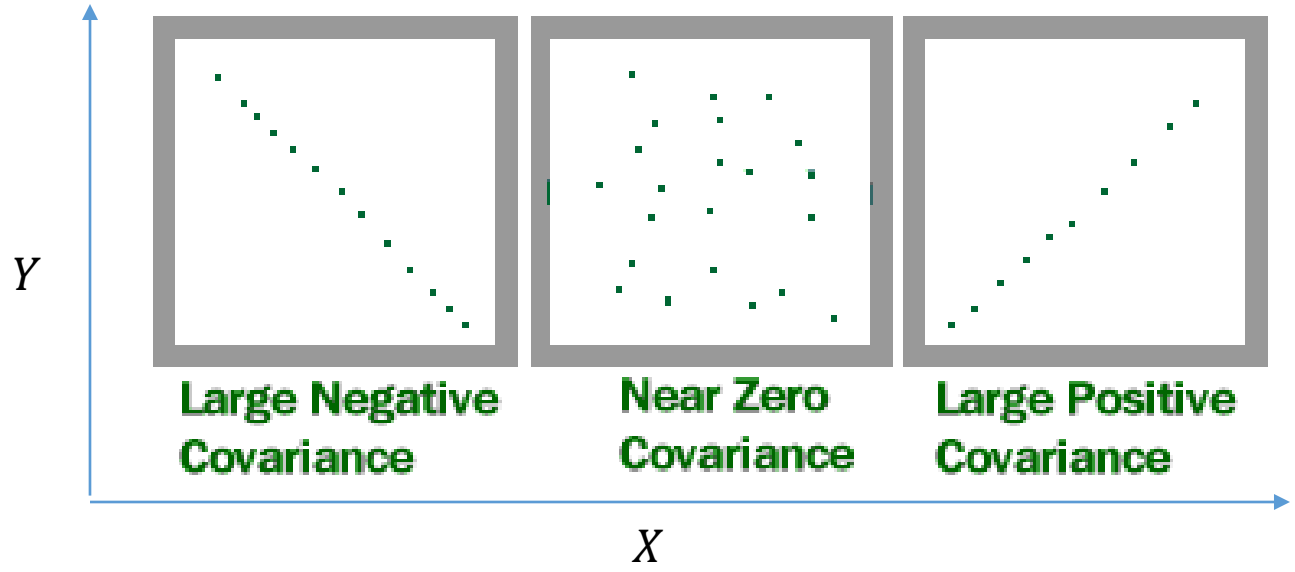
- Definition
  - Assign empirical values to model parameters from databases and review of the literature
- Methods
  - Systematic search of the literature and expert knowledge
    - Evidence-level 1 (peer-reviewed journal article of meta-analysis, systematic reviews, randomized trials, cohort studies)
    - Evidence-level 2 (peer-reviewed journal article of cross-sectional studies)
    - Evidence-level 3 (internal analysis of publicly or privately available data)

Step 3: Model Parametrization

- Types of Data
  - Measures of occurrence and precision
    - Means (continuous variables)
    - Proportions (categorical variables)
    - Variance and standard deviations
  - Measures of association
    - Correlations/Covariance
    - Regression coefficients
      - Mean Differences
      - Risk ratios
      - Odds ratio

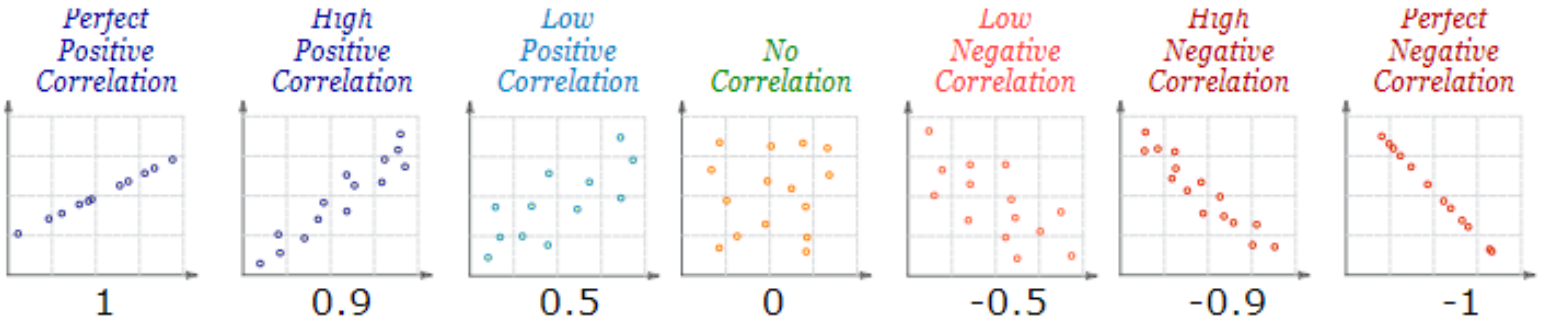
Step 3: Model Parametrization

- **Covariance**
  - $Cov(X,Y) = E[(X - \mu_x)(Y - \mu_x)]$
  - $Cov(X,Y) = E(XY) - \mu_x\mu_y$



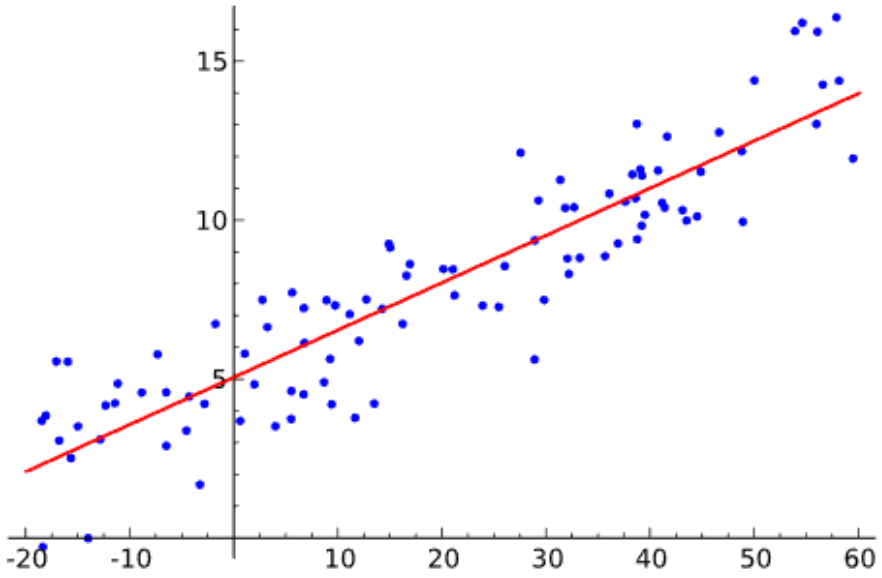
Step 3: Model Parametrization

- **Correlation**
  - $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x * \sigma_y}$



Step 3: Model Parametrization

- **(Linear) Regression**
  - $Y = a + b_X x + \varepsilon$
  - $b_X = \frac{Corr(X,Y) * \sigma_y}{\sigma_x} = \frac{Cov(X,Y)}{\sigma_x^2}$

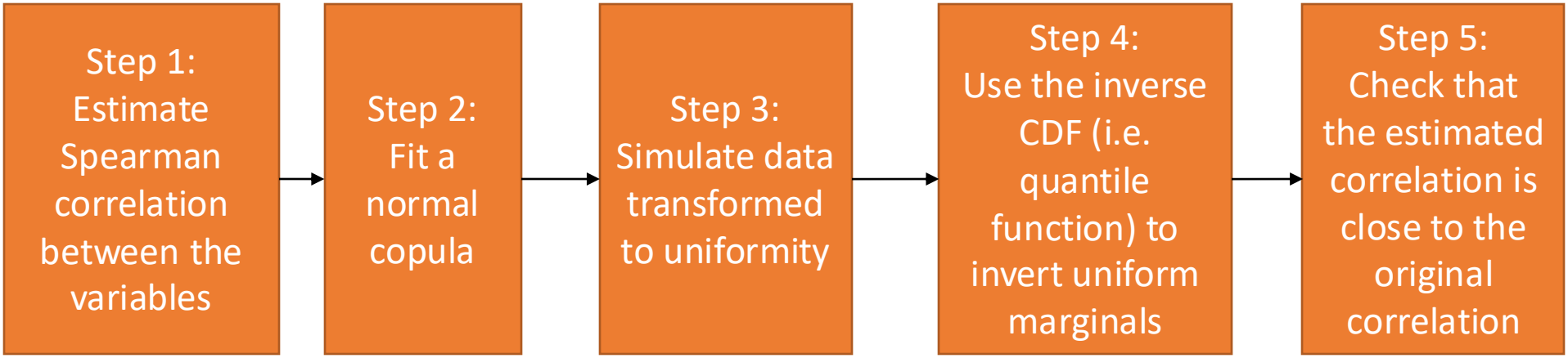


Step 3: Model Parametrization

- Generating data from correlated data
  - We can use copula functions to generate data from correlated data
  - “copula” means to link or join
  - It allows one to create a multivariate distribution by joining univariate marginal distributions

Step 3: Model Parametrization

Generating data from correlated data



Step 3: Model Parametrization

- Parameter conversion

Converting relative risks to risk =>

$$p_1 = \mathrm{RR} \times p_0 = \left(\frac{p_1}{p_0}\right) \times p_0.$$

$$p_1 \approx \left(\frac{p_{1\_\mathrm{adjusted}}}{p_{0\_\mathrm{adjusted}}}\right) \times p_{0\_\mathrm{unadjusted}}.$$

$$r = \frac{-\ln(1-p)}{t},$$

Converting rates to probabilities =>

$$p = 1 - \exp(-rt).$$

$$p = 1 - (1-p)^{1/n}.$$

Probability (p), rate (r) and time (t)

Step 4a: Calibration (Optimization)

- Definition
  - Concerned with assigning empirical baseline or trend characteristics (also known as input parameters) to the virtual neo-system

Step 4a: Calibration (Optimization)

- Calibrate/Optimize the model

$$\hat{Y} = a + b_X x$$

Calibration-in-the-large

- Fine tuning of intercept (a)
- Grid search

$\text{mean}(\hat{Y}_{predicted}) = \text{mean}(Y_{observed})$

Step 4a: Calibration (Optimization)

- Step 1: Define an objective function to assess goodness of fit (e.g. comparing predicted vs observed outcome)
  - Distance => minimize
    - RMSE and SSE
    - MAE
- Step 2: Determine a list of parameters to calibrate and range of parameters to search through
  - Grid search,
  - Random search
  - Iterative search (Nelder-Mead)

Some distance functions

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

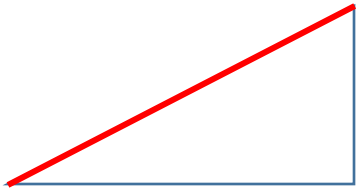
$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$||x||_2 = \sqrt{\left(\sum_i x_i^2\right)} = \sqrt{x_1^2 + x_2^2 + \ldots + x_i^2}$$

$$||x||_1 = \sum_i |x_i| = |x_1| + |x_2| + \ldots + |x_i|$$

L2-norm

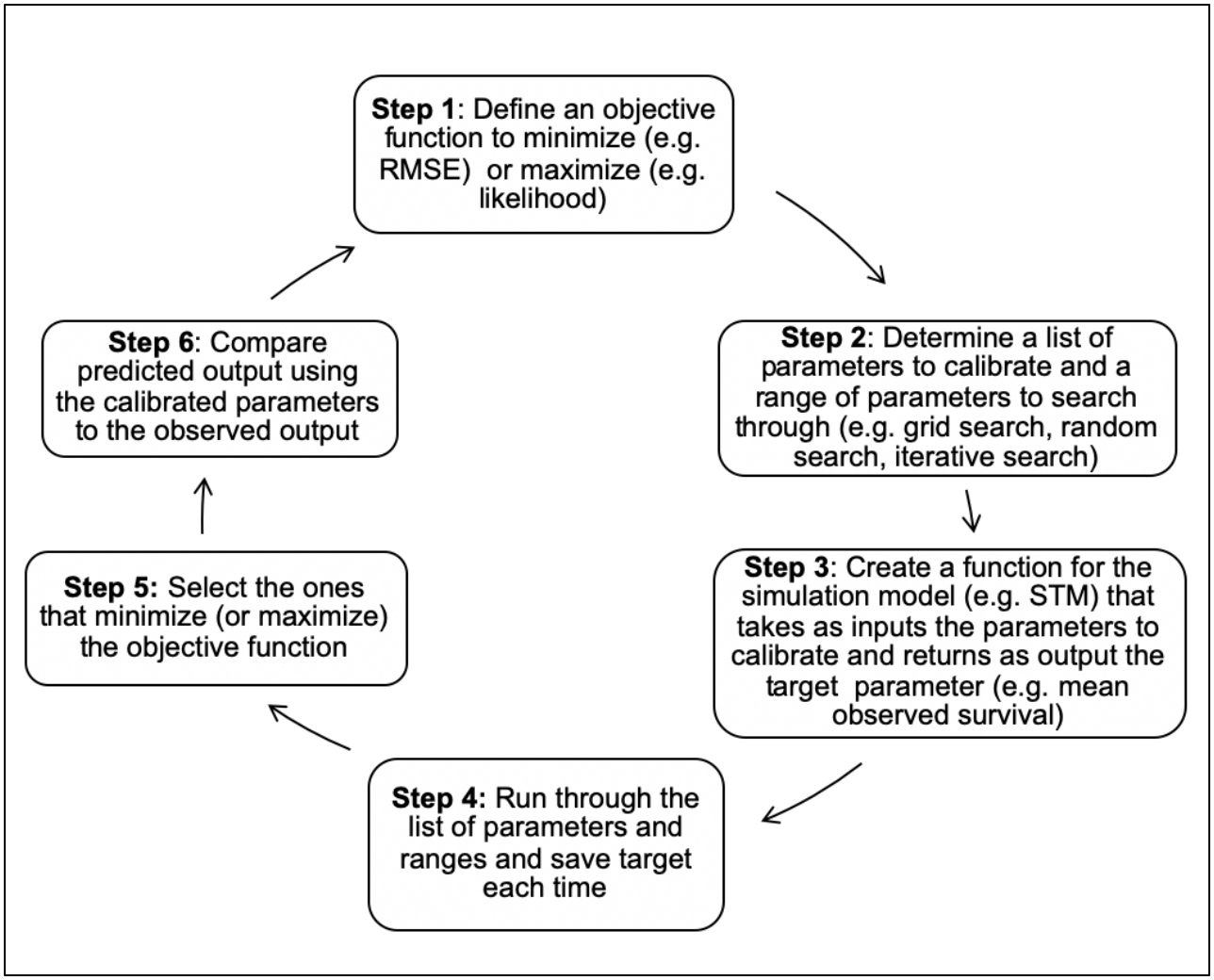
Euclidian norm



L1-norm

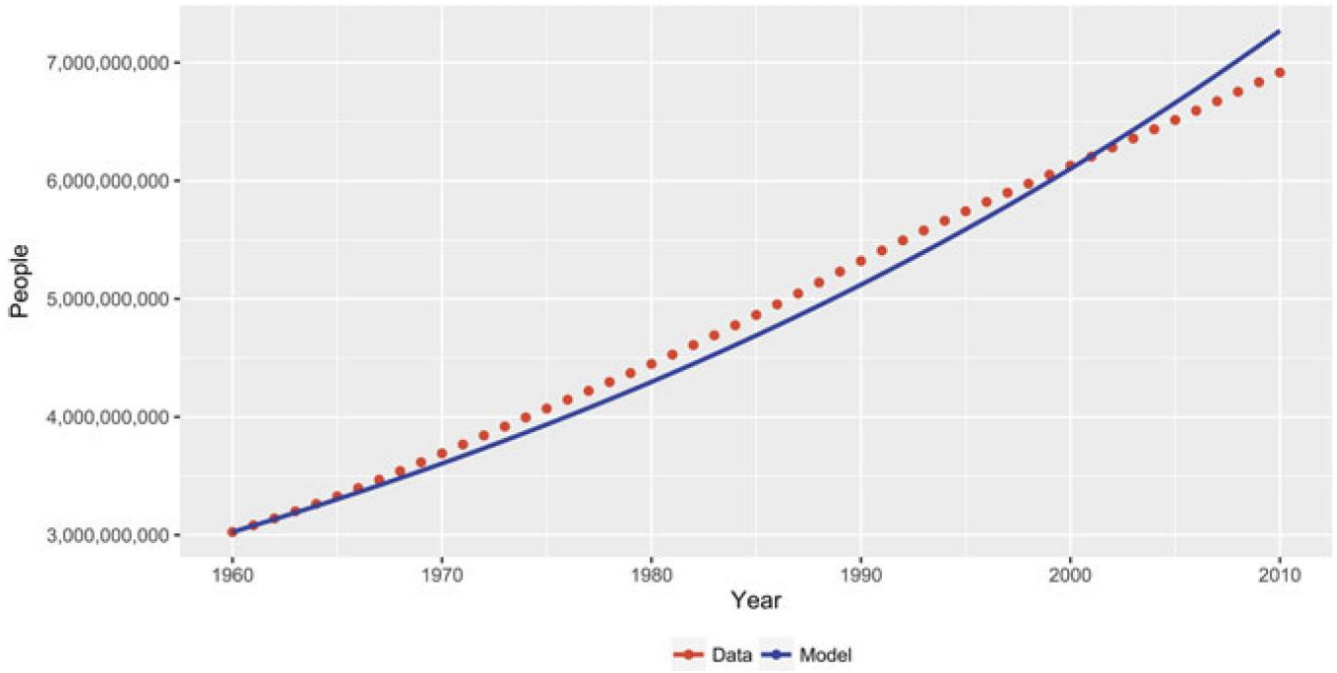
Taxicab norm or Manhattan norm

Calibration steps



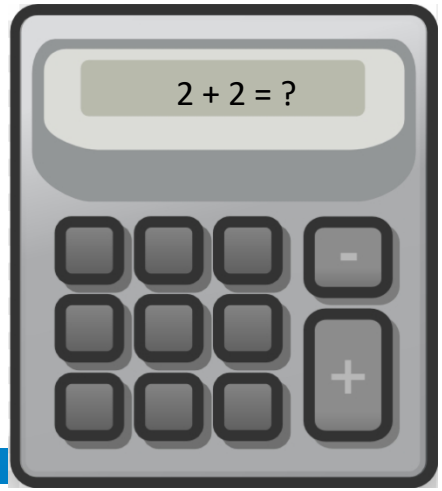
Step 4a: Calibration (Optimization)

- Calibration plot



Step 4b: Verification (Testing)

- Definition
  - Concerned with debugging the model, checking for errors in coding and making sure that the model does what it is intended to do (e.g. correct calculation)
- Methods
  - Baseline output operation of the codes compared to the expectations stated in the design documents.
  - Structured code walk-throughs
  - Debugging walk-throughs



Code review

American Journal of Epidemiology

Vol. 190, No. 10
https://doi.org/10.1093/aje/kwab092
Advance Access publication:
April 8, 2021

Practice of Epidemiology

Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research

Anusha M. Vable*, Scott F. Diehl, and M. Maria Glymour

* Correspondence to Dr. Anusha M. Vable, Department of Family and Community Medicine, Zuckerberg San Francisco General Hospital, Building 80, Ward 83, 995 Potrero Avenue, San Francisco, CA 94110 (e-mail: anusha.vable@ucsf.edu).

A Need for Change! A Coding Framework for Improving Transparency in Decision Modeling

Fernando Alarid-Escudero[1] · Eline M. Krijkamp[2] · Petros Pechlivanoglou[3] · Hawre Jalal[4] · Szu-Yu Zoe Kao[5] · Alan Yang[6] · Eva A. Enns[5]

Table 3 Recommended prefixes in variable names that encode data and variable type

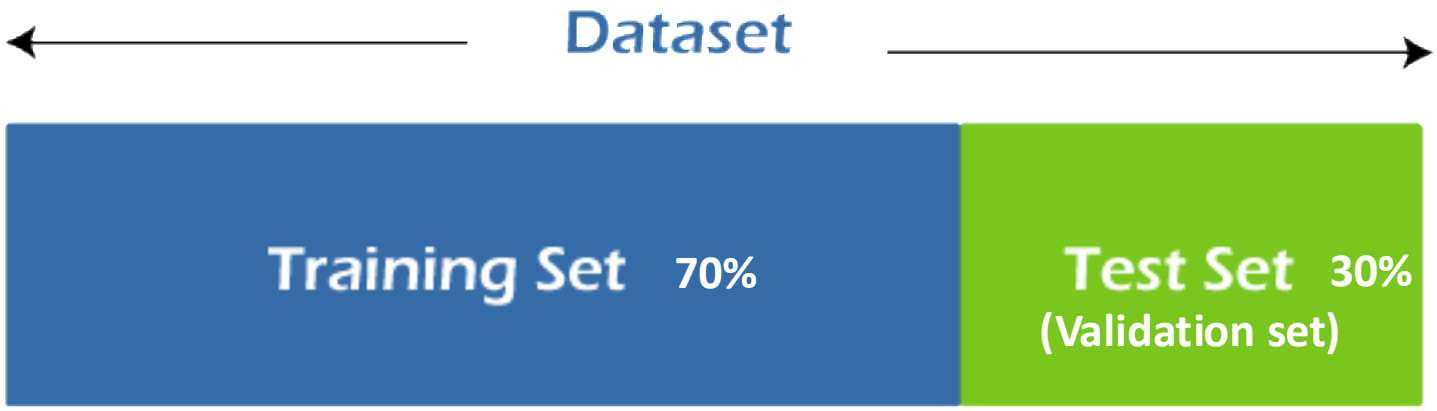
Prefix	Data type	Prefix	Variable type
<> (no prefix)	scalar	n	Number
v	vector	p	Probability
m	matrix	r	Rate
a	array	u	Utility
df	data frame	c	Cost
dtb	data table	hr	Hazard ratio
l	list	rr	Relative risk
		ly	Life years
		q	QALYs
		se	Standard error

QALYs quality-adjusted life-years

Step 4c: Validation

- Definition
  - Concerned with how accurate the virtual system reflects the real-word system
  - Similar to the calibration procedures but uses external data (e.g. different type period, different sites, testing/validation set) to compare predicted output to observed output
- Methods
  - *Predictive validation* (comparing the predicted baseline output data to the original real-world data)
  - *Historical data validation (Comparing* the predicted output to historical data whenever available
  - *Face validity*
  - *Process Validation*

Step 4c: Validation



Apply calibration algorithm

Evaluate model
(interval validity or internal quantitative credibility)
Comparing model predicted output to observed data in training set

Evaluate model
(external validity or external quantitative credibility)
Comparing model predicted output to observed data in test/validation set

Step 5: Main Experiment

- Observational
  - Forecasting (stochastic prediction)
    - Modeling *what the future could be given what we know today.*
    - Examples of metrics include predicted risk, rates, survival and cases

Step 5: Main Experiment

- Interventional
  - Simulating counterfactual "what if" scenarios to assess
  - Modeling *what the future could be if we could change something today*
    - Comparative effectiveness, i.e., comparing the impact of two interventions (examples of metrics commonly reported include risk ratio, risk difference and numbers needed to intervene on)
    - Cost-effectiveness, i.e,. comparing the cost-effectiveness of two interventions (examples of metrics commonly reported include net monetary benefit (NMB) and incremental cost-effectiveness ratio (ICER))

Step 5: Main Experiment

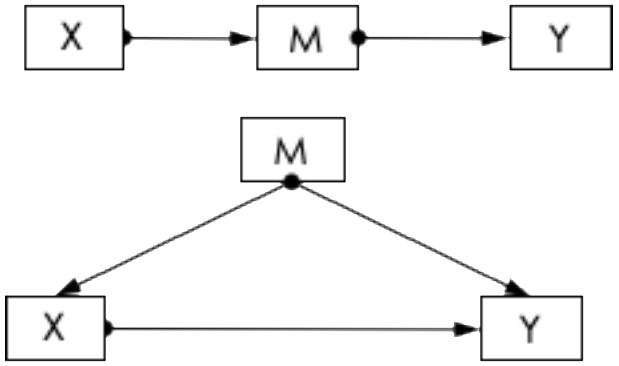
- Evaluating statistical methods or theory
  - Evaluate statistical methods
    - Comparing relatively novel methods to conventional or previously used methods
    - Examples of metrics commonly reported include the amount of bias relative to the “truth”, precision estimate (e.g. variance, standard error) and coverage probability
  - Evaluate theory
    - Testing a theory
    - These models are often stylized but can be empirically based as well.
    - The types of metrics will vary depending on the subject

Step 6: Uncertainty and Sensitivity Analysis

- Definition
  - The process by which we handle the uncertainty. Allows to test how your model output changes over ranges of uncertain
- Type
  - Model/structural uncertainty
  - Parameter uncertainty

Step 6: Uncertainty and Sensitivity Analysis

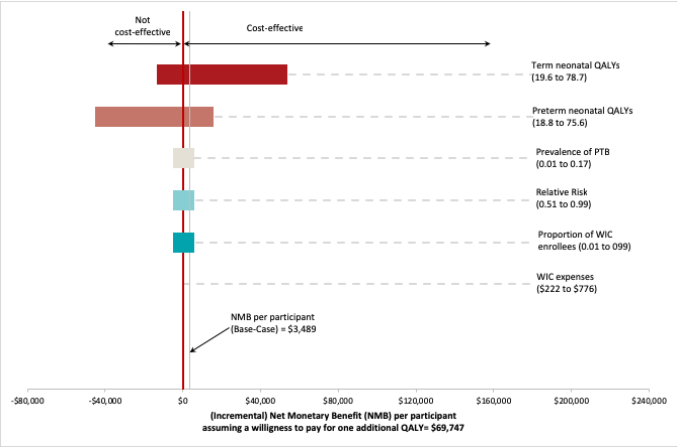
- Model or structural uncertainty

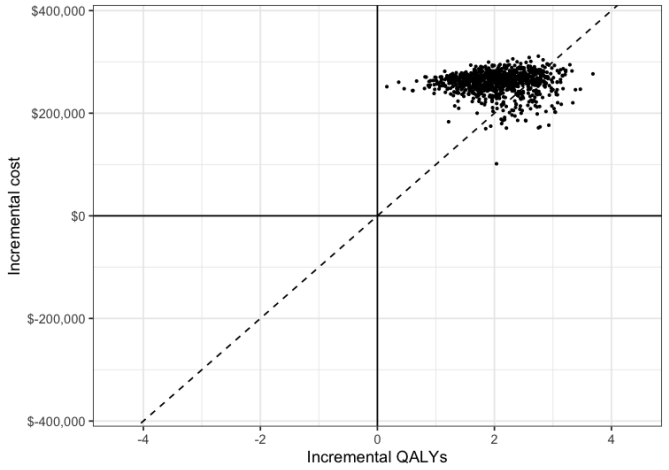


Step 6: Uncertainty and Sensitivity Analysis

- Parameter uncertainty

	Univariate	Multivariate
Deterministic (fixed)	One-way sensitivity analysis	Multivariate sensitivity analysis (scenario sensitivity analysis)
Probabilistic (stochastic)	Univariate PSA (rarely done)	Multivariate PSA





Simulation variation and resampling (bootstrap)

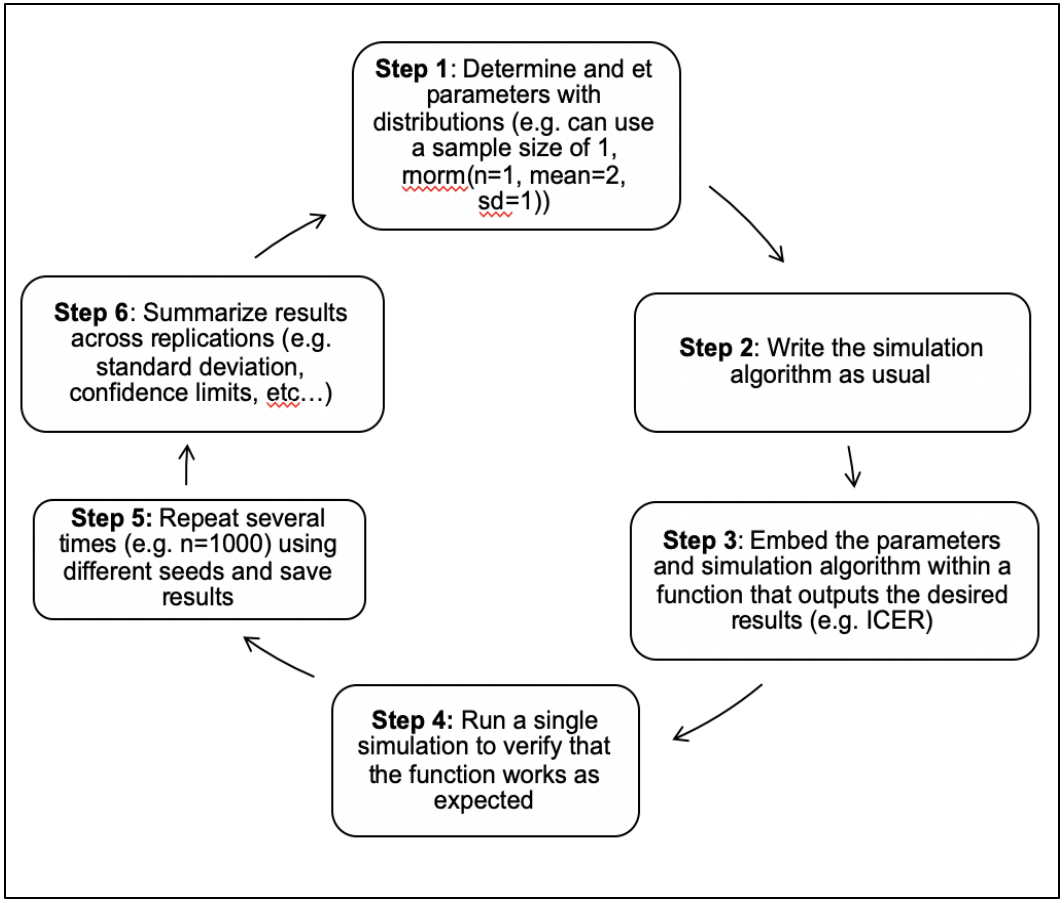
- One can assess the distribution of a parameter by
  - repeating the ”base simulation” several times → the simulation approach or
  - resampling with replacement an original dataset → the bootstrap approach
- This generates some random errors that can be used to estimate the
  - simulation interval (simulation approach) or
  - confidence interval (bootstrap approach)
- The intervals can then be computed by
  - obtaining the percentiles (2.5^{th} and 97.5^{th}) for the distribution or
  - using a normal approximation: CL = Mean +/- 1.96*SE

Approaches

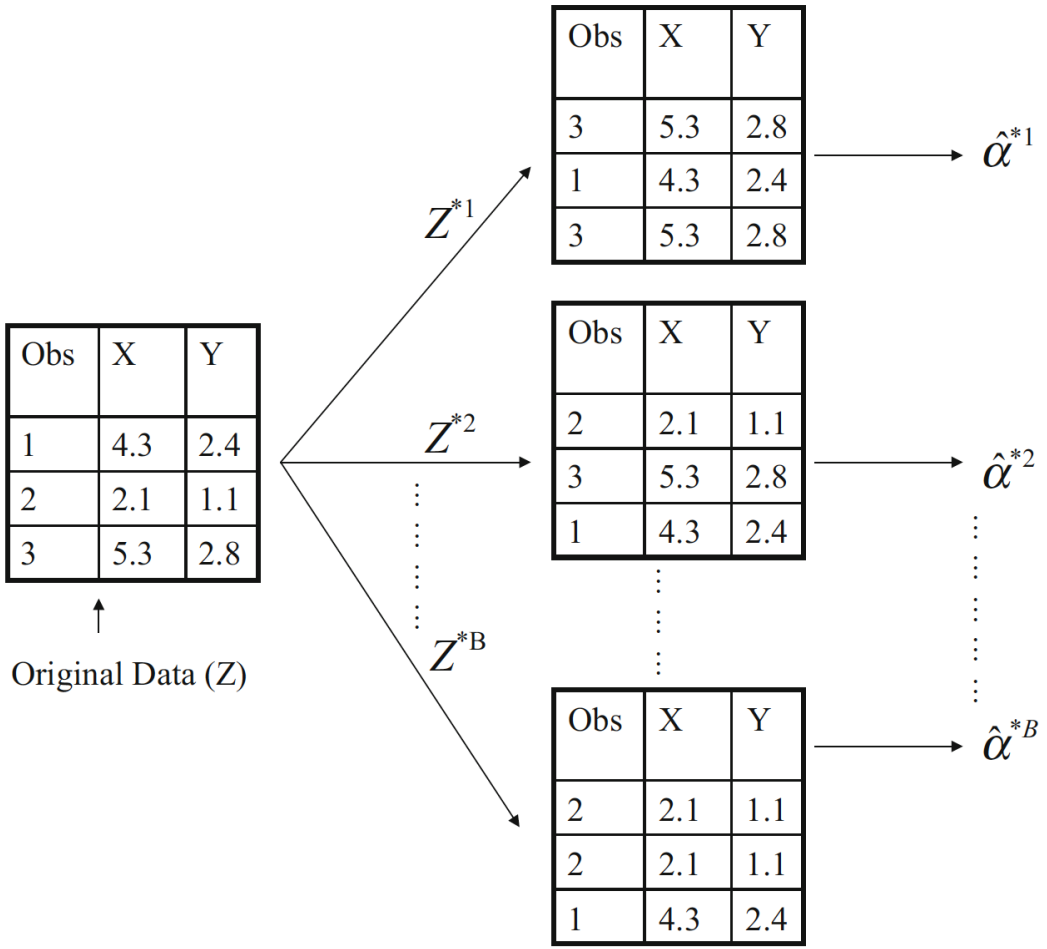
Simulation approach

- 1) Create a function for the data generating process for a single simulation
  - The function should ideally return the parameter you would like to estimate
- 2) Repeat this process several times (e.g. rep=10,000) and save the rep number of the parameters
- 3) Obtain mean and the percentiles (2.5^{th} and 97.5^{th}) of the distribution

Probabilistic sensitivity analysis (PSA)



Bootstrap


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Bootstrap

Bootstrap approach

- 1) Create a function for a single sample with replacement
  - The function should ideally return the parameter you would like to estimate
- 2) Repeat this process several times (e.g. rep=10,000) and save the rep number of the parameters
- 3) Obtain mean and the percentiles (2.5^{th} and 97.5^{th}) of the distribution

Good modeling practices and reporting

Special cases:
- Agent-based modeling
- Economic evaluation

General simulation

Good modeling practices and reporting

Types of modeling

EQUATOR network

Agent-based model

ODD protocol

Other types of modeling

- STRESS for SDM, STM, ABM, MSM
- CONSORT and STROBE simulation

Purpose of modeling

Decision analysis

- CHEERS
- Second Panel on CEA

Forecasting and Prediction

- EPIFORGE

Others

- STRESS
- CONSORT and STROBE simulation

Good modeling practices and reporting

- General themes
  1. Transparency
  2. Justification
  3. Systematic documentation
  4. Clarity
  5. Parsimony
  6. Conceptual diagram (e.g. Markov diagram)
  7. Visualization of outputs
  8. Sharing codes or algorithms and decision rules?

Good modeling practices and reporting

Table 2. The six principles of reporting simulation studies.

1. State the purpose of the study and the model's intended use
2. Provide enough detail to reproduce the results of the base run of the model and any simulation experiments conducted as part of the study
3. Ensure that descriptions of the model are software and hardware *independent*
4. Include data for verification and parameter values. Where proprietary or ethical issues prevent the inclusion of data, "hypothetical" test data should be included for verification purposes
5. Document all software and where necessary hardware-specific implementation
6. Provide additional visualisation of model logic or algorithms using a recognised diagramming approach

Some additional tips

Tips

- Setting a seed for reproducibility
  - set.seed()
- Be faithful to the hypothesized data-generating mechanism
- Be consistent when naming variables
- Use large number of repetition to reduce the Monte Carlo standard error
- Use vectors and matrices whenever possible
- Start small and increase in complexity
- Start with hypothetical parameters before searching for empirical data

Tips

- Benchmarking to assess the performance (time) of various portion of the program
- Reduce size of parameter grid search (e.g. power, varying standard deviations and sample size)
- Go parallel if needed. Set up independent simulations and estimate the time it takes to run one

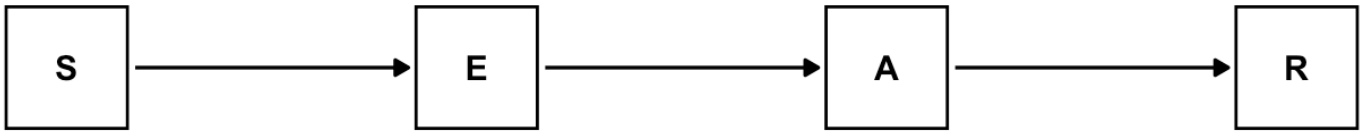
Lab

Lab

- We will explore a ”non-infectious” disease process => Susceptible-Exposed-Adopted-Reduced
- In this system, everyone is overweight/obese but there is one peer educator in the population
- The peer educator can convince his/her peers to adopt a behavior change (e.g. exercise or diet) for weight reduction.
- Once a behavior is adopted, there will be some time before individuals start to lose weight
- This model is primarily based of the diffusion of information theory

Lab

- Four states
  - Susceptible (S): Individuals who are unaware of the information but are potentially able to learn about it.
  - Exposed (E): Individuals who have been introduced to the information (e.g., heard about it) but are not yet acting on it to change their behavior.
  - Adopted (A): Individuals who adopted the behavior change, i.e., changed behavior
  - Reduced weight or Weight reduction (R): Individuals who's weight have reduced



Lab

- Four modeling approaches
  - Systems dynamics model
  - State transition model
  - Microsimulation model
  - Agent-based model