Agent-based modeling goes mainstream

Ben Klemens
Nonresident Fellow, Brookings
Senior Statistician, Mood and Affective Disorders, NIMH

Challenge(s) in agent-based modeling (ABM)

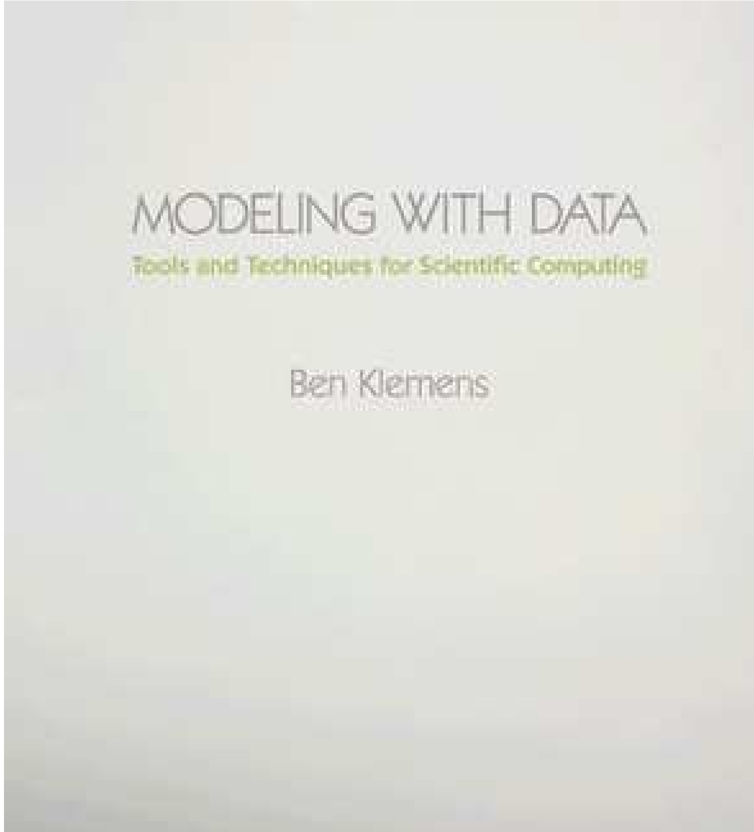
Bring the model and the data closer together.

The literature slide

- *Agent-based social simulation: a method for assessing the impact of seasonal climate forecast applications among smallholder farmers*, Ziervogel, Bithell, et al.
- *An In Silico Transwell Device for the Study of Drug Transport and Drug–Drug Interactions*, Garmire, Garmire, et al.
- *Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review*, Parker, Manson, et al.
- Today's presentations

The literature slide (self-citation)

- Modeling with Data: Tools and Techniques for Statistical Computing
- `http://modelingwithdata.org`



MODELING WITH DATA

Tools and Techniques for Scientific Computing

Ben Klemens

The outline slide

- Defining a model
- Defining probability
- Applying statistical technique to agent-based models
- An example: Finding the Sierpinski triangle

What is a model?

- Ask the OED:
  - A person employed to wear clothes for display, or to appear in displays of other goods.
  - *euphem.* A prostitute.
- No help at all, so here's mine:
  A function (probably intended to mirror a real-world situation) that expresses the likelihood of a given set of data and parameters.

Models are a statistical frame

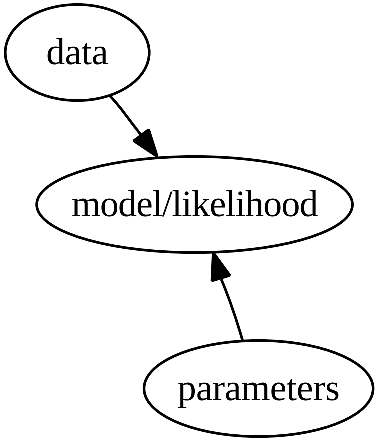
- Normal distribution.
  - inputs: mean μ, variance σ^2, your observation x
  - output: $P(x, \mu, \sigma)$.

OLS (Ordinary Least Squares)

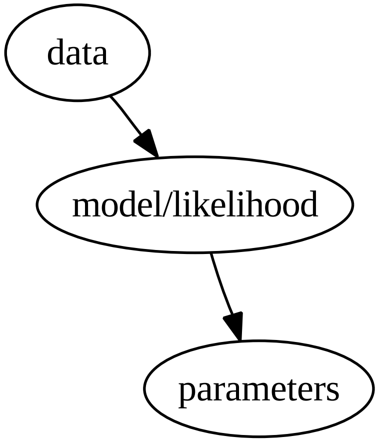
- inputs: vector of params $\boldsymbol{\beta}$, your observed dependent variable y, your observed independents $\mathbf{x}$.
- output: $P(\mathbf{x}, \boldsymbol{\beta}, y)$.
- To find $P(\mathbf{x}, \boldsymbol{\beta}, y)$, look up $\epsilon = (y - \mathbf{x}\boldsymbol{\beta})$ on the Normal distribution tables.
- OLS minimizes squared distance $(y - \mathbf{x}\boldsymbol{\beta})^2$, which is a monotonic transformation of probability.
- A type of “best fit” model—see below.
- Usually we don’t have $\boldsymbol{\beta}$ and find the **most likely** $\boldsymbol{\beta}$.

One model, taken different ways

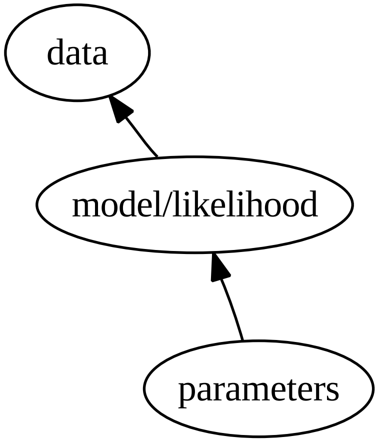
Probability



Max likelihood



Expected value, RNG



- At this level, regressions and ABMs are identical.
- ‘But Ben’, you retort, ‘the traditional model outputs a probability, while ABM outputs are not based in observed frequencies.’ [i.e., these models can’t be verified.]

Probability is problematic

The frequentist approach is not useful

- Repeat a test enough times, and count the percent success.
  - Die rolling. Coin flipping.
- This breaks quickly.
  - If the die rolls are ‘identical’, why do we get different results?
  - What about clearly non-replicable events like the weather?

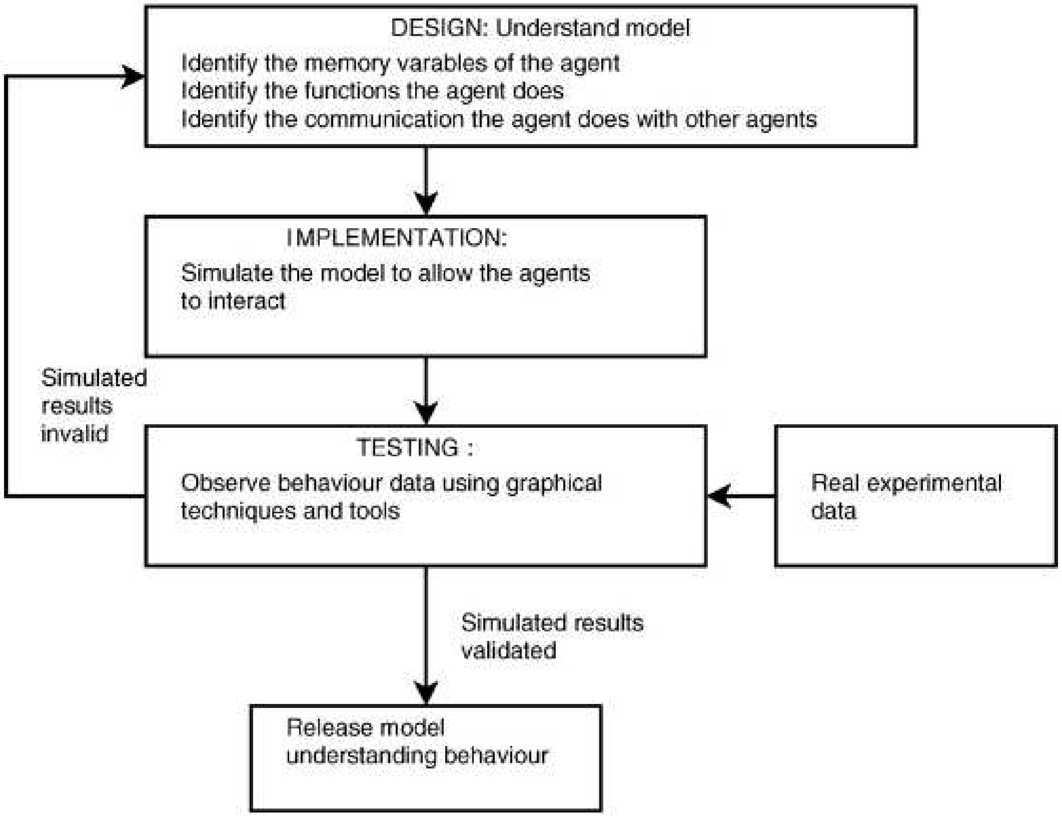
What’s the chance of rain tomorrow?

- The weatherman is always right.
  - There is no verifiable, objective probability.
  - *There’s a 99% chance of rain* and *There’s a 1% chance of rain* are equally impossible to verify.
- With enough information, couldn’t we develop an objective measure?
- ›› already knows with certainty.
- The solution is to add more structure. Given:
  - Prior data listing R_t=rain on date t, H_t=humidity, B_t=pressure
  - $R = \text{probit}(\alpha + \beta_H H + \beta_B B)$
- Now the question is meaningful.
- But there’s a sleight-of-notation: we’re not talking about P(real event), but P(event in model).

Probability statements come from the calibrated model

- Models define probabilities: *There's a 20% chance* is shorthand for *I have a model that states that there is a 20% chance.*
- Even confidence levels and p-values are derived from the model.
- Which brings us back to agent-based modeling and simulation.

Design, implement, validate against the data



DESIGN: Understand model
Identify the memory varables of the agent
Identify the functions the agent does
Identify the communication the agent does with other agents
IMPLEMENTATION:
Simulate the model to allow the agents to interact
Simulated results invalid
TESTING :
Observe behaviour data using graphical techniques and tools
Real experimental data
Simulated results validated
Release model understanding behaviour

[*Validation and discovery from computational biology models*, Kiran, Coakley, et al.]

We can use graphical tools *and* statistical tools.

- E.g., say that we seek a target pattern
  - I observe residential segregation.
  - I observe fox and hare populations oscillating.
- Define a distance between model outcome ($\widehat{x}$) given parameters and the target (x).
- It is natural to say that smaller distance = larger likelihood.[*]
- E.g., $P(\widehat{x}) \propto \frac{1}{1+D(x,\widehat{x})}$

[*] E.g., as with OLS.

It’s a statistical model!

- The likelihood function is a model that defines the probability of given parameters and data.

- But it’s not really a probability measure!

- Sure it is! $P(A) \geq 0$. $P(A \cap B) = P(A) + P(B)$. $\int_{\forall x} P(x)dx = 1$.

- But there may be alternate re-scalings!

- The invariance principle: don’t sweat the details!
  - A number and its square have the same quantity of information.

- But the model is *ad hoc*!

- So is OLS! Being from the early 1900s does not make a model objective. Nor does invoking limited mathematical facts like the CLT.

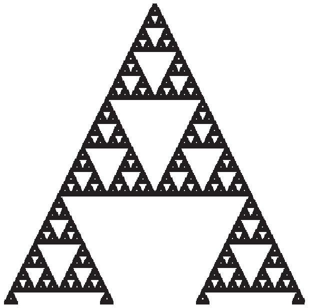
So what?

- Almost every procedure that can be applied to a traditional statistical model can be applied to an ABM.
- Find the most likely parameters.
- Forecasting: Once you fit existing data, produce a new output distribution given changes in data or parameters.
- Find the variance of the parameters (i.e., robustness of output given $\triangle$ parameter).
- $\Rightarrow$ Find confidence intervals or p-values for the parameters
- Hierarchical modeling: Use a local ABM for each group; regress the outputs from all ABMs.
- Bayesian update: Normal distribution + your model $\Rightarrow$ a new histogram expressing a distribution.

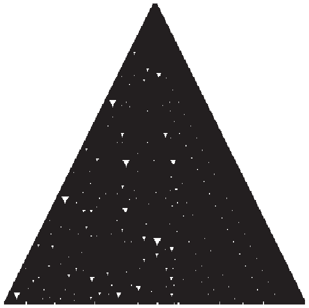
An example: the Sierpinski triangle

- There are seven rules (=parameters). Select each as on or off.
  - In binary:
    0101001=41
    0101011=43
    1101001=105
- See Wolfram or *Finding Optimal ABMs* @ SSRN.com for details.

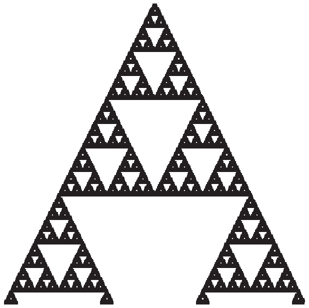
Configuration 41



Configuration 43



Configuration 105



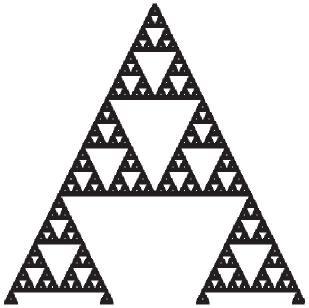
Our procedure

- This is a small space, so run every possibility.
- Measure the distance between the output and the Sierpinski triangle.
- Calculate the matrix of differentials (i.e., value with bits (i, j) minus the value without).
- Use the Cramér-Rao Lower Bound: invert the square of the differential matrix to calculate the variance in output given a change in input.

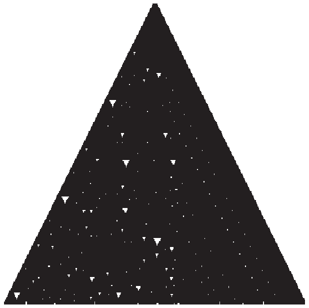
The variances

rule	variance
1: (0, 0, 1)	4.790
2: (0, 1, 0)	3.541
3: (0, 1, 1)	14.402
4: (1, 0, 0)	4.788
5: (1, 0, 1)	15.994
6: (1, 1, 0)	14.403
7: (1, 1, 1)	20.471

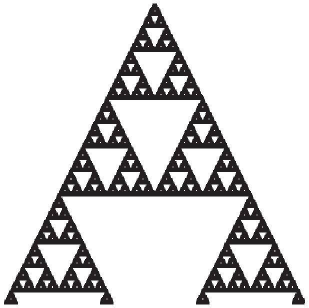
Configuration 41



Configuration 43



Configuration 105



In conclusion

- Agent-based models are increasingly quantitative.
- Agent-based models are first-class models, and we can use them as such, for both descriptive and inferential work.