Narrative Modeling

Ben Klemens

Center for Statistical Research and Methodology
U.S. Census Bureau

28 August 2014

This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Thanks

- Harlan Harris
- Marck Vaisman
- Sean Gonzalez
- Ben Sweezy

The Storyline

The storyline:

- We think in stories.
- We want to bring them to data, fit and evaluate them.
- What would a tool set to formalize narrative stories look like?

The implementation:

- 10 mins: the state of social science modeling
- 5 mins: an implementation sidebar
- 10 mins: building model objects
- 5 mins: transformations of model objects
- 20 mins: increasingly elaborate examples

Formalizing the story (Probability)

The storyline:
Start with $100. Every day, it grows or shrinks by some percent.
What does the total look like after ten days?

Formalizing the story (Probability)

The storyline:
Start with \$100. Every day, it grows or shrinks by some percent.
What does the total look like after ten days?

The formalized model:

- outcome $= 100 \cdot k_1 \cdot k_2 \cdot \dots \cdot k_{10}$
- $log(100 \cdot k_1 \cdot \dots \cdot k_{10}) = log(100) + log(k_1) + \dots + log(k_{10})$
- Assume ks are independent and identically distributed
- Apply the Central Limit Theorem: log(out) is Normally distributed
- $\Rightarrow$ The outcome has a Lognormal(μ, σ) distribution.

Formalizing the story (Physics)

The storyline:
There are particles in a box. They bump into each other. How much heat does the box emit?

Formalizing the story (Physics)

The storyline:
There are particles in a box. They bump into each other. How much heat does the box emit?

The formalized model:
Assume N particles in a constrained subset of $\mathbb{R}^3$. They collide according to Newton's laws.

Formalizing the story (Economics)

The storyline:

- Capital (human, social, physical) accumulates with age; capital affects income.
- Gender affects income.

Formalizing the story (Economics)

The storyline:

- Capital (human, social, physical) accumulates with age; capital affects income.
- Gender affects income.

The formalized model (we'll get to gender in a second):

$$\ln(I) = \beta_0 + A\beta_1 + \epsilon \tag{1}$$

Why do women earn less?

Bosses are jerks:

Why do women earn less?

Bosses are jerks:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

Why do women earn less?

Bosses are jerks:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

They accumulate capital more slowly:

Why do women earn less?

Bosses are jerks:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

They accumulate capital more slowly:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

Why do women earn less?

Bosses are jerks:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

They accumulate capital more slowly:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \tag{2}$$

There's a risk that they'll produce babies:

Why do women earn less?

Bosses are jerks:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \qquad (2)$$

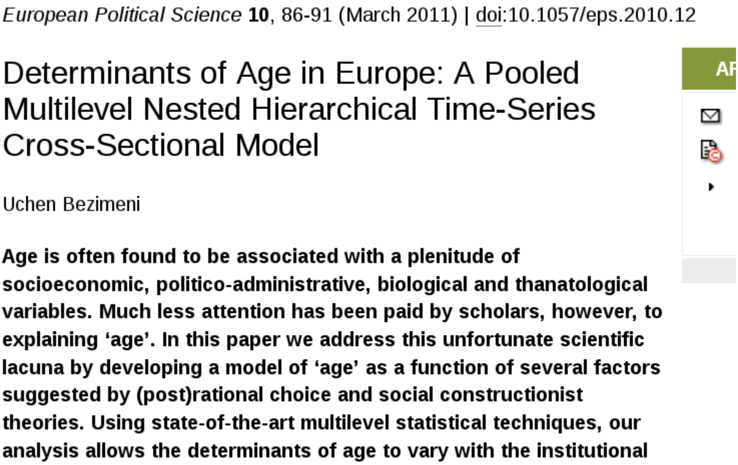
They accumulate capital more slowly:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \qquad (2)$$

There's a risk that they'll produce babies:

$$\ln(I) = \beta_0 + A\beta_1 + S\beta_2 + \epsilon \qquad (2)$$

Regression technology advances



European Political Science **10**, 86-91 (March 2011) | doi:10.1057/eps.2010.12

Determinants of Age in Europe: A Pooled Multilevel Nested Hierarchical Time-Series Cross-Sectional Model

Uchen Bezimeni

Age is often found to be associated with a plenitude of socioeconomic, politico-administrative, biological and thanatological variables. Much less attention has been paid by scholars, however, to explaining 'age'. In this paper we address this unfortunate scientific lacuna by developing a model of 'age' as a function of several factors suggested by (post)rational choice and social constructionist theories. Using state-of-the-art multilevel statistical techniques, our analysis allows the determinants of age to vary with the institutional

The regression model is not no structure

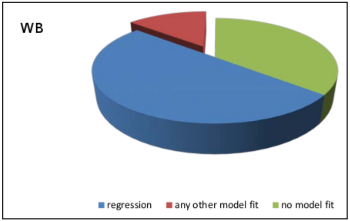
- If I'm k% older, expected log income is $\beta_1 \cdot k$% more, $\forall k$.
- Age and sex are structurally symmetric.

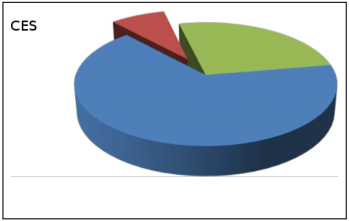
100 working papers

	WB	CES
papers with regressions only	25	33
papers with any other model	7	4
papers w/no model fitting	18	13

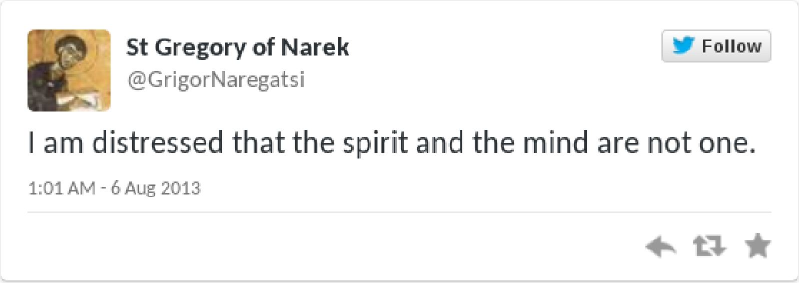
WB = World Bank

CES = Census Center for Economic Studies





Social science are the harder sciences; why are our models so much simpler?



St Gregory of Narek
@GrigorNaregatsi

I am distressed that the spirit and the mind are not one.

1:01 AM - 6 Aug 2013

Why do we do use proxy models we don’t believe?

- I do this.

Why do we do use proxy models we don’t believe?

- I do this.
- You can run a regression in under five minutes, and that may be good enough.

Why do we do use proxy models we don’t believe?

- I do this.
- You can run a regression in under five minutes, and that may be good enough.
- In the 1970s, it was the peak of computing power ⇒ well-researched models.

Why do we do use proxy models we don't believe?

- I do this.
- You can run a regression in under five minutes, and that may be good enough.
- In the 1970s, it was the peak of computing power ⇒ well-researched models.
- Nobody ever got fired for running a linear regression.

Other alternatives

- structural equation modeling
- Bayesian hierarchies
- graphical models
- causal networks

Sidebar: implementation

We have the technology

Sidebar: Apophenia

A library of stats functions.

- Pretty stable at this point.
- Open source.
- In process to be Debian, Fedora packages.

Sidebar: A sample program with Apophenia

```
#include <apop.h>

int main( ) {
    apop_data *data = apop_text_to_data("data.txt",
                                         .has_row_names='y');
    apop_model *est = apop_estimate(data, apop_ols);
    apop_model_show(est);
}
```

Side-sidebar: An essay on programming language choice

Side-sidebar: An essay on programming language choice

It's the vocabulary.

Side-sidebar: good enough for me

Side-sidebar: not good enough for you?

I still ♡ you. Implement the algebraic system in your favorite platform.

Side-sidebar: not good enough for you?

I still ♡ you. Implement the algebraic system in your favorite platform. Please.

The algebraic system

The idea

- There is a space of models, $\mathbb{M}$.
- Every transformation maps from $\mathbb{M} \to \mathbb{M}$, or $(\mathbb{M}, \mathbb{M}) \to \mathbb{M}$.

What is a model?

- At its core, a likelihood: $(\mathbb{D}, \mathbb{P}) \to \mathbb{R}^+$
  - $\mathbb{D} =$ a data space
  - $\mathbb{P} =$ a parameter space
    - Regression tree: $\mathbb{P} =$ the space of bifurcations.
    - "Nonparametric" model: $\mathbb{P}$ has infinite/indeterminate dimension.

What is a model?

- At its core, a likelihood: $(\mathbb{D}, \mathbb{P}) \to \mathbb{R}^+$
  - $\mathbb{D} =$ a data space
  - $\mathbb{P} =$ a parameter space
    - Regression tree: $\mathbb{P} =$ the space of bifurcations.
    - "Nonparametric" model: $\mathbb{P}$ has infinite/indeterminate dimension.
- In practice, we want more:
  - a method for estimating parameters from data
  - a method for making random draws
  - a method for testing claims

Back to that sample code

```
apop_model *est1 = apop_estimate(data, apop_ols);
apop_model *est2 = apop_estimate(data, apop_logit);
```

- OLS's estimation method is well known $[(X'X)^{-1}X'Y]$
- Logit doesn't have an `estimate` method. But given a likelihood, I can estimate via MLE.

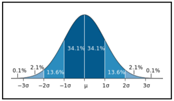
Given part of a model...

. . . we can build the whole thing.

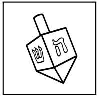
- Likelihood → Parameter estimation: MLE
- Likelihood → RNG: MCMC, ARMS
- RNG → Likelihood: build an empirical PMF
- RNG → CDF: Monte Carlo integration
- CDF → Likelihood: Numeric deltas

Now functions don't have to care what's inside the black box.

$y \sim$

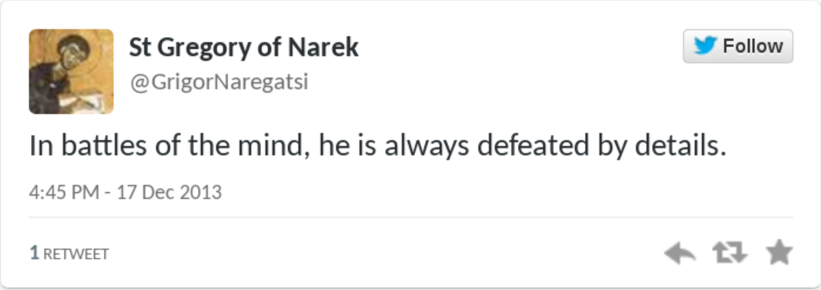


$x \sim$



$z \sim \boxed{\text{Update(prior=y, likelihood=x)}}$

Don't be oversold


St Gregory of Narek
@GrigorNaregatsi

In battles of the mind, he is always defeated by details.

4:45 PM - 17 Dec 2013

1 RETWEET

Here's a 49-page paper: `tinyurl.com/ModelSet`
Apophenia is ∼15,000 nontrivial lines of code.

Transformations

Truncation

- Maps a model to a truncated model; cut off everything less than x.
  - $\mathbb{D}$: same as before, but truncated
  - $\mathbb{P}$: same as before
  - Likelihood: drop all weight less than x; rescale
  - RNG: if draw $< x$, try again
  - CDF: Redistribute weight under cutoff
  - Estimation: You have a likelihood. Go fish.
- It's still a model $\in \mathbb{M}$.

Data-parameter composition

AKA Bayesian updating: $P(\beta|D) \propto P(D|\beta)P(\beta)$.

- Start with M_p, M_L, $\rho \in \mathbb{P}_p$, $D \in \mathbb{D}_L$.
- Draw p from the RNG for M_p given params ρ [$\propto P(\beta)$].
- Evaluate $P_L(D, p)$ using the likelihood from M_L [$P(D|\beta)$].
- Write this as $DPcompose(M_p, M_L) \in \mathbb{M}$.
- Likelihood of (D, ρ) depends on both model likelihoods.

Story problems

A waiting narrative

- People have different mean wait times.
- Their mean waiting time is a sum of iid glitches. What is the distribution of waiting times?
  - For now, say those glitches generate a $\mathcal{N}(3, 1)$
- But those with negative mean wait times are already out of the population

Encode the narrative

$$M_{wait} \equiv DPcompose\left(Trunc\left(\mathcal{N}(3,1)\right), \mathcal{E}xp\right)$$

```
//The constraint function.
double over_zero(apop_data *in, apop_model *m){
    return apop_data_get(in) > 0;
}

apop_model *norm = apop_model_set_parameters(apop_normal, 3, 1);
apop_model *orm = apop_model_dconstrain(.base_model=norm,
                                        .constraint=over_zero);
apop_model *posterior=apop_update(.data=wait_data,
                                  .prior=orm,
                                  .likelihood=apop_exponential);
```

Estimating with data

$$M_{wait} \equiv DPcompose\left(Trunc\left(\mathcal{N}(\mu, \sigma)\right), \mathcal{E}xp\right)$$

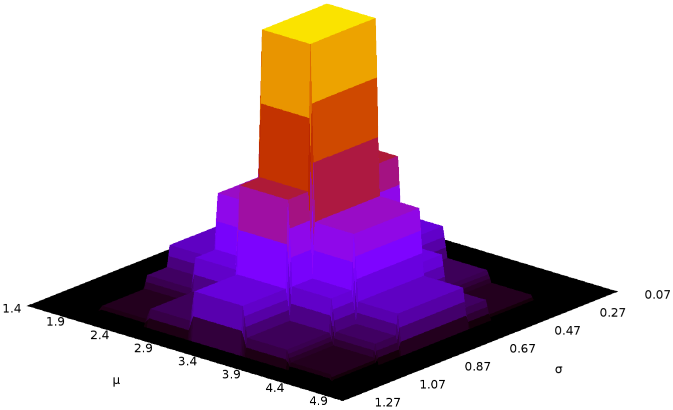
```
apop_model *rm = apop_model_dconstrain(.base_model=apop_normal,
                                        .constraint=over_zero);
apop_model *wait = apop_model_dpcompose(rm, apop_exponential);
apop_model *optimum = apop_estimate(data, wait);
apop_data *cov = apop_model_bootstrap_cov(data, wait);
```

More models, more transformations

- M_{PMF}
  - Probability Mass Function
  - $PMF(0) \equiv$ a point mass at zero
- JACOBI: invertible coordinate transformation
  - Box sides are $M_{\mathcal{N}} \Rightarrow$ box volume is $Jacobi_{x^3}(M_{\mathcal{N}})$
- MIX
  - Weighted sum of input models
- CROSS
  - Independent draws from input models. For example...

The Ziggurat distribution

```
#define unif(a, b) apop_model_set_parameters(apop_uniform, a, b)
apop_model *zig = apop_model_stack(
  apop_model_mixture(unif(3.0, 3.6), unif(2.5, 4.1), unif(2.0, 4.6)),
  apop_model_mixture(unif(.6, .9), unif(0.5, 1.0), unif(0.3, 1.2))
);
```



The dinner party

- Some try to be on time, but hit delays at λ/minute.
- Some are 'fashionably late': shoot for 30 minutes late, but are imprecise.
- Nobody is early. If early, hang out at the Rite Aid until the right time.

Build it!

- First group:

$$M_A = Jacobi_{x/\lambda}(M_{Exp})$$

- Second group:

$$M_B = Mix_*(Trunc(\mathcal{N}(\mu, \sigma)), PMF(0))$$

- Together: $M_C = Mix_w(M_A, M_B)$
  - $\mathbb{P} = \lambda, \mu, \sigma, w$
  - $\mathbb{D} =$ minutes late

Build it!

- First group:

$$M_A = Jacobi_{x/\lambda}(M_{Exp})$$

- Second group:

$$M_B = Mix_*(Trunc(\mathcal{N}(\mu, \sigma)), PMF(0))$$

- Together: $M_C = Mix_w(M_A, M_B)$
  - $\mathbb{P} = \lambda, \mu, \sigma, w$
  - $\mathbb{D} =$ minutes late
- Priors
  - λ : $Trunc(\mathcal{N}(\mu_1, \sigma_1))$
  - μ : $\mathcal{N}(\mu_2, \sigma_2)$
  - σ : $Jacobi_{\sqrt{x}}(M_{\chi^2})$
  - w : $Uniform(0, 1)$

The dinner party, built

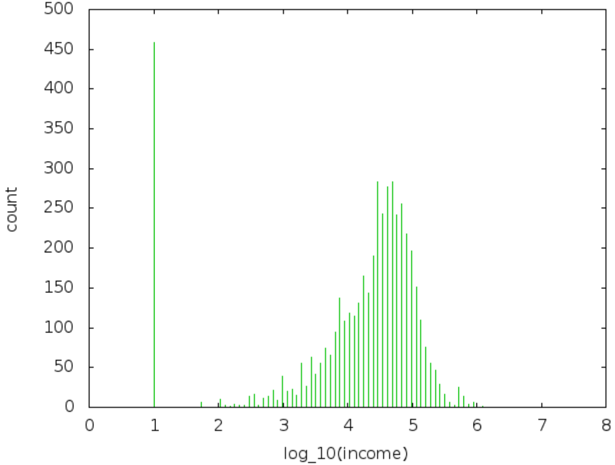
$$
\begin{aligned}
M = DPCompose(&\\
&Cross(\\
&\qquad Trunc(\mathcal{N}(\mu_1, \sigma_1)),\\
&\qquad \mathcal{N}(\mu_2, \sigma_2),\\
&\qquad Jacobi_{\sqrt{x}}(M_{\chi^2})\\
&\qquad Uniform(0, 1)\\
&),\\
&Mix_w(\\
&\qquad Jacobi_{1/\lambda}(M_{Exp}),\\
&\qquad Mix(Trunc(\mathcal{N}(\mu, \sigma)), PMF(0))\\
&)\\
)&
\end{aligned}
$$

Probabilistic programming implementation sidebar II

- That was a single expression.
- Algebraic transformations are the core of “functional” programming.
- Some statistics (e.g., *monoids*) are very amenable to this treatment.
- How can we adapt the functional paradigm to describe random processes?
    - BLOG
    - Church
    - Venture
    - HLearn for Haskell

Estimating incomes

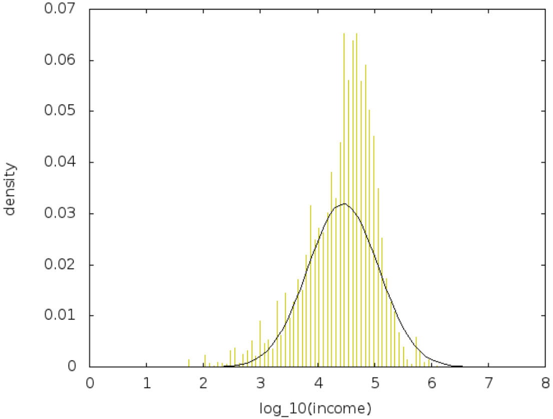
Income, DC Public use micro sample (PUMS)



- Ages 0-14 are N/A, excluded

Linear regression

$\log_{10}(I) = 4.406 + 0.005A - 0.119S + \epsilon$

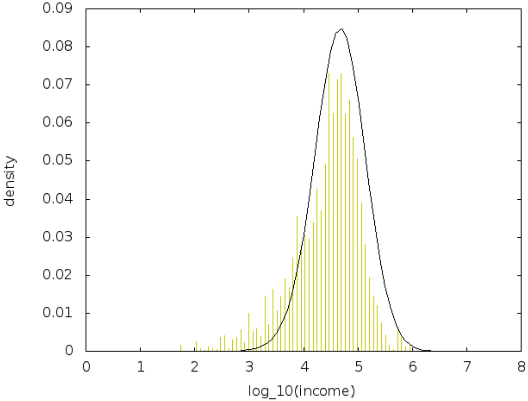


- $\mathbb{P}$: Model self-reports all βs significantly $\neq 0$.
- $\mathbb{D}$: $AIC_c = 451$

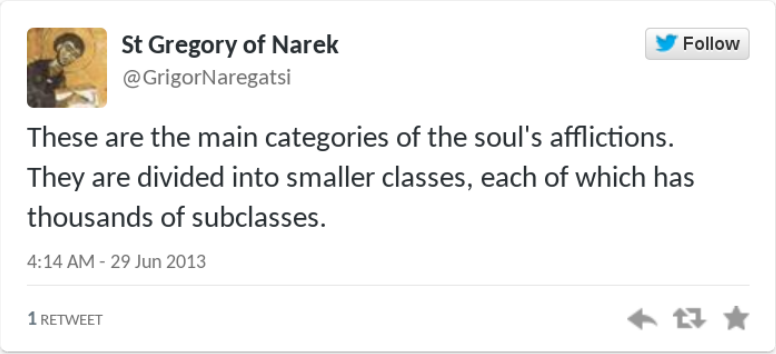
Ways to improve this

Keep adding elements to the linear combination:

var	β	p-val
1	4.29	(0.00)
age	0.01	(0.000)
sex	-0.10	(0.12)
citizen	0.005	(0.78)
hisp	-0.002	(0.74)
lang@home	-0.001	(0.007)
# autos	0.008	(0.86)
education	0.049	(0.000)
speak English	0.045	(0.41)



Other ways to improve this



St Gregory of Narek
@GrigorNaregatsi

These are the main categories of the soul's afflictions. They are divided into smaller classes, each of which has thousands of subclasses.

4:14 AM - 29 Jun 2013

1 RETWEET

A production model

- Agents have capital (human, social, physical)
    - Capital accumulates or decays each period

$$W_{t+1} = W_t d$$

    - d is iid
    - Run for ten periods.

A production model

- Agents have capital (human, social, physical)
  - Capital accumulates or decays each period

$$W_{t+1} = W_t d$$

  - d is iid
  - Run for ten periods.
- Each year, agents use the capital they have to get an income
  - Capital → income

$$ln(I) = W^{\beta}$$

A production model

- Agents have capital (human, social, physical)
  - Capital accumulates or decays each period

$$W_{t+1} = W_t d$$

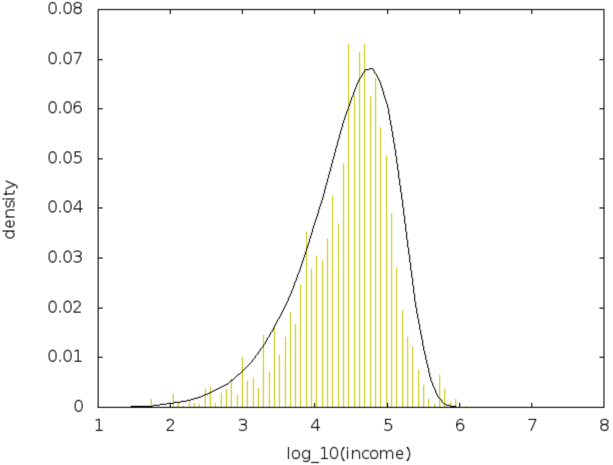
  - d is iid
  - Run for ten periods.
- Each year, agents use the capital they have to get an income
  - Capital → income

$$ln(I) = W^\beta$$

- $$ln(I) \sim Jacobi_{x^\beta}(M_{\mathcal{N}}(\mu, \sigma))$$

  - $\mathbb{D}$: income data (> 0)
  - $\mathbb{P} = [\mu, \sigma, \beta]$

Production model I: how'd we do?



- $\mathbb{P}$: AIC$=$ 1114 (vs. OLS: 451)

The bad luck model

- As before, capital accumulates or decays each period

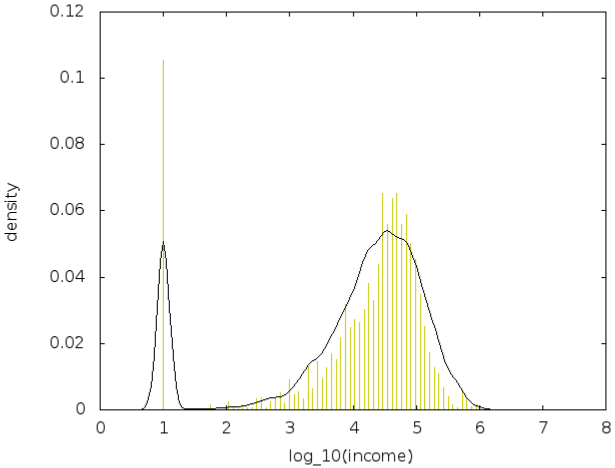
$$W_{t+1} = W_t d$$

- $d \sim \mathcal{N}(\mu, \sigma)$
- Run for 10 periods.

There is a cutoff k, below which we drop wealth to zero.

- $\mathbb{D}$: income data (≥ 0)
- $\mathbb{P}$: μ, σ, k

Production model II: how'd we do?

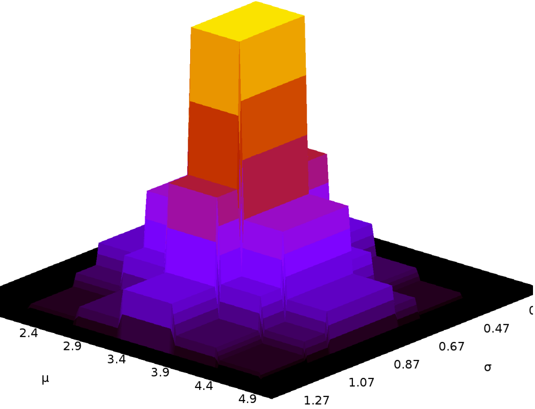


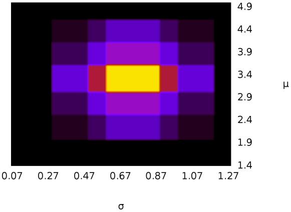
$\mathbb{P}: \mu = 3.03 \pm .002\sigma = 1.22 \pm .0024k = 1.12 \pm .007$

$\mathbb{D}: AIC_0 = 403$ (versus 451)

It’s not a complete model until there are priors

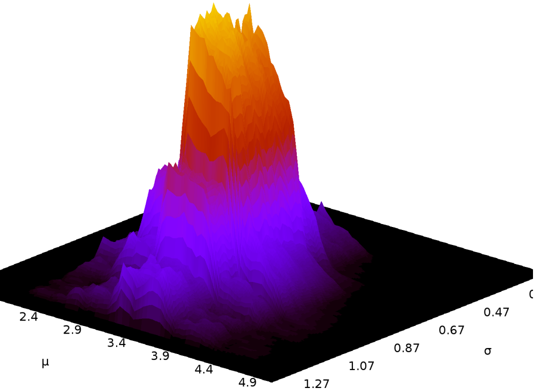
Add priors: `apop_model_stack(ziggurat, ziggurat, ziggurat)`

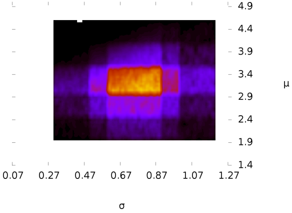




It’s not a complete model until there are priors

Update using the model given data: `apop_update(dc_data, ziggy, sim)`

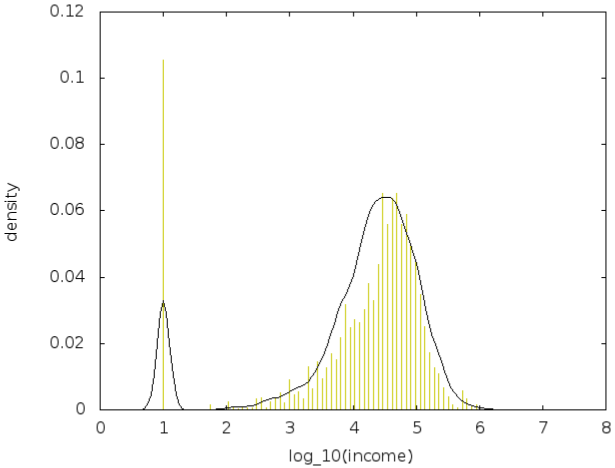




Production model III: does age matter?

- Same setup as Production model II, but:
- Agents evolve for p·(age-24)+ $(1 - p)$·10 periods.
    - $p = 0$: everybody gets ten draws.
    - $p = 1$: everybody gets (age-14) draws.

Production model III: how'd age do?



$\mathbb{P}$:$\mu = 3.06; \sigma = 0.93; k = 1.22;\ p = 0.036 \pm .002$

$\mathbb{D}$: $\mathrm{AIC}_0 = 436$ (versus PM1=403; OLS=451)

Why do women make less?

- Bosses are jerks
  - The model: Subtract some amount from women's income, post capital-to-income transformation.

Why do women make less?

- Bosses are jerks
  - The model: Subtract some amount from women's income, post capital-to-income transformation.
- They accumulate capital more slowly
  - The model: men draw from one capital shock distribution; women from another.

Why do women make less?

- Bosses are jerks
  - The model: Subtract some amount from women's income, post capital-to-income transformation.
- They accumulate capital more slowly
  - The model: men draw from one capital shock distribution; women from another.
- There's a risk that they'll produce babies
  - The model: add a glitch to the ABM that some women fall out of the game for some months.

Summary regarding the examples

- If your model doesn't fit, don't just throw more covariates in; ask whether the structure could better fit the narrative.
- We can grow models incrementally. Start with no parameters, add age, add sex, ...
    - Simulations can have comprehensible structures; don't have to have 1,000 parameters
- There can be a smooth transition from closed-form to open-form models.
- Evaluate all models equally:
    - $\mathbb{P}$: Put confidence intervals, priors on parameters.
    - $\mathbb{D}$: Compare model and actual distribution, measure information loss.

Conclusion

Calls to action

- Please be discontent with the Social Science models we have today.
    - Please model the story, not its shadow.
- Please build tools that allow the formalization of structure.
- Please write transformations as $\mathbb{M} \to \mathbb{M}$ or $(\mathbb{M}, \mathbb{M}) \to \mathbb{M}$, so we can develop models of arbitrary complexity.

終
end [fin]