Statistics Research and Careers

Sheridan Grant

Statistics PhD Candidate

slgstats@uw.edu

October 13, 2020

Where do I want to be, and what might it look like?

Jobs You're Considering (Statistically Speaking)

1. Statistics PhD

Jobs You're Considering (Statistically Speaking)

1. Statistics PhD
2. Data Scientist

Jobs You're Considering (Statistically Speaking)

1. Statistics PhD
2. Data Scientist
3. Other applied math/stats (clinical trials, econometrics, etc.)

Jobs You're Considering (Statistically Speaking)

1. Statistics PhD
2. Data Scientist
3. Other applied math/stats (clinical trials, econometrics, etc.)
4. Software
5. Other

Other applied math/stats

Energy Modeling at E3

- A mix of econometrics, engineering, and consulting roles
- "Predict the profit stream from this potential transmission line so we can decide if it's worth building"
- "Compile a dossier on environmental threats to the electric grid"

Other applied math/stats

Energy Modeling at E3

- A mix of econometrics, engineering, and consulting roles
- "Predict the profit stream from this potential transmission line so we can decide if it's worth building"
- "Compile a dossier on environmental threats to the electric grid"
- **"Predict the energy flows and prices in real time in the Western US for the next year"**

Other applied math/stats

Energy Modeling at E3

- A mix of econometrics, engineering, and consulting roles
- "Predict the profit stream from this potential transmission line so we can decide if it's worth building"
- "Compile a dossier on environmental threats to the electric grid"
- **"Predict the energy flows and prices in real time in the Western US for the next year"**

Electric Grid Optimization

Minimize total energy cost given demand time series, subject to these constraints:

- generator output $\in \{0\} \cup [250, 500] GW$
- transmission line capacity $\in [-250, 500] GW$
- generator startup cost $= 250,000$
- excess production cost $= 1000/GW$

Mixed integer-linear program is NP hard, need to balance accuracy vs. compute

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups
- Use linear models to describe conditional associations, summarize/interpret ML models w/r/t race

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups
- Use linear models to describe conditional associations, summarize/interpret ML models w/r/t race

R Tidyverse Applications

"Is race associated with home sale price within US counties?"

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups
- Use linear models to describe conditional associations, summarize/interpret ML models w/r/t race

R Tidyverse Applications

"Is race associated with home sale price within US counties?"

- Which block group and county was a home sale in? "Spatial join" home lat/lon with geographic polygons (`sf::st_join`)

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups
- Use linear models to describe conditional associations, summarize/interpret ML models w/r/t race

R Tidyverse Applications

"Is race associated with home sale price within US counties?"

- Which block group and county was a home sale in? "Spatial join" home lat/lon with geographic polygons (`sf::st_join`)
- Filter data to remove misleading transactions without biasing results (`dplyr::filter`)

Data Science

Algorithmic Fairness at Zillow

- Investigated ML models for racial disparities (cannot give details)
- Data sets: all home transactions in US over 5-year period, racial demographics for all 250K census block groups
- Use linear models to describe conditional associations, summarize/interpret ML models w/r/t race

R Tidyverse Applications

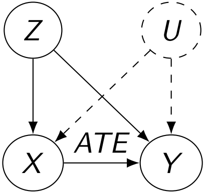
"Is race associated with home sale price within US counties?"

- Which block group and county was a home sale in? "Spatial join" home lat/lon with geographic polygons (`sf::st_join`)
- Filter data to remove misleading transactions without biasing results (`dplyr::filter`)
- Compute county- and block group-level sale averages (`dplyr::group_by`, `dplyr::summarize`)

Statistics PhD Research

Causal Inference

Given a causal graph:

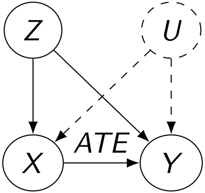


and observational data, *estimate* the Average Treatment Effect $ATE_{X \to Y}$.

Statistics PhD Research

Causal Inference

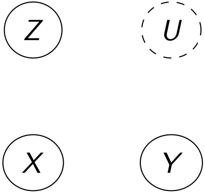
Given a causal graph:



and observational data, *estimate* the Average Treatment Effect $ATE_{X \to Y}$.

Causal Discovery

Given variables of interest and observational data,
fill in the graph:



Application: Causal Fairness in Peer Review

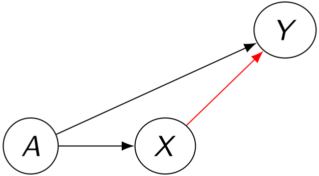


Figure: Forward review procedure

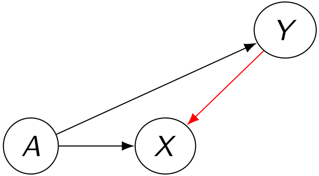


Figure: Backward review procedure

Goal: estimate direct effect of A (race) on Y (overall score) in the presence of X (criterion score)

Application: Causal Fairness in Peer Review

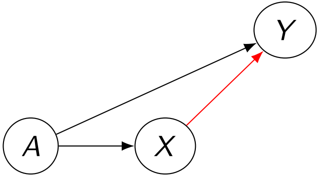


Figure: Forward review procedure

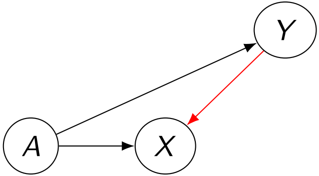


Figure: Backward review procedure

Goal: estimate direct effect of A (race) on Y (overall score) in the presence of X (criterion score)

- Fit $Y = \beta_0 + \beta_A A + \beta X + \epsilon$; β_A is direct effect of race under forward procedure

Application: Causal Fairness in Peer Review

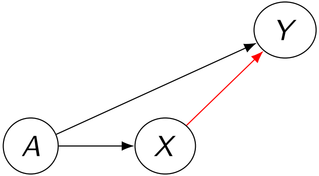


Figure: Forward review procedure

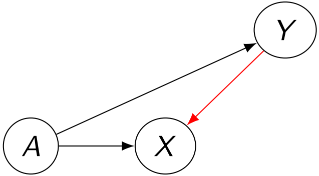


Figure: Backward review procedure

Goal: estimate direct effect of A (race) on Y (overall score) in the presence of X (criterion score)

- Fit $Y = \beta_0 + \beta_A A + \beta X + \epsilon$; β_A is direct effect of race under forward procedure
- Under backward procedure, however, need to fit $Y = \beta_0 + \beta_A A + \epsilon$ for β_A to be direct effect

Bayesian Model Selection

1. Consider models for the data $M_1, \ldots, M_k$

Bayesian Model Selection

1. Consider models for the data $M_1, \ldots, M_k$
2. You learn in the 340s that if these models are "nested," LRT helps you distinguish between them

Bayesian Model Selection

1. Consider models for the data $M_1, \ldots, M_k$
2. You learn in the 340s that if these models are "nested," LRT helps you distinguish between them
3. You learn in 435 to choose a model based on out-of-sample prediction

Bayesian Model Selection

1. Consider models for the data $M_1, \ldots, M_k$
2. You learn in the 340s that if these models are "nested," LRT helps you distinguish between them
3. You learn in 435 to choose a model based on out-of-sample prediction
4. What if the models aren't nested and you want to quantify how plausible each model is given some data $\{X, Y\}_n$?

Bayesian Model Selection

1. Consider models for the data $M_1, \ldots, M_k$
2. You learn in the 340s that if these models are "nested," LRT helps you distinguish between them
3. You learn in 435 to choose a model based on out-of-sample prediction
4. What if the models aren't nested and you want to quantify how plausible each model is given some data $\{X, Y\}_n$?

Use Bayes' Rule! $\mathrm{P}(M_k|\{X, Y\}_n) = \frac{\mathrm{P}(\{X,Y\}_n|M_k)\mathrm{P}(M_k)}{\mathrm{P}(\{X,Y\}_n)}$
We make an *assumption* about how cause and effect are related (a "model"), but the models M_k describe causal relationships.

(Nonlinear) Additive Noise Models (ANMs)

Identification

Suppose (WLOG)

$$
\begin{aligned}
&Y = f(X) + \epsilon, \\
&\epsilon \perp\!\!\!\perp X, \\
&f \text{ nonlinear.}
\end{aligned}
$$

Then $\nexists g, \eta$ such that

$$
\begin{aligned}
&X = g(Y) + \eta \\
&\eta \perp\!\!\!\perp Y
\end{aligned}
$$

(Hoyer et al. 2009).

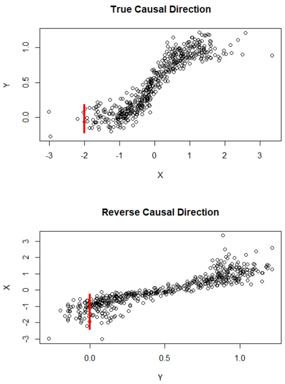


Figure: Data generated from a logistic ANM with Gaussian noise.

Nonlinearity and Identifiability

$\gamma\backslash\sigma_Y$	0.5	1	2
1/3	1.0 (11.4)	0.85 (2.2)	0.77 (0.9)
1/2	1.0 (8.2)	0.81 (1.4)	0.68 (0.6)
1	0.50 (0.1)	0.57 (0.0)	0.54 (0.1)
2	0.99 (14.8)	0.93 (3.8)	0.54 (0.3)
3	1.0 (31.0)	1.0 (8.5)	0.68 (0.8)

Table: Correct causal discovery rates (log average Bayes Factor) for Bayesian ANM over 100 replications and $n = 100$.

- For linear Gaussian model, Bayes Factors close to zero as expected
- For nonlinear models, greater nonlinearity and smaller noise yield higher Bayes Factors and greater accuracy
- Small sample size and model misspecification don't destroy ability to learn plausibility of identification assumption under each model

How do I decide? How do I get there?

General Advice

- There is a (partial) trade-off between *learning* new stuff, and *producing* new stuff. You may be eager to do the latter, but I recommend pushing harder on the former. You have plenty of time to produce stuff, but learning stuff gets harder.

General Advice

- There is a (partial) trade-off between *learning* new stuff, and *producing* new stuff. You may be eager to do the latter, but I recommend pushing harder on the former. You have plenty of time to produce stuff, but learning stuff gets harder.
- Learn more math than you "need." Learn only as much programming as you need. Any job you get will teach you to code, but opportunities to learn math are rare. Math will make you better at everything else, even if you never "use" it.

General Advice

- There is a (partial) trade-off between *learning* new stuff, and *producing* new stuff. You may be eager to do the latter, but I recommend pushing harder on the former. You have plenty of time to produce stuff, but learning stuff gets harder.
- Learn more math than you "need." Learn only as much programming as you need. Any job you get will teach you to code, but opportunities to learn math are rare. Math will make you better at everything else, even if you never "use" it.
- The world will show you what you're good at—you get to *help* decide. 5 years ago, I thought I would be a great mathematician. Turns out, I'm better at making statistics transparent and accessible.

What should I do when I graduate?

What should I do when I graduate?

- “I want a job”: sounds good, find a job

What should I do when I graduate?

- "I want a job": sounds good, find a job
- "I love nothing more than theoretical math and I don't care about money": sounds good, go to grad school

What should I do when I graduate?

- “I want a job”: sounds good, find a job
- “I love nothing more than theoretical math and I don’t care about money”: sounds good, go to grad school
- “I love nothing more than applied statistics and I don’t care about money”: do whatever, you’ll probably get money anyway

What should I do when I graduate?

- "I want a job": sounds good, find a job
- "I love nothing more than theoretical math and I don't care about money": sounds good, go to grad school
- "I love nothing more than applied statistics and I don't care about money": do whatever, you'll probably get money anyway
- "It all sounds good/I'm not sure": give yourself the chance to find out what you're good at. This may take time, money, or commitment—1 year at a job may not be enough (it wasn't for me), a PhD is a 5-year commitment, and a Master's degree isn't cheap.

What should I do when I graduate?

- “I want a job”: sounds good, find a job
- “I love nothing more than theoretical math and I don’t care about money”: sounds good, go to grad school
- “I love nothing more than applied statistics and I don’t care about money”: do whatever, you’ll probably get money anyway
- “It all sounds good/I’m not sure”: give yourself the chance to find out what you’re good at. This may take time, money, or commitment—1 year at a job may not be enough (it wasn’t for me), a PhD is a 5-year commitment, and a Master’s degree isn’t cheap.

Master’s degrees:

- Expensive ($100K or more for 2 years)
- Useful for progressing in salary/level in industry
- Useful for improving grad school resume, particularly for international students (UW Stats master’s explicitly preps you for PhD applications)