More interactions and generalized linear models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

April 5, 2021

Interaction example from last time

The ”Judgement of Princeton”

The Judgement of Princeton

- 9 judges, 20 wines
- Wines split between red and white, NJ or France
- Judges split between American or French/Belgium

Predictors:

- Wine color: red or white
- Wine origin: NJ or France
- Judge nationality: US or EU

Interactions

Potential for interactions between all predictors:

- Interaction between origin and judge: judge bias.
  Judge bias might depend upon color.
- Interaction between color and judge: taste preference.
  Taste preference might depend upon origin.
- Interaction between origin and color: relative advantage.
  Advantage might depend upon judge.

Let's build up some models for this data.

Generalized linear models

GLMs in a nutshell

Basic idea of a GLM:

- Want the mechanics of a linear regression, but outcomes aren't normally distributed
  - outcomes may be discrete/categorical
  - outcomes may have heavier tails than a normal distribution
- So, use an outcome distribution dependent on an expectation parameter $E[y]$ and model

$$g(E[y_i]) = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots$$

- What's g? The link function

Link functions

Link functions:

- Transform the linear model so that it takes on sensible values
- e.g., probabilities lie in $[0, 1]$, rates lie in $[0, \infty)$
- Most common include:
    - logit (common for binomial outcomes)
    - log (common for Poisson outcomes)
    - probit (similar to logit, but different tails)

Logistic regression

Most familiar GLM: logistic regression

- Binomial outcome, logit link
- Underlying parameter

$$y_i \sim \mathrm{Binomial}(p, n_i)$$
$$\mathrm{logit}(p) = \alpha + \boldsymbol{\beta} \cdot \mathsf{x}$$

Funding data for NWO grants

Example: funding data for NWO grants

- NWO (Dutch research council) awards funding to researchers in many fields
- We have a data set of application and approval counts for NWO grants, stratified by field and by applicant gender (in this data set, male or female)
- Research question: is there bias toward male applicants?

A simple model

A model:

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha_{\mathrm{gender}(i)} \\
\alpha &\sim \mathrm{Normal}(0, 2)
\end{aligned}
$$

Prior on α: quite vague, prefer log-odds between ± 4

A DAG

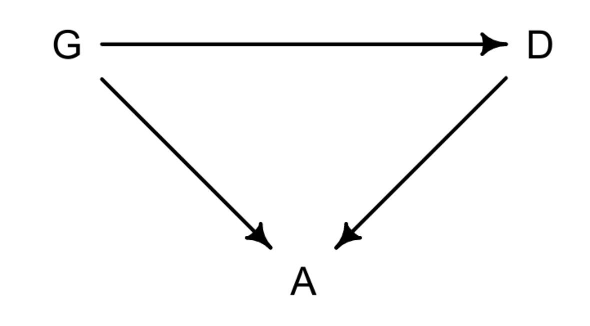
The computation suggests a noticeable gap between men and women: 3 percentage points on average, but with funding rates quite low, 3 percentage points is not so small.

But is this a direct causal effect, or mediated by an intermediate variable?

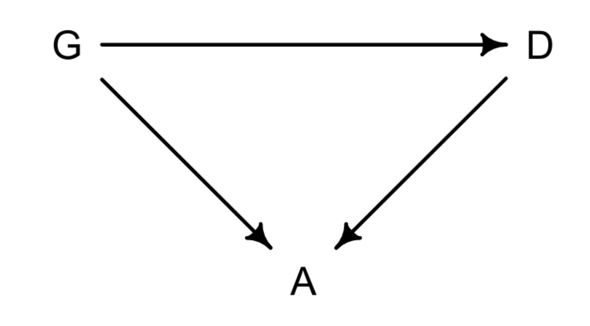
A DAG

The computation suggests a noticeable gap between men and women: 3 percentage points on average, but with funding rates quite low, 3 percentage points is not so small.

But is this a direct causal effect, or mediated by an intermediate variable?



A DAG



Two causal paths:

- Direct path $G \rightarrow A$
- Indirect path $G \rightarrow D \rightarrow A$

Previous model measured the two combined. Question about bias: is the direct effect nonzero?

A simple model

A model including discipline:

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha_{\mathrm{gender}(i)} + \beta_{\mathrm{discipline}(i)} \\
\alpha_j &\sim \mathrm{Normal}(0, 2) \\
\beta_j &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

A multilevel model

Since the number of applications varies widely across disciplines (almost a factor of 10 from the least (physics) to most (social sciences)), we can also introduce partial pooling:

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha_{\mathrm{gender}(i)} + \beta_{\mathrm{discipline}(i)} \\
\alpha_j &\sim \mathrm{Normal}(0, 2) \\
\beta_j &\sim \mathrm{Normal}(0, \tau) \\
\tau &\sim \mathrm{HalfCauchy}(5)
\end{aligned}
$$

What else?

We’ve tried:

- A simple GLM
- A multivariate GLM
- A multilevel GLM with partial pooling

What’s the last thing clearly missing on the posterior predictive plots?

What else?

We've tried:

- A simple GLM
- A multivariate GLM
- A multilevel GLM with partial pooling

What's the last thing clearly missing on the posterior predictive plots?

- Effect of gender conditional on discipline

Summary

Today:

- Interaction wrap-up
- GLM intro

Next time:

- Multilevel regression
- Assembling more complex models