# More interactions and generalized linear models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 18, 2021

# 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)

# Poisson regression

## Salamander counting

- Salamanders like to hide from predators – they want lots of forest coverage
- The average age of trees in the forest might have an effect?
- Predict salamander count as a function of coverage

# Salamander counting

$$
\begin{aligned}
y_i &\sim Poisson(\lambda_i) \\
\log \lambda_i &= \alpha + \beta C_i \\
\alpha &\sim ? \\
\beta &\sim ?
\end{aligned}
$$

- Salamander count distributed as Poisson RV
- $C_i$ – forest coverage
- Log link – $\lambda$ is a rate, must be positive
- Priors on model coefficients to be determined

# Salamander counting

```
with pm.Model() as cover_model:
    alpha = pm.Normal('alpha', 0, 1)
    bcov = pm.Normal('bcov', 0, 1)

    rate = pm.Deterministic('rate',
                            pm.math.exp(alpha + bcov * salamanders['PCTCOVER']))

    y_ = pm.Poisson('y', rate, observed = salamanders['SALAMAN'])
    cover_trace = pm.sample()
    cover_repl = pm.sample_posterior_predictive(cover_trace)
```

# Necessary sermon on priors

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \beta C_i \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta &\sim \text{Normal}(0, 1)
\end{aligned}
$$

| SITE | SALAMAN | PCTCOVER | FORESTAGE |
|---|---|---|---|
| 1 | 13 | 85 | 316 |
| 2 | 11 | 86 | 88 |
| 3 | 11 | 90 | 548 |
| 4 | 9 | 88 | 64 |
| 5 | 8 | 89 | 43 |

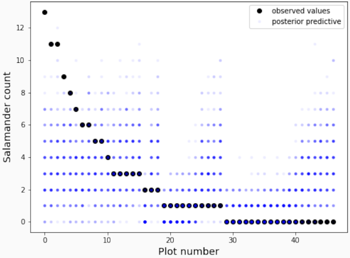
# After standardizing

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \beta C_i \\
\alpha &\sim \mathrm{Normal}(0, 1) \\
\beta &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

| | SALAMAN | PCTCOVER | FORESTAGE |
|---|---|---|---|
| **SITE** | | | |
| **1** | 13 | 0.85 | 0.760627 |
| **2** | 11 | 0.86 | -0.417586 |
| **3** | 11 | 0.90 | 1.959512 |
| **4** | 9 | 0.88 | -0.541609 |
| **5** | 8 | 0.89 | -0.650129 |

# Posterior predictive check

| | mean | sd | hdi_3% | hdi_97% |
|---|---|---|---|---|
| **alpha** | -0.897 | 0.333 | -1.489 | -0.272 |
| **bcov** | 2.524 | 0.403 | 1.796 | 3.254 |




- Model detects a strong influence of coverage on salamander rate
- Posterior predictive check doesn't look great

# Back to the DAG

As soon as we're considering using both variables, let's write a DAG:

# Model with two variables

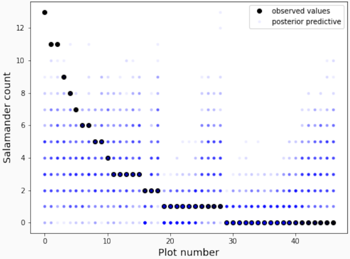
```
with pm.Model() as age_model:
    alpha = pm.Normal('alpha', 0, 1)
    bcov = pm.Normal('bcov', 0, 1)
    bage = pm.Normal('bage', 0, 1)

    rate = pm.Deterministic('rate',
                            pm.math.exp(alpha
                                        + bage * salamanders['FORESTAGE']
                                        + bcov * salamanders['PCTCOVER']))

    y_ = pm.Poisson('y', rate, observed = salamanders['SALAMAN'])
    age_trace = pm.sample()
    age_repl = pm.sample_posterior_predictive(age_trace)
```

# Posterior predictive check

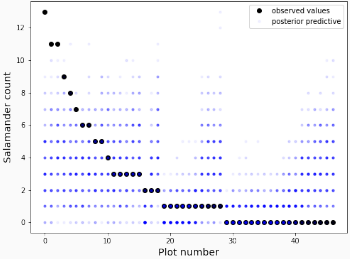
| | mean | sd | hdi_3% | hdi_97% |
|---|---|---|---|---|
| **bcov** | 2.432 | 0.432 | 1.615 | 3.237 |
| **bage** | 0.050 | 0.098 | -0.135 | 0.226 |




- Not much better, though we do get estimates for slopes

# Posterior predictive check

| | mean | sd | hdi_3% | hdi_97% |
|---|---|---|---|---|
| **bcov** | 2.432 | 0.432 | 1.615 | 3.237 |
| **bage** | 0.050 | 0.098 | -0.135 | 0.226 |



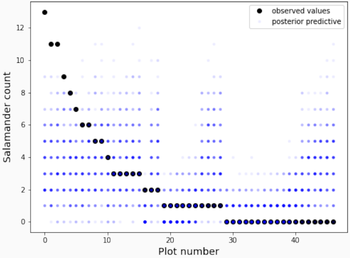

- Not much better, though we do get estimates for slopes
- Try looking at an interaction?

# Model with interactions

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \beta_C C_i + \beta_A A_i + \beta_{C,A} A_i C_i \\
\alpha &\sim \mathrm{Normal}(0, 1) \\
\beta_A &\sim \mathrm{Normal}(0, 1) \\
\beta_C &\sim \mathrm{Normal}(0, 1) \\
\beta_{C,A} &\sim \mathrm{Normal}(0, 0.5)
\end{aligned}
$$

# Posterior predictive check

| | mean | sd | hdi_3% | hdi_97% |
|---|---|---|---|---|
| **bcov** | 2.297 | 0.468 | 1.390 | 3.152 |
| **bage** | 0.317 | 0.366 | -0.402 | 0.995 |
| **bint** | -0.314 | 0.409 | -1.105 | 0.467 |

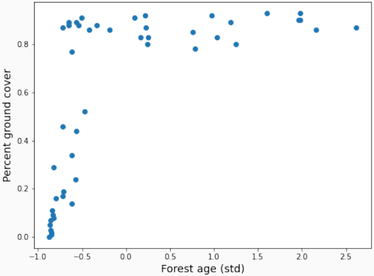



- Effect of age is positive, cover positive
- Interaction effect negative – how can we interpret this coefficient?

# Model with interactions

Another question about this interaction:

- what is the relationship between forest age and forest cover?



# A new DAG

Including our new variable in the DAG:

# Including the new variable in the model

Separate out parameters for the two clusters:

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha_{\mathrm{BURNED}[i]} + \beta_{C,\mathrm{BURNED}[i]} C_i + \beta_{A,\mathrm{BURNED}[i]} A_i \\
\alpha_i &\sim \mathrm{Normal}(0, 1) \\
\beta_C &\sim \mathrm{Normal}(0, 1) \\
\beta_A &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

## Varying intercepts

- We still don't have a great predictive model here – posterior predictions don't resemble real data
- Inference: there is some difference between plots that depends on unobserved variables
- Add parameters to account for these unobserved variables
    - Let each forest plot get its own intercept $\alpha_i$

# Overdispersion in Poisson regression

This is a common issue in Poisson models:

- Poisson distribution depends only on a rate parameter $\lambda$
- Mean and variance are both equal to $\lambda$
- If there is more variation than the Poisson allows (overdispersion), we're stuck
- The *offsets* $\alpha_i$ allow for this extra variation
- Think of this as letting the likelihood actually be a mixture of Poissons with differing $\lambda$

# Varying intercepts need regularization

Danger of varying intercepts:

- Particularly when we have a single observation per intercept
- Model equation is overdetermined:

$$\log \lambda_i = \alpha_i + \beta_C C_i + \beta_A A_i$$

- If $\alpha_i$ is allowed a lot of freedom, then just set $\beta_C = \beta_A = 0$ and fit each $\alpha_i$

Solution: regularize

# Add some multilevel structure

For regularization, we'll add multilevel structure

- Varying intercepts $\alpha_i$ account for overdispersion
- $\alpha_i$ drawn from a common distribution
- Allow the mean, SD of this common distribution to be learned

# Multilevel model

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \beta_C C_i + \beta_A A_i \\
\alpha_i &\sim \mathrm{Normal}(\mu, \sigma) \\
\beta_C &\sim \mathrm{Normal}(0, 1) \\
\beta_A &\sim \mathrm{Normal}(0, 1) \\
\mu &\sim \mathrm{Normal}(0, 1) \\
\sigma &\sim \mathrm{Exponential}(1)
\end{aligned}
$$

# Putting it all in the pot

We can combine our multilevel model with all the other features:

- Interactions between age and cover
- Segregation into two clusters (interaction between burned-ness and effects)

Let's go take a look...

# Binomial regression

# Logistic regression

Most familiar GLM: binomial regression (aka logistic regression)

- Binomial outcome, logit link
- Underlying parameter

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

- In PyMC3, use `pm.math.invlogit`

## 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 $\alpha$: quite vague, prefer log-odds between $\pm 4$

# A DAG

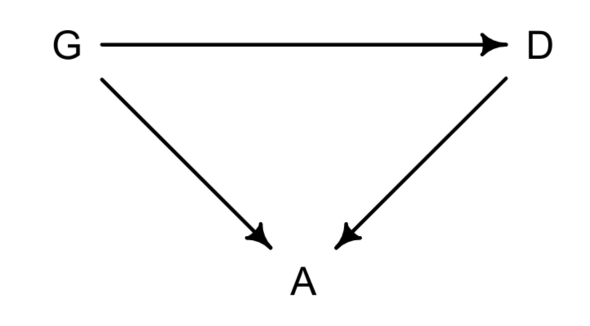
The computation suggests a noticeable gap between men and women: 3 percentage points on average, but with funding rates quite low, 3 percentage points is not so small.

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

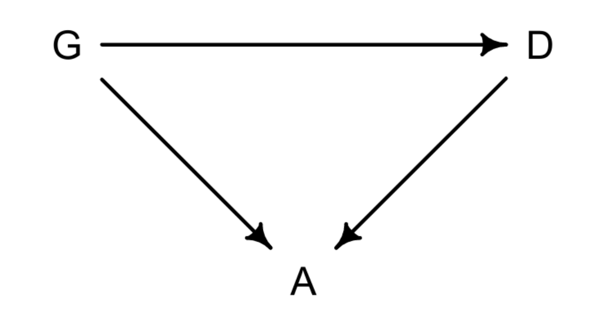
# A DAG

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 $G \to A$
- Indirect path $G \to D \to A$

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:

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

## Summary

Today:

- GLM intro
- Poisson regression example

Next up:

- Multilevel linear regression
- Assembling more complex models