Today

- Individual project requirements
  - 00_individual_project.md
- Simulating data and study designs
- Thursday
  - precorded lecture
  - individual meetings
- Next week
  - longer individual meetings

Simulating data

- Gain understanding
  - science, design, algorithm
- Test for correct setup
  - math, code, recover known parameters
- Explore study design
  - how many reps? etc
- Does the fitted model look right?
  - generate data like the real data?

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \, \text{latitude}_j$$

Data story: pseudocode

for each site j
    latitude determines broad-scale richness
    but there is some stochasticity about this

for each plot i
    habitat modifies local richness
    then stochasticity determines the final richness

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data story: pseudocode

for each site j
    latitude determines broad-scale richness
    but there is some stochasticity about this

for each plot i
    habitat modifies local richness
    then stochasticity determines the final richness

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data story: pseudocode

for each site j
    latitude determines broad-scale richness
    but there is some stochasticity about this

for each plot i
    habitat modifies local richness
    then stochasticity determines the final richness

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data story: pseudocode

for each site j
    latitude determines broad-scale richness
    but there is some stochasticity about this

for each plot i
    habitat modifies local richness
    then stochasticity determines the final richness

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data generating algorithm

for each site j
    generate expected ln(richness) by latitude
    generate stochasticity about this

for each plot i
    generate expected ln(richness) by habitat
    generate richness with stochasticity

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data generating algorithm

for each site j
    generate expected ln(richness) by latitude
    generate stochasticity about this

for each plot i
    generate expected ln(richness) by habitat
    generate richness with stochasticity

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data generating algorithm

for each site j
    generate expected ln(richness) by latitude
    generate stochasticity about this

for each plot i
    generate expected ln(richness) by habitat
    generate richness with stochasticity

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}\left(\ln(\mu_j), \sigma_\alpha^2\right)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Data generating algorithm

for each site j
    generate expected ln(richness) by latitude
    generate stochasticity about this

for each plot i
    generate expected ln(richness) by habitat
    generate richness with stochasticity

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \, \text{latitude}_j$$

Vectorized R code

```r
# For each site, generate an expected ln(richness) based on latitude
mu_alpha <- b_0 + b_2 * latitude

# For each site, generate stochasticity around this expectation (Eq.
# (note how this value will be the same for both plots at a site)
alpha <- rnorm(22, mu_alpha, sigma_alpha)

# For each plot, generate an expected ln(richness) based on habitat
# (we use j to extract the appropriate alpha and latitude values)
ln_mu <- alpha[j] + b_1 * forest + b_3 * forest * latitude[j]

# For each plot, generate richness with stochasticity (Eq. line 1)
# (we use the inverse link function to obtain mu)
y <- rpois(44, exp(ln_mu))
```

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Vectorized R code

```r
# For each site, generate an expected ln(richness) based on latitude
mu_alpha <- b_0 + b_2 * latitude

# For each site, generate stochasticity around this expectation (Eq.
# (note how this value will be the same for both plots at a site)
alpha <- rnorm(22, mu_alpha, sigma_alpha)

# For each plot, generate an expected ln(richness) based on habitat
# (we use j to extract the appropriate alpha and latitude values)
ln_mu <- alpha[j] + b_1 * forest + b_3 * forest * latitude[j]

# For each plot, generate richness with stochasticity (Eq. line 1)
# (we use the inverse link function to obtain mu)
y <- rpois(44, exp(ln_mu))
```

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Vectorized R code

```r
# For each site, generate an expected ln(richness) based on latitude
mu_alpha <- b_0 + b_2 * latitude

# For each site, generate stochasticity around this expectation (Eq.
# (note how this value will be the same for both plots at a site)
alpha <- rnorm(22, mu_alpha, sigma_alpha)

# For each plot, generate an expected ln(richness) based on habitat
# (we use j to extract the appropriate alpha and latitude values)
ln_mu <- alpha[j] + b_1 * forest + b_3 * forest * latitude[j]

# For each plot, generate richness with stochasticity (Eq. line 1)
# (we use the inverse link function to obtain mu)
y <- rpois(44, exp(ln_mu))
```

Ants Poisson GLMM

Plot $y_i \sim \text{Poisson}(\mu_i)$

$$\ln(\mu_i) = \alpha_{j[i]} + \beta_1 \text{forest}_i + \beta_3 \text{forest}_i \times \text{latitude}_{j[i]}$$

Site $\alpha_j \sim \text{Normal}(\ln(\mu_j), \sigma_\alpha^2)$

$$\ln(\mu_j) = \beta_0 + \beta_2 \text{ latitude}_j$$

Vectorized R code

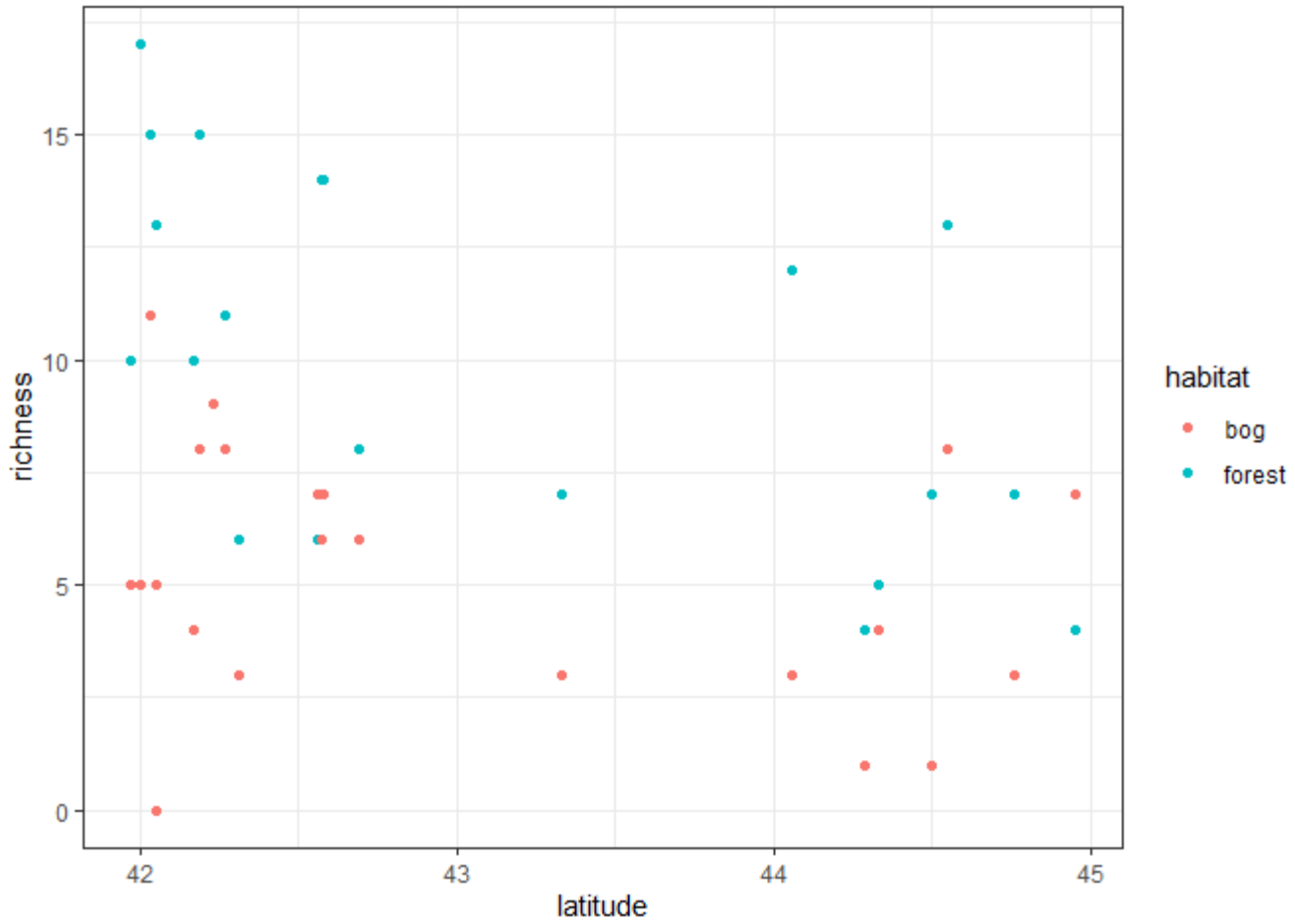
```r
# For each site, generate an expected ln(richness) based on latitude
mu_alpha <- b_0 + b_2 * latitude

# For each site, generate stochasticity around this expectation (Eq.
# (note how this value will be the same for both plots at a site)
alpha <- rnorm(22, mu_alpha, sigma_alpha)

# For each plot, generate an expected ln(richness) based on habitat
# (we use j to extract the appropriate alpha and latitude values)
ln_mu <- alpha[j] + b_1 * forest + b_3 * forest * latitude[j]

# For each plot, generate richness with stochasticity (Eq. line 1)
# (we use the inverse link function to obtain mu)
y <- rpois(44, exp(ln_mu))
```

Parameter values

- What values?
- Depends on what you want to do
- Ballparks are sufficient for most things
  - exploring study design
  - testing algorithms
  - scenarios

Simulated data

Code: ants_simulated_data.md



Does it work?

- ants_simulated_data.md
- recover parameter values?
- how reliable?

What to do when simulated data goes wrong?

Three possibilities

1. Math wrong
2. Simulation wrong
   - code error (bug) or translation of math to code
3. Fit wrong
   - training algorithm issue or model code wrong

Strategy: build from simple models

1. Make it as simple as possible at first
2. Build complexity incrementally