Today

- Tidy data
  - reshaping with tidyr
  - data manipulation with dplyr
- Ants GLM, Bayesian
  - ulam: equations, incl priors
  - working with posterior samples
- Model matrix

Independent project

- Homework: ideas submitted today
- I'll look at your ideas tomorrow
- Questions?
- Also chat after class

Tidy data

Variables in columns

Observations in rows

Values in cells

pivot_longer() - tidy a variable that is in multiple columns
pivot_wider() - tidy an observation that is in multiple rows

Base R: reshape(), stack(), unstack(), strsplit(), paste()

dplyr - working with data

filter() - pick observations by their values
select() - pick columns by name
arrange() - reorder rows
mutate() - create new variables from existing variables
summarize() - collapse values to a summary statistic
group_by() - all the above split by group

|> pipe to combine (or %>%)

Base R: subset(), order(), sort(), table(), aggregate(), |>

tibbles (tables with NZ accent)

Data frame; optimized printing for large datasets.
You probably spent a lot of time collecting data.
Wouldn't you want to spend a few minutes to inspect each row and column? That's how you catch errors.

Convert to dataframe

```
my_tibble |> data.frame()
```

Keep tibble, but print all the data by default

```
options(tibble.width=Inf)
options(tibble.print_max=Inf)
options(max.print=1500)
```

`?print.tbl` - see options

Other ways to inspect tibbles

`View(my_tibble)` – only works in Rstudio

`glimpse(my_tibble)` - summary

dplyr vs base

How many trees with known mortality status are missing a diameter in 2013?

```
tree_dat |>
    filter(year == 2013) |>
    filter(!is.na(mortality)) |>
    mutate(diam_missing=is.na(diam)) |>
    summarize(sum(diam_missing))

sum(is.na(subset(tree_dat, year == 2013 &
    !is.na(mortality))$diam))
```

dlpyr, complex queries

How many species observed per habitat fragment in each year per treatment?

```
beetle_species |>
    group_by(species, year, fragment,
             treatment) |>
    summarize(mean_abun = mean(abundance)) |>
    mutate(present = mean_abun > 0) |>
    group_by(year, fragment, treatment) |>
    summarize(richness = sum(present)) |>
    group_by(year, treatment) |>
    summarize(mean_richness = mean(richness))
```

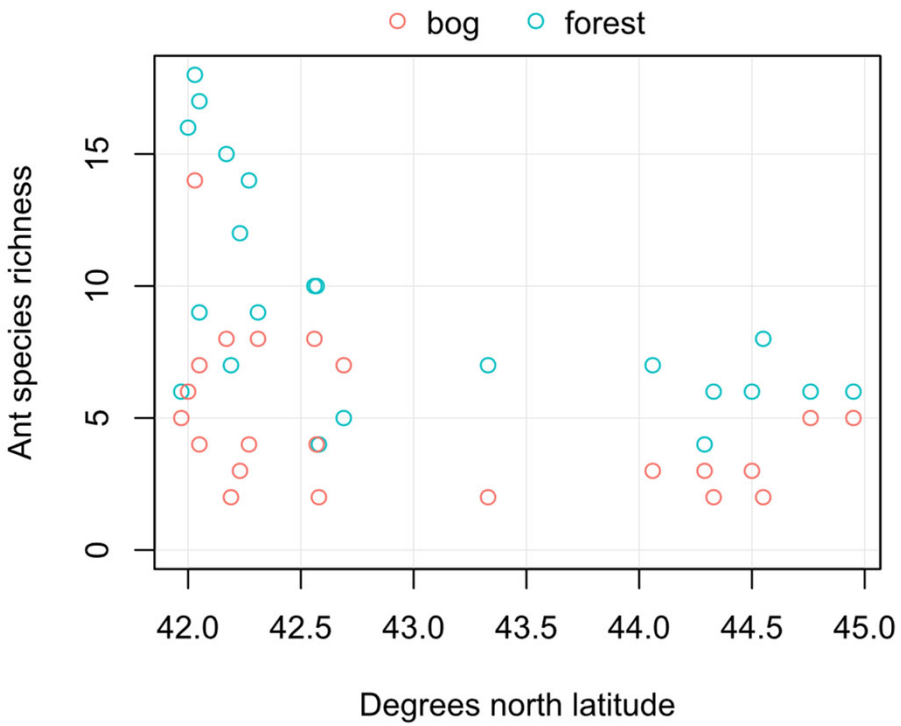
Most common models

Normal + Identity link	Poisson + Log link	Binomial + Logit link
$y_i \sim \text{Normal}(\mu_i, \sigma)$	$y_i \sim \text{Poisson}(\mu_i)$	$y_i \sim \text{Binomial}(\mu_i, n)$
$\mu_i = \beta_0 + \beta_1 x_i$	$\log(\mu_i) = \beta_0 + \beta_1 x_i$	$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i$

Key properties:

y: -∞ to ∞, continuous *μ*: -∞ to ∞, continuous	*y*: 0 to ∞, discrete, integer *μ*: 0 to ∞, continuous	*y*: 0, 1, discrete, binary *μ*: 0 to 1, probability

Which makes the most sense?



Poisson
+
Log link

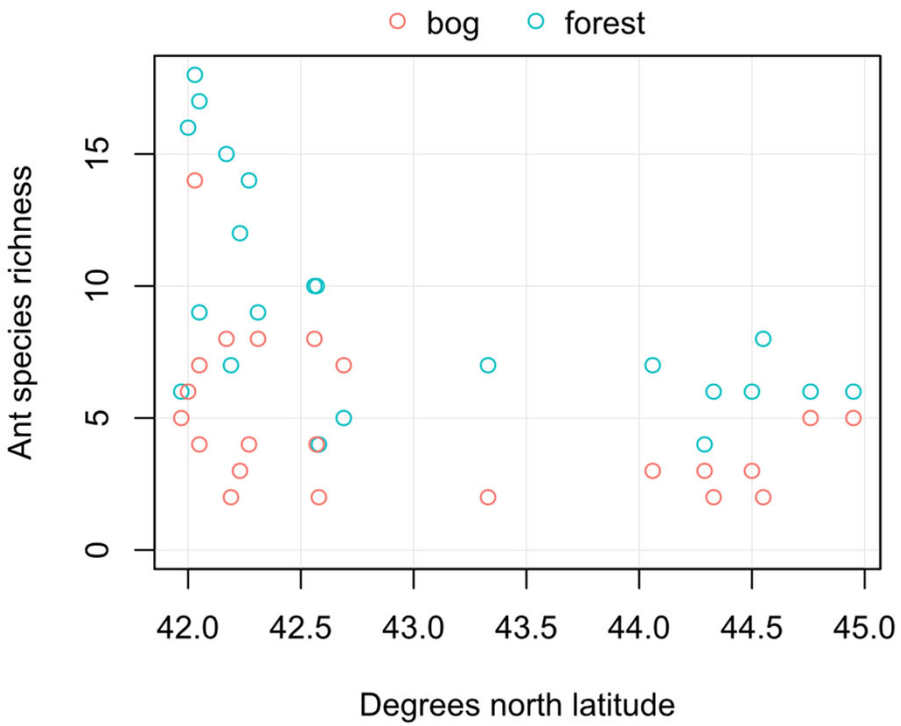
$$y_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \ldots$$

$$\eta_i = \beta_0 + \beta_1 x_i + \ldots$$

$$\mu_i = e^{\eta_i}$$

Write the full linear predictor



Poisson
+
Log link

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

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \ldots$$

$$\eta_i = \beta_0 + \beta_1 x_i + \ldots$$

$$\mu_i = e^{\eta_i}$$

$$\eta_i = \beta_0 + \beta_1 forest_i + \beta_2 latitude_i + \beta_3 forest_i \times latitude_i$$

Model matrix (design matrix)

Data

habitat	latitude	richness
forest	42	16
forest	42.56	10
forest	43.33	7
forest	44.76	6
bog	42.17	8
bog	42.57	4
bog	44.06	3
bog	44.95	5

$$\eta_i = \beta_0 + \beta_1 forest_i + \beta_2 latitude_i + \beta_3 forest_i \times latitude_i$$

Model matrix (design matrix)

Data

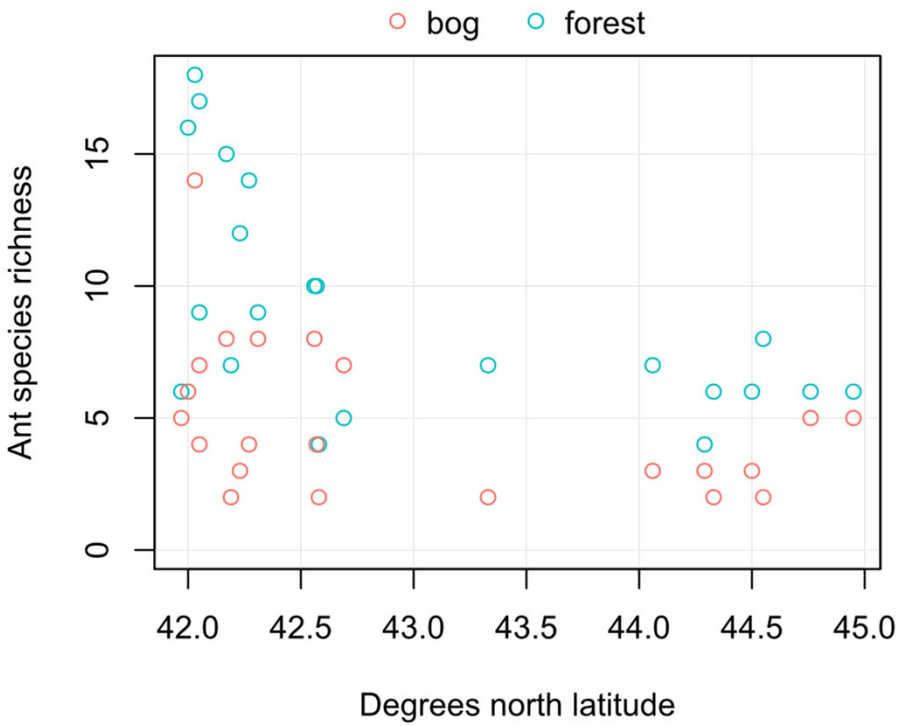
habitat	latitude	richness
forest	42	16
forest	42.56	10
forest	43.33	7
forest	44.76	6
bog	42.17	8
bog	42.57	4
bog	44.06	3
bog	44.95	5

Model matrix

intercept	forest	latitude	forest:latitude
1	1	42	42
1	1	42.56	42.56
1	1	43.33	43.33
1	1	44.76	44.76
1	0	42.17	0
1	0	42.57	0
1	0	44.06	0
1	0	44.95	0

$$\eta_i = \beta_0 intercept_i + \beta_1 forest_i + \beta_2 latitude_i + \beta_3 forest_i \times latitude_i$$

Bayesian GLM: Priors?



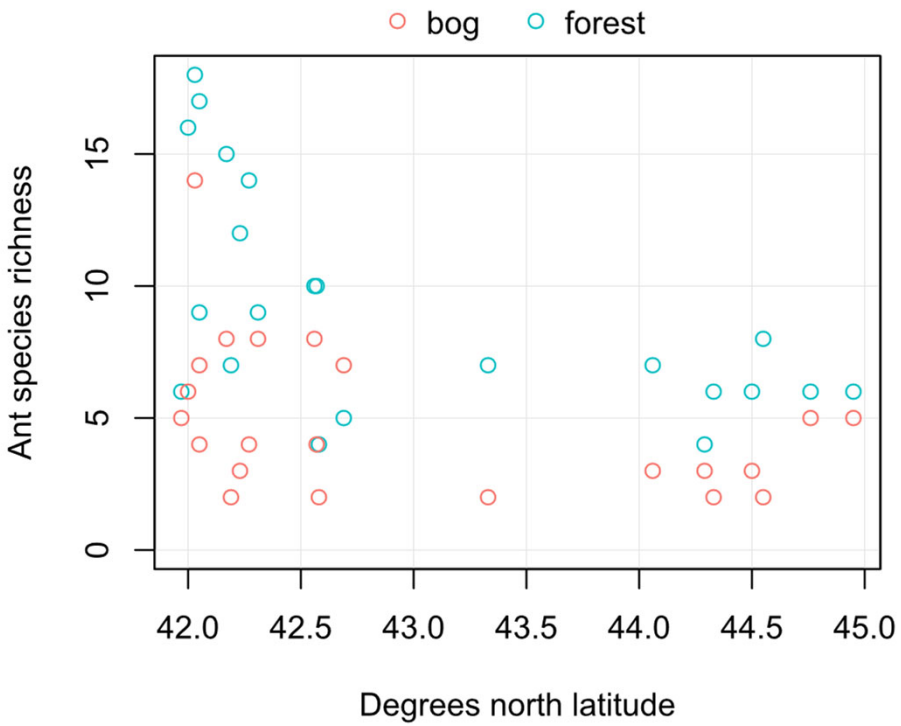
Poisson
+
Log link

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

$$\log(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 forest_i + \beta_2 latitude_i + \beta_3 forest_i \times latitude_i$$

Bayesian GLM: Code



Poisson
+
Log link

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

$$\log(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 forest_i + \beta_2 latitude_i + \beta_3 forest_i \times latitude_i$$

Write the code for ulam()