Today

- Finish off model checking
- Data manipulation with dplyr
- Generalized linear models (GLM)
  - McElreath Ch 9

Independent project

- Complete analysis (EDA through inference & conclusions)
- ggplot, dplyr
- Preferably hierarchical model:
  - rstanarm: stan_glmer or stan_lmer
- Submit .md from .R or .Rmd
- Due end of semester

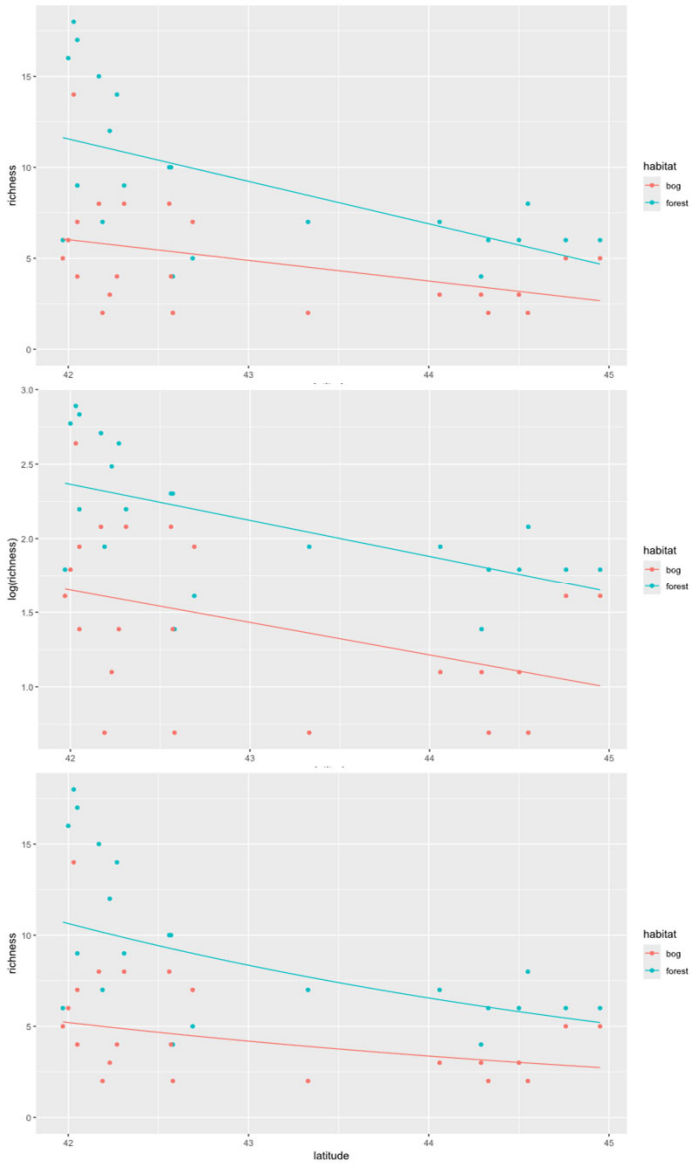
Model checking

- When is normality of errors important?
  - Not critical for inference about means
    - Frequentist: sampling distribution will still be approximately normal
    - Bayesian: posterior distribution will still be approximately normal or insensitive
  - Can be important for inference about prediction
    - Because: data generating process
- Good information for improving model

Model checking

- How can I know what to look for in a diagnostic plot?
  - simulate it!
- Continue coding demo

How to proceed?



Linear

Log-linear

Nonlinear

Scientific questions:

How does species richness vary with latitude?

Is this relationship different between habitats?

How different is species richness between habitats?

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 in each habitat fragment in each year?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))
```

Main points McElreath Ch 9

- Generalized linear models

Data

Parameters

$$y_i \sim \text{Distribution}(\mu_i, \ldots)$$

$$g(\mu_i) = \eta_i = \underbrace{\beta_0 + \beta_1 x_i}$$

Link function

Linear predictor
"eta"

Linear model

Main points McElreath Ch 9

- Generalized linear models

Data

Parameters

$$y_i \sim \text{Distribution}(\mu_i, \ldots)$$

$$g(\mu_i) = \eta_i = \underbrace{\beta_0 + \beta_1 x_i}$$

Link function

Linear predictor "eta"

Linear model

or

$$\mu_i = g^{-1}(\beta_0 + \beta_1 x_i)$$

Inverse link function

Main points McElreath Ch 9

- Exponential family (some)
  - Exponential, Gamma, Normal, Poisson, Binomial
- Other distributions
  - with write-your-own or Bayesian, this doesn't have to be from the exponential family
  - even more generalized!
- Link functions (some)
  - identity, log, logit

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

Most common models

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

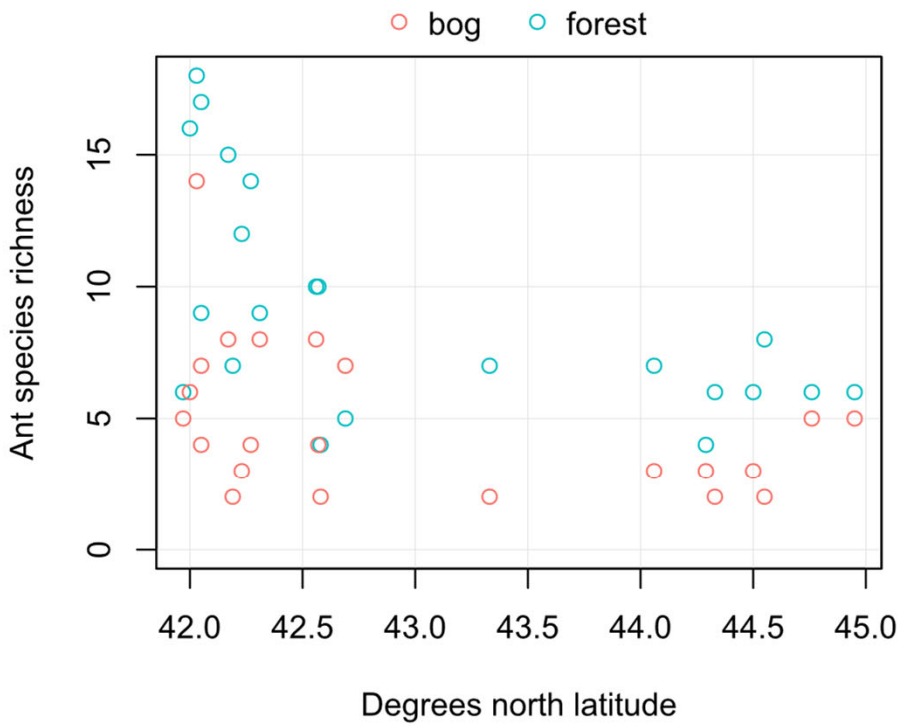
Linear predictor (always the same):

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

Inverse link function:

$\mu_i = \eta_i$	$\mu_i = e^{\eta_i}$	$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

Which makes the most sense?



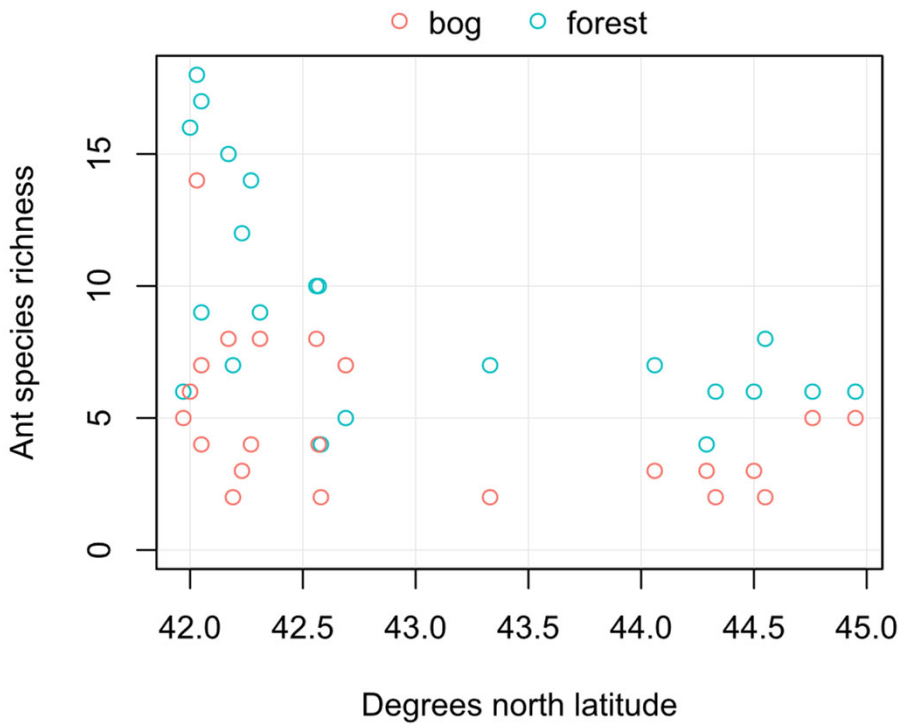
Most common models

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

Inverse link functions:

$\mu_i = \eta_i$	$\mu_i = e^{\eta_i}$	$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

Which makes the most sense?



Poisson
+
Log link

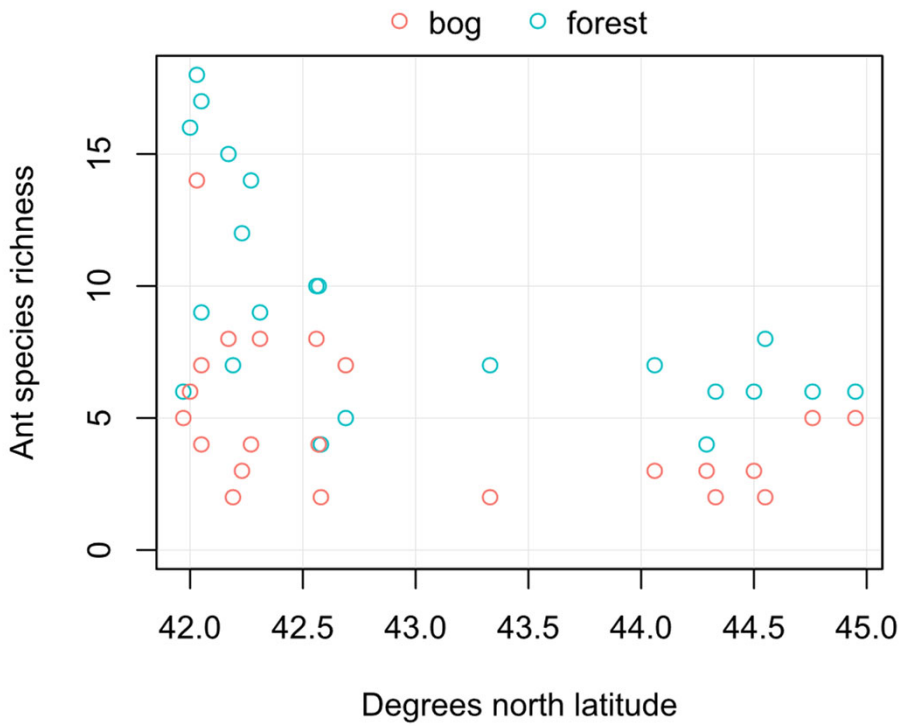
$$y_i \sim \text{Poisson}(\mu_i)$$

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

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

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

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 formula

```
fit <- glm(richness ~ habitat + latitude + habitat:latitude,
           family=poisson(link="log"), data=ant)
```

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

Model formula

```
fit <- glm(richness ~ habitat + latitude + habitat:latitude,
           family=poisson(link="log"), data=ant)
```

Equivalent:

```
richness ~ habitat * latitude

richness ~ 1 + habitat + latitude + habitat:latitude
```

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

Model matrix (design matrix)

```
fit <- glm(richness ~ habitat + latitude + habitat:latitude,
           family=poisson(link="log"), data=ant)
```

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

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

Model matrix (design matrix)

```
fit <- glm(richness ~ habitat + latitude + habitat:latitude,
           family=poisson(link="log"), data=ant)
```

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(fit)`

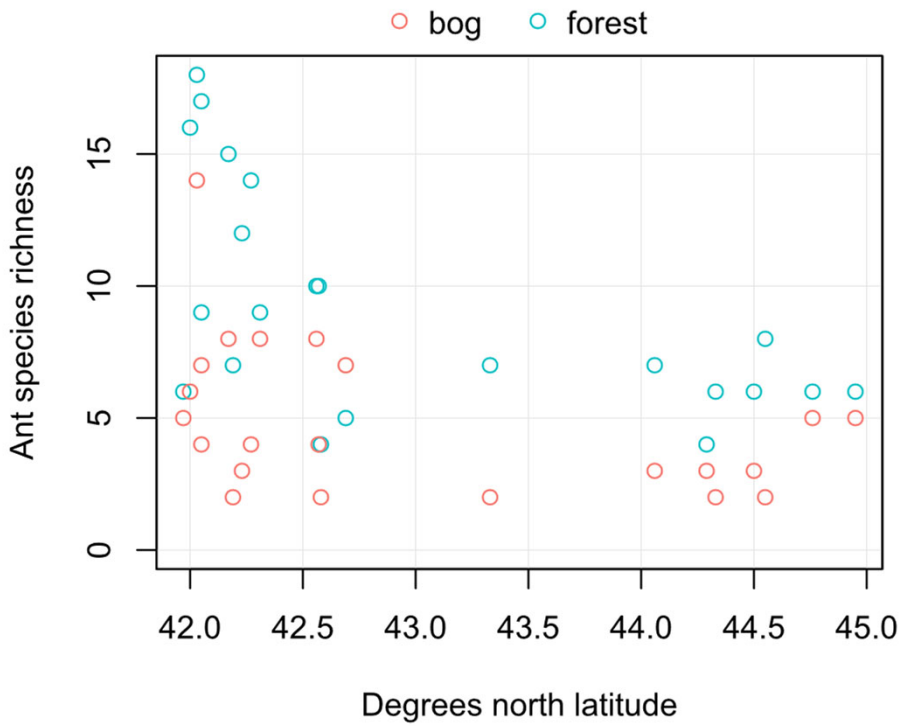
Model matrix

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

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

```
eta ~   1        + forest  + latitude + forest:latitude
```

Bayesian GLM: Priors?



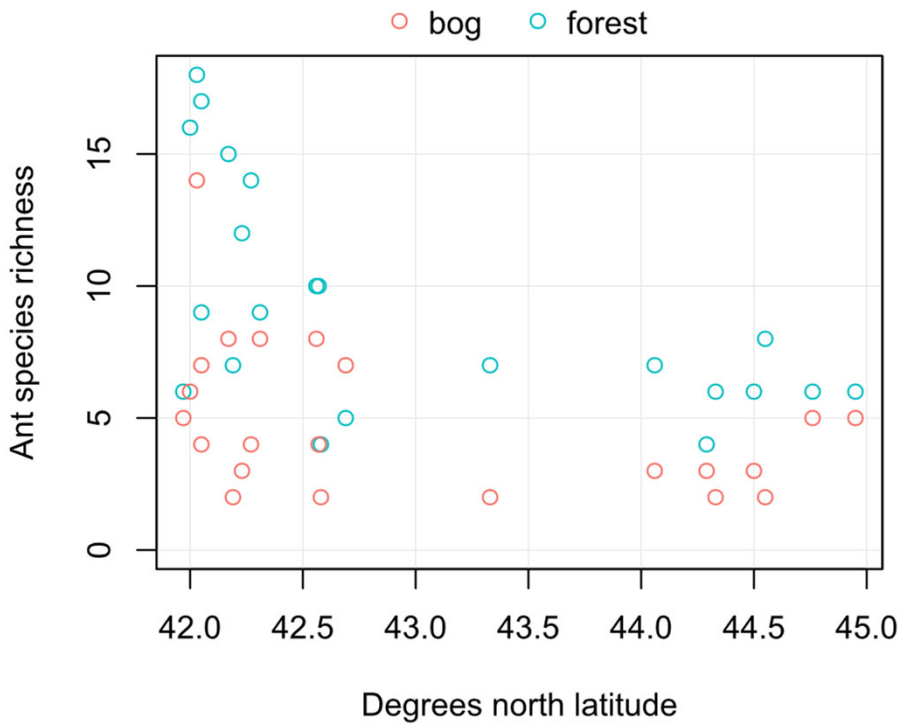
Poisson
+
Log link

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

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

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

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()