Activity 14: Statistical reasoning 6: generalized linear and multilevel models

Abbie & Gabe

1. Generalized linear models

1.1 Introduction

```
library(tidyverse) # For data wrangling
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.2.0     v readr     2.2.0
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.2     v tibble    3.3.1
v lubridate 1.9.5     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(brms) # For stats
```

```
Loading required package: Rcpp
Loading 'brms' package (version 2.23.0). Useful instructions
can be found by typing help('brms'). A more detailed introduction
to the package is available through vignette('brms_overview').

Attaching package: 'brms'

The following object is masked from 'package:stats':
```

```
    ar
```

```
library(ggeffects) # for plotting model predictions
# Note: I needed to also install the `insight` and `see` packages to get `modelbased` to inst
# install.packages('modelbased') # if you need to install this package
library(modelbased) # for plotting model predictions. supports the link scale (ggeffects does
```

```
Attaching package: 'modelbased'

The following objects are masked from 'package:ggeffects':

    collapse_by_group, pool_predictions, residualize_over_grid
```

```
# install.packages('faraway') # if you need to install this package
library(faraway) # For data on galapagos species richess
```

```
Attaching package: 'faraway'

The following object is masked from 'package:brms':

    epilepsy
```

Let's look at the (huge) variety of distribution families that are available to use:

```
?brmsfamily
```

Conceptual practice

For Q1.1a and b, consider the following response variables:

1. Counts of Clarkia flowers in a meadow
2. Whether or not a female elephant seal gives birth
3. The percent cover of red algae in the intertidal
4. Growth of a tree from one year to the next
5. The spatial area of a forest in square meters

Q1.1a What values can each of the response variables take on?

1. Counts of Clarkia flowers in a meadow

   Positive real integers

2. Whether or not a female elephant seal gives birth

   Binary– 0/1

3. The percent cover of red algae in the intertidal

   Non negative– fractions

4. Growth of a tree from one year to the next

   Positive, not bounded by an integer can be fraction

5. The spatial area of a forest in square meters

   Positive, not bounded by an integer can be fraction

Q1.1b Choose a distribution that fits each of the response variables

1. Counts of Clarkia flowers in a meadow

   Poisson

2. Whether or not a female elephant seal gives birth

   Binomial

3. The percent cover of red algae in the intertidal

   Beta

4. Growth of a tree from one year to the next

   Poisson

5. The spatial area of a forest in square meters

   Poisson

Q1.2 Choose a distribution that fits your final project response variable

1. Enzymatic activity is the response variable in Gabe's final project.
2. The values cannot be lower than 0 and do not have to be whole numbers
3. Therefore fits a poisson distribution.

1.2 GLM with a log link

Explore the data

```
# Read in the pre-stored data
data("gala")
# Check out the first 6 rows
head(gala)
```

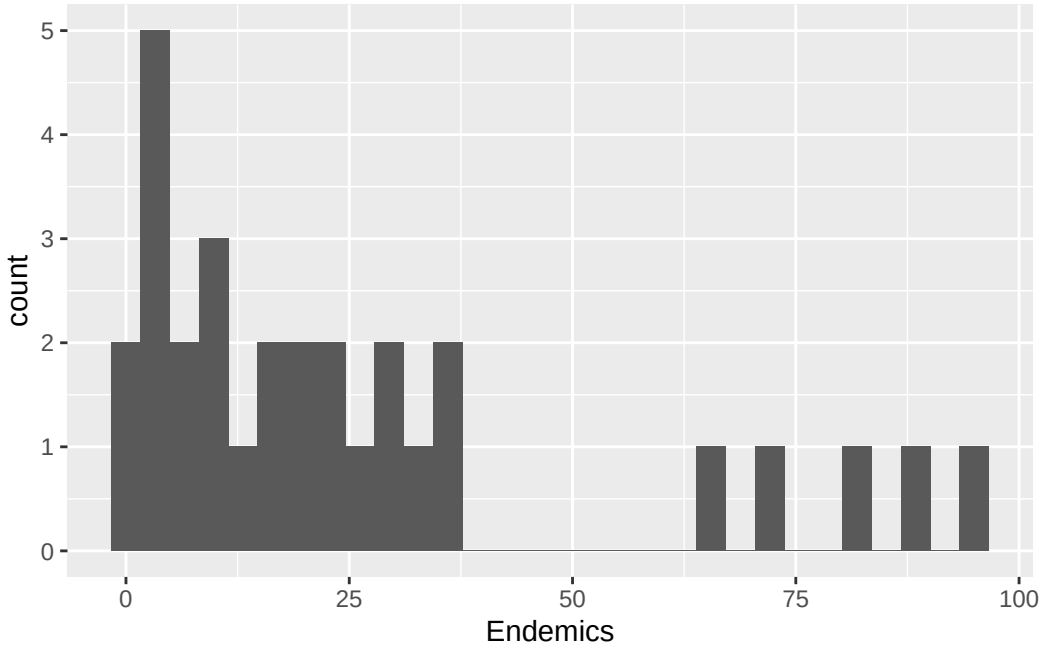
```
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        9  0.10        46     1.9  47.4     0.18
Coamano            2        1  0.05        77     1.9   1.9   903.82
Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
```

```
?gala
```

Q1.3 Plot a histogram of the response variable `Endemics`

```
ggplot(gala, aes(x = Endemics)) +
  geom_histogram()
```

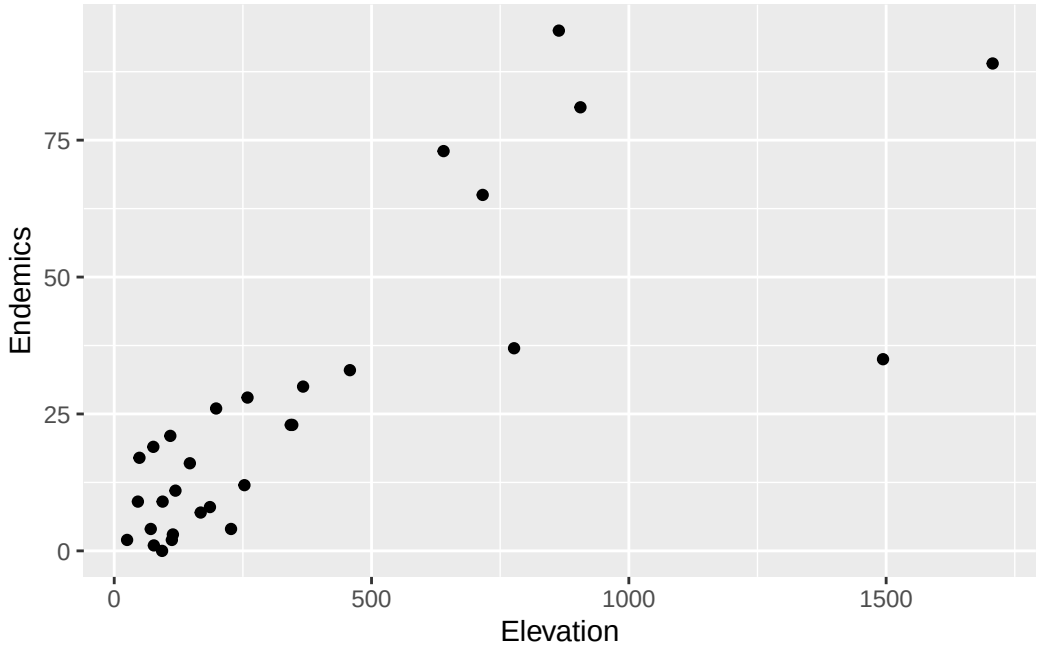
```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



This does not look like a normal distribution. This looks quite skewed toward lower numbers of endemic species. There are no numbers below zero and numbers are all whole.

Q1.4 Plot Endemics ~ Elevation

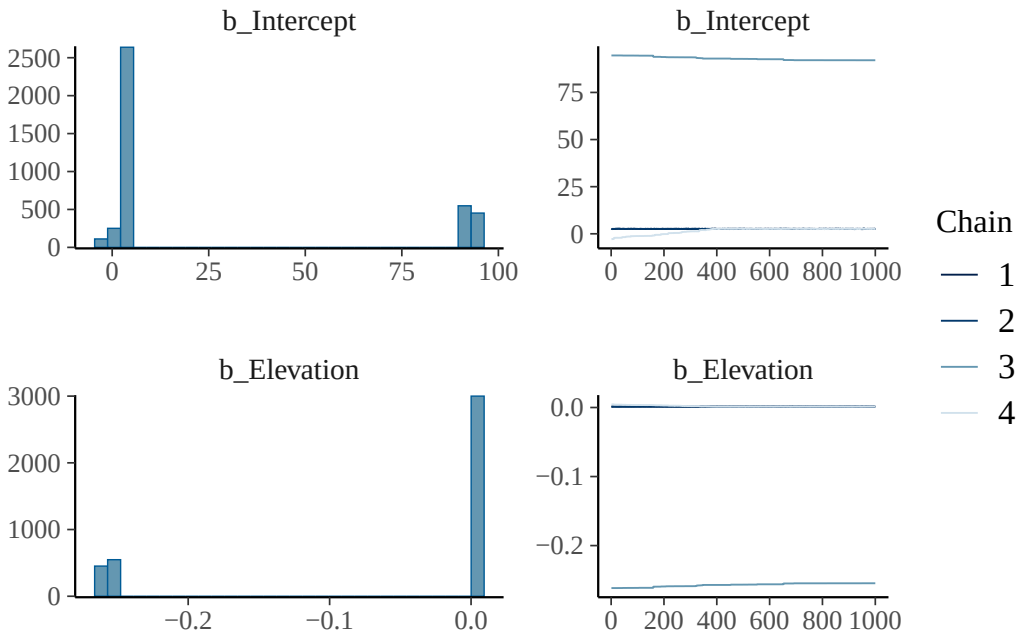
```
ggplot(gala, aes(y=Endemics, x= Elevation)) +
  geom_point()
```



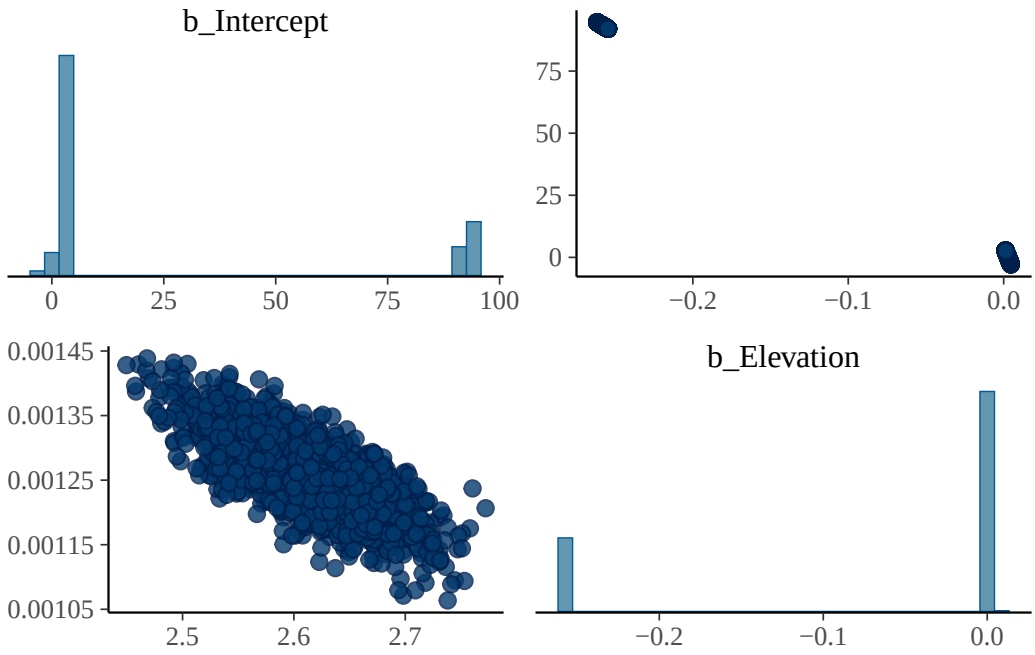
Run the model

```
# Endemics ~ Elevation
m.elev <-
  brm(data = gala, # Give the model the penguins data
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),
      # Specify the model here.
      Endemics ~ 1 + Elevation,
      # Here's where you specify parameters for executing the Markov chains
      # We're using similar to the defaults, except we set cores to 4 so the analysis runs fa
      iter = 2000, warmup = 1000, chains = 4, cores = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.elev")
```

```
plot(m.elev)
```


b_Intercept
b_Intercept
b_Elevation
b_Elevation
Chain
1
2
3
4

```
pairs(m.elev)
```


b_Intercept
b_Elevation

```r
summary(m.elev)
```

```
Warning: Parts of the model have not converged (some Rhats are > 1.05). Be
careful when analysing the results! We recommend running more iterations and/or
setting stronger priors.

 Family: poisson
  Links: mu = log
Formula: Endemics ~ 1 + Elevation
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    24.94     39.31    -1.19    94.55 1.92        6       11
Elevation    -0.06      0.11    -0.26     0.00 1.94        6       11

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Q1.5 Evaluate the output

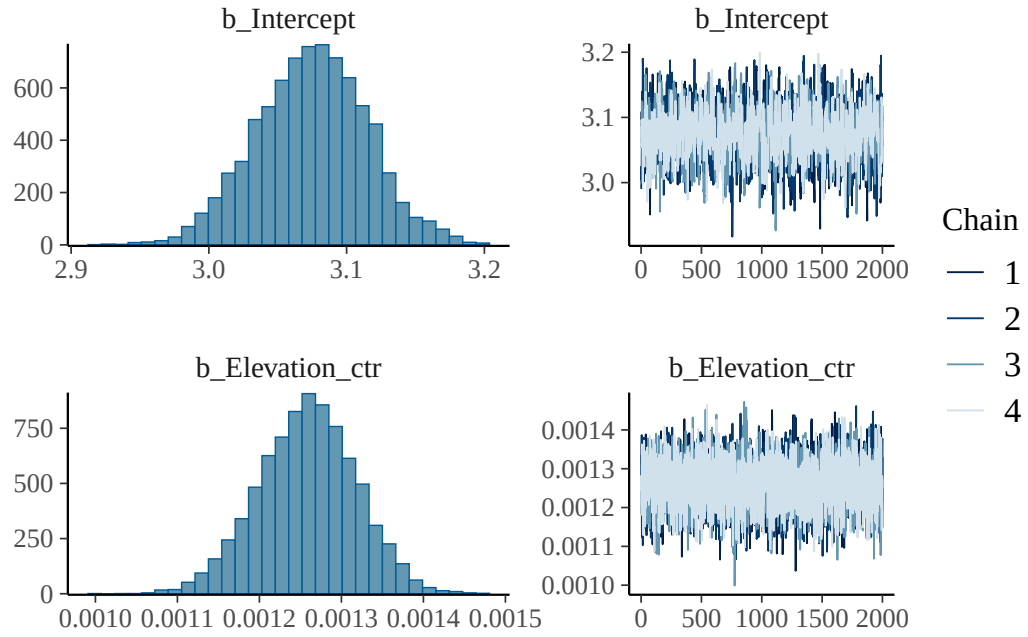
Looks like the model did not converge, the Rhat value is greater than 1.0. Chains are non overlapping and posterior distributions are not normally distributed.

Q1.6 Center the predictors

```r
#Adding in new column
gala <- gala %>%
  mutate(Elevation_ctr = Elevation - mean(Elevation))
```

```r
#New Model
m.elev2 <-
  brm(data = gala,
      family = poisson(link = "log"),
      Endemics ~ 1 + Elevation_ctr,
      iter = 6000, warmup = 4000, chains = 4, cores = 4,
      file = "output/m.elev2")
```

```
plot(m.elev2)
```



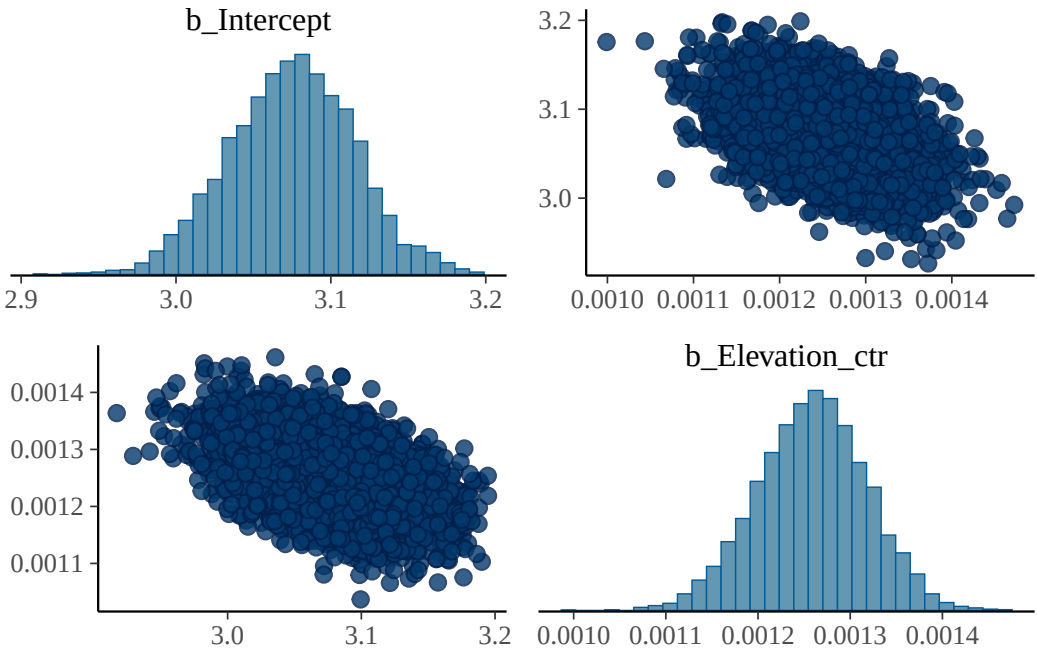

```
summary(m.elev2)
```

```
 Family: poisson
  Links: mu = log
Formula: Endemics ~ 1 + Elevation_ctr
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 6000; warmup = 4000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        3.07      0.04     3.00     3.16 1.00     1444     1496
Elevation_ctr    0.00      0.00     0.00     0.00 1.00     3691     4450

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
print(m.elev2, digits = 4)
```

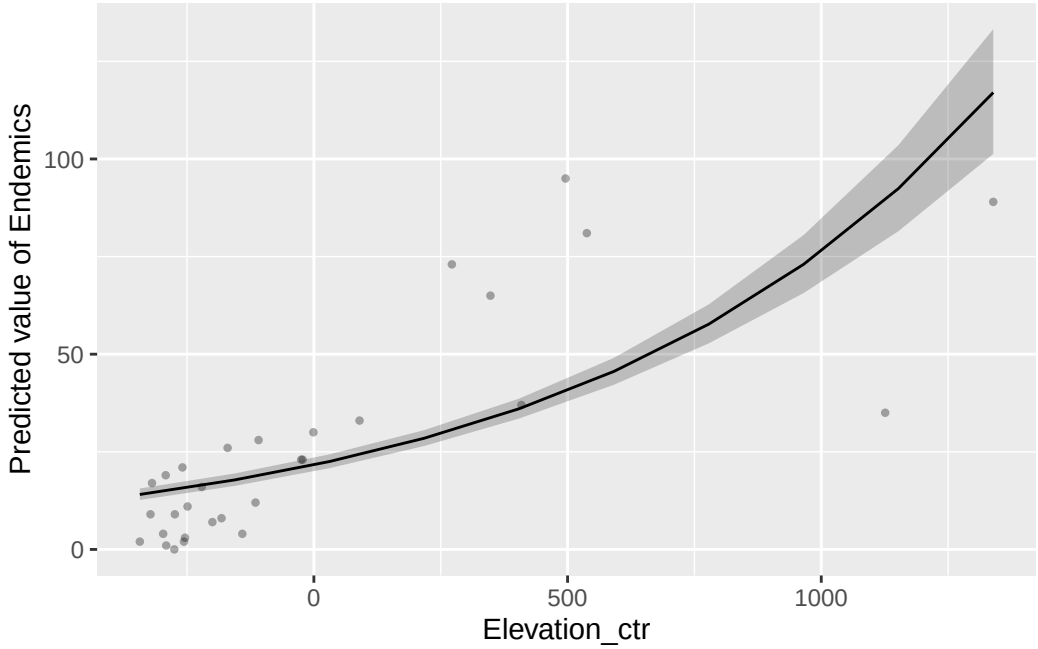
```
 Family: poisson
  Links: mu = log
Formula: Endemics ~ 1 + Elevation_ctr
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 6000; warmup = 4000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
              Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
Intercept       3.0747    0.0404   2.9951   3.1552 1.0036     1444     1496
Elevation_ctr   0.0013    0.0001   0.0011   0.0014 1.0012     3691     4450

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
pairs(m.elev2)
```

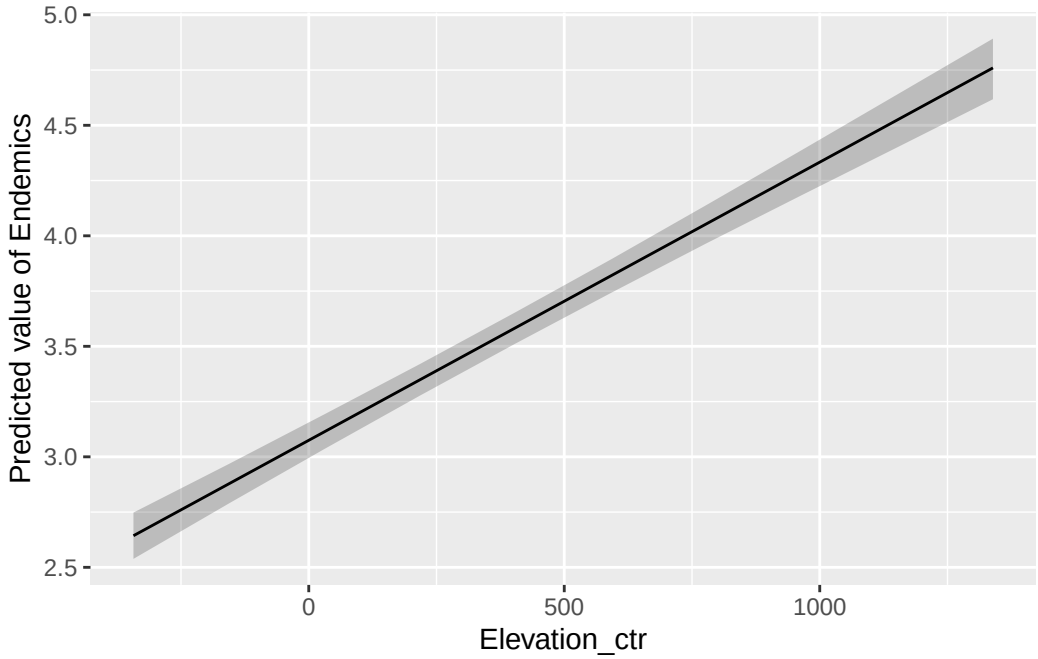


Plot the posterior

```
preds <- estimate_expectation(m.elev2, by = 'Elevation_ctr')
plot(preds, show_data = TRUE)
```




```
predslog <- estimate_expectation(m.elev2, by = 'Elevation_ctr', predict = 'link')
plot(predslog)
```



Interpreting link scale coefficients

```
print(m.elev2, digits = 4)
```

```
 Family: poisson
  Links: mu = log
Formula: Endemics ~ 1 + Elevation_ctr
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 6000; warmup = 4000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
             Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
Intercept       3.0747    0.0404   2.9951   3.1552 1.0036     1444     1496
Elevation_ctr   0.0013    0.0001   0.0011   0.0014 1.0012     3691     4450

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

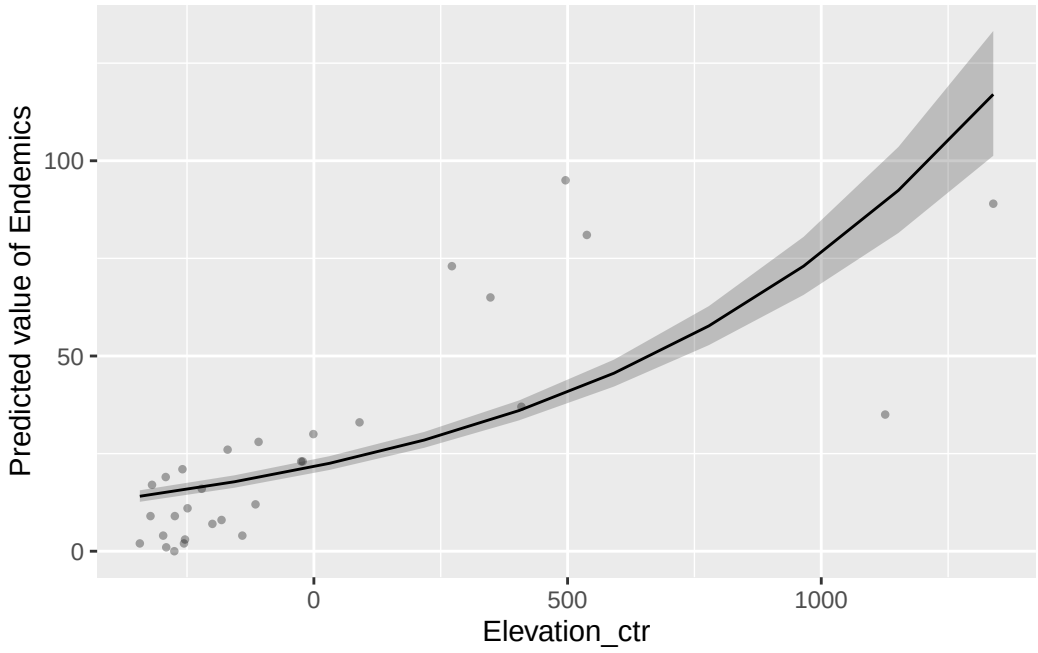
Backtransform:

```
exp(0.0013)
```

```
[1] 1.001301
```

Interpret

```
plot(preds, show_data = TRUE)
```



Q1.7 What is the percent change on the response scale?

1. Number of Clarkias blooming as a function of temperature in Celsius: 1.09

   ```
   exp(1.09)
   ```

   ```
   [1] 2.974274
   ```

   For every 1 degree increase in Celsius, there is a 2.97 times increase in number of Clarkias blooming as the previous degree.

2. Density of sea urchins per square meter in a quadrat as a function of number of sea otters: -2.5

```
exp(-2.5)
```

```
[1] 0.082085
```

For every sea otter present, there is a 8.2085% decrease in sea urchins densities per square meter.

3. Number of tomatoes per plant as a function of kg of fertilizer: 6.24

```
exp(6.24)
```

```
[1] 512.8585
```

For every kg of fertilizer, there is a 512.8 times increase in number of tomatoes per plant. (lol what?!)

DIY: Run a model of non-endemic species ~ distance from Santa Cruz Island

Q1.8 Create a non-endemic column

New column

```
gala <- gala %>%
  mutate(non_Endemics = Species - Endemics)
```

Plot of non endemics

```
ggplot(gala, aes(x= Scruz, y = non_Endemics)) +
  geom_point()
```

Q1.9 Run a model of non Endemics ~ distance from Santa Cruz Island

```
m.non_endem <-
  brm(data = gala,
      family = poisson(link = "log"),
      non_Endemics ~ 1 + Scruz,
      iter = 6000, warmup = 4000, chains = 4, cores = 4,
      file = "output/m.non_endem")
```

Q1.10 Evaluate the output

```
summary(m.non_endem)
```

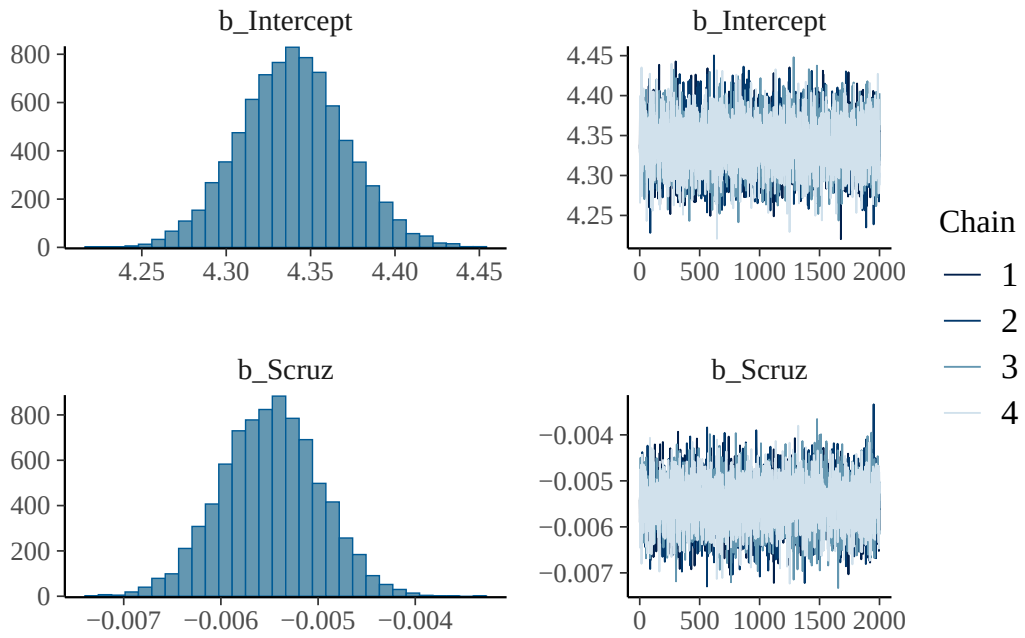
```
 Family: poisson
  Links: mu = log
Formula: non_Endemics ~ 1 + Scruz
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 6000; warmup = 4000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
```

```
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     4.34      0.03     4.28     4.40 1.00     5919     5755
Scruz        -0.01      0.00    -0.01    -0.00 1.00     6629     5609

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
print(m.non_endem, digits = 4)
```

```
 Family: poisson
  Links: mu = log
Formula: non_Endemics ~ 1 + Scruz
   Data: gala (Number of observations: 30)
  Draws: 4 chains, each with iter = 6000; warmup = 4000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
Intercept   4.3395    0.0316   4.2774   4.4022 1.0002     5919     5755
Scruz      -0.0055    0.0005  -0.0065  -0.0045 1.0004     6629     5609

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
plot(m.non_endem)
```




```
exp(-0.01)
```

```
[1] 0.9900498
```

Looks like the model ran correctly, Rhat is 1, the chains are overlapping, and the posterior distributions are normally distributed.

Q1.11 Interpret the output

1. What is the effect of distance from Santa Cruz Island on number of non-endemic species? Report the a) original output on the log scale, b) your backtransformed value, and c) the percent change that this translates to. Describe the effect using the proper units.

a) -0.01

b)

```
0.9900498
```

c) For every 1 km distance from SC island, the amount of non-endemic species is changing at a rate of 0.99.

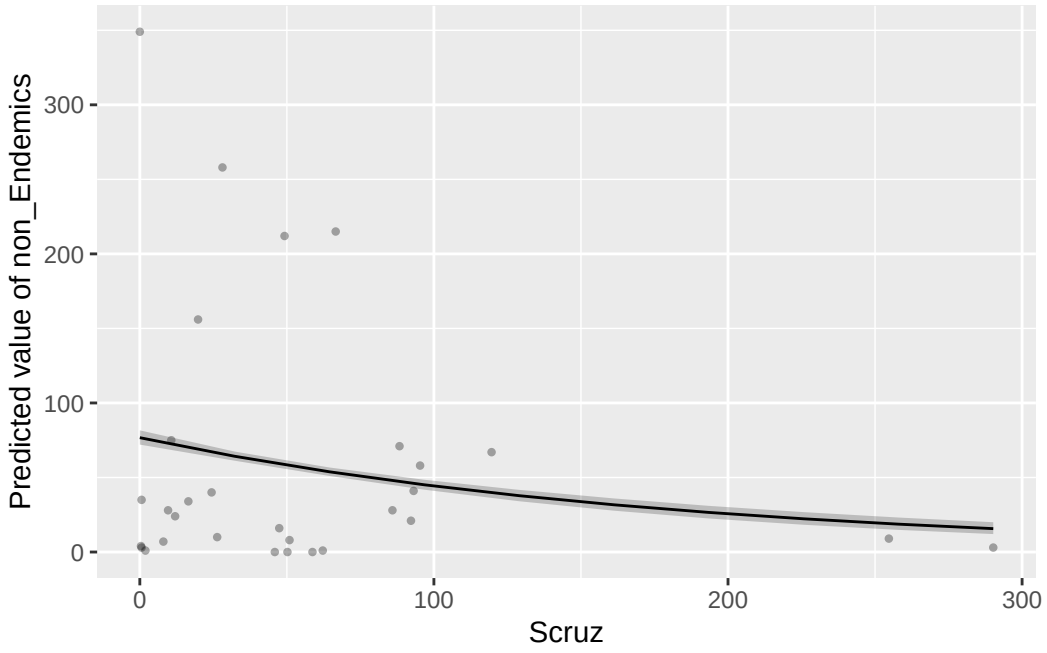
2. Does it seem like the slope estimate is different from zero? Why?

   Yes, it does seem like the slope estimate is different from zero, but only barely because the compatibility intervals are only just outside of the range of zero.

Q1.12 Plot the posterior

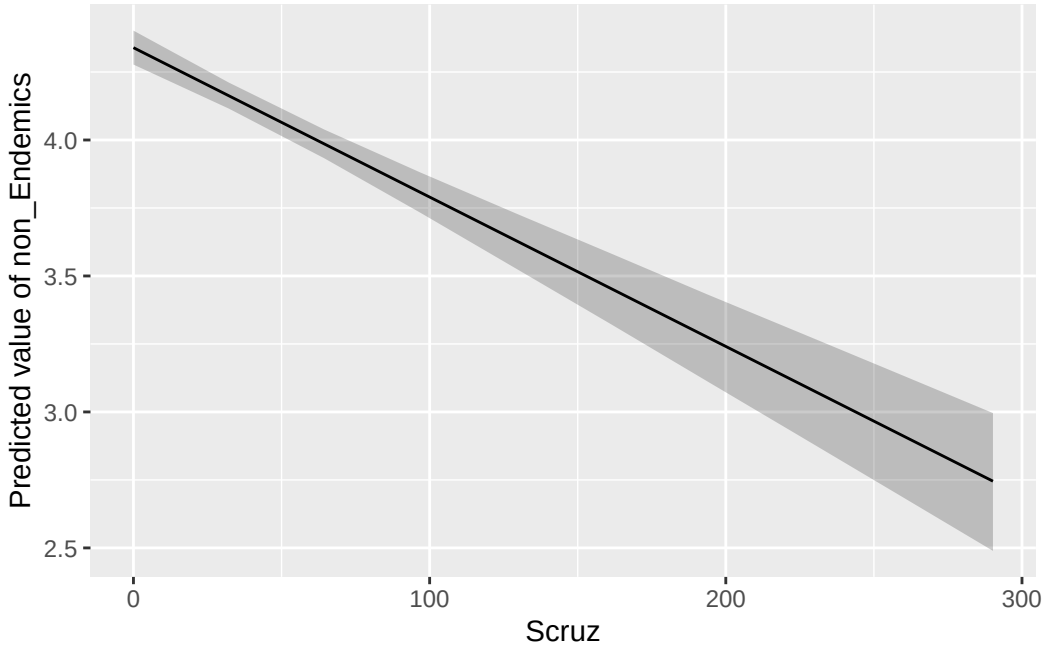
Response scale

```
preds_endem <- estimate_expectation(m.non_endem, by = 'Scruz')
plot(preds_endem, show_data = TRUE)
```



Log Scale

```
preds_endemlog <- estimate_expectation(m.non_endem, by = 'Scruz', predict = 'link')
plot(preds_endemlog
    )
```

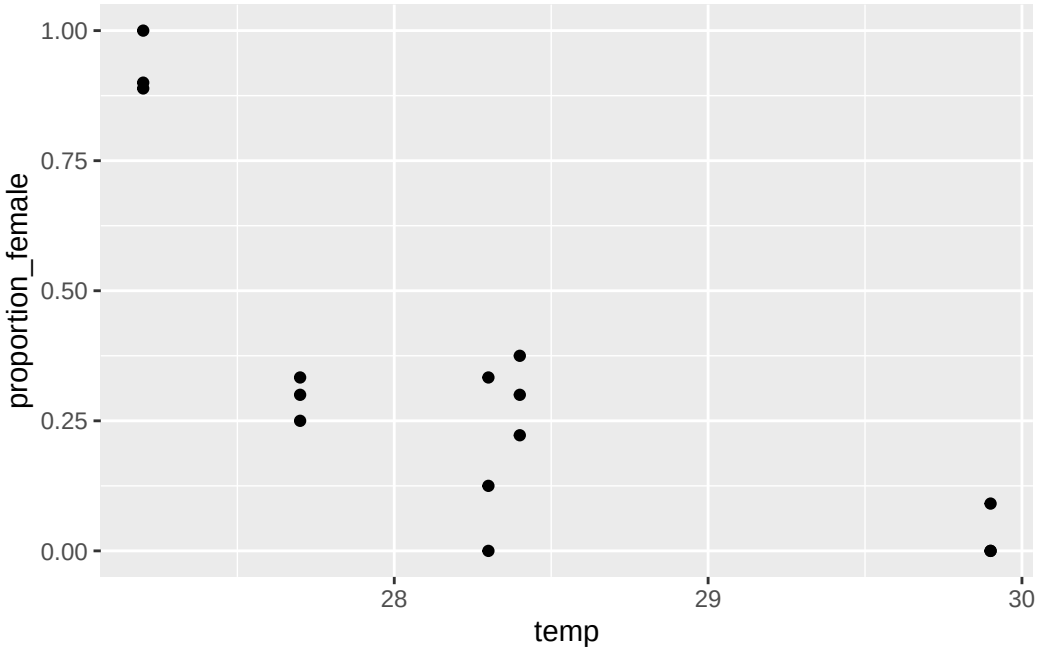


1.3 GLM with a logit link

Bring in turtle data

```
turtle <- faraway::turtle %>%
  mutate(total_turtles = male + female,
         proportion_female = female/total_turtles)
```

```
turtle %>%
  ggplot(aes(x = temp, y = proportion_female)) +
  geom_point()
```



Run model

```
m.turt <-
  brm(data = turtle, # Give the model the data
      # Choose a binomial distribution - THIS IS THE NEW PART!
      family = binomial(link = "logit"),
      # Specify the model here.
      female | trials(total_turtles) ~ 1 + temp,
      # Here's where you specify parameters for executing the Markov chains
      # We're using similar to the defaults, except we set cores to 4 so the analysis runs fa
      iter = 4000, warmup = 1000, chains = 4, cores = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.turt")
```

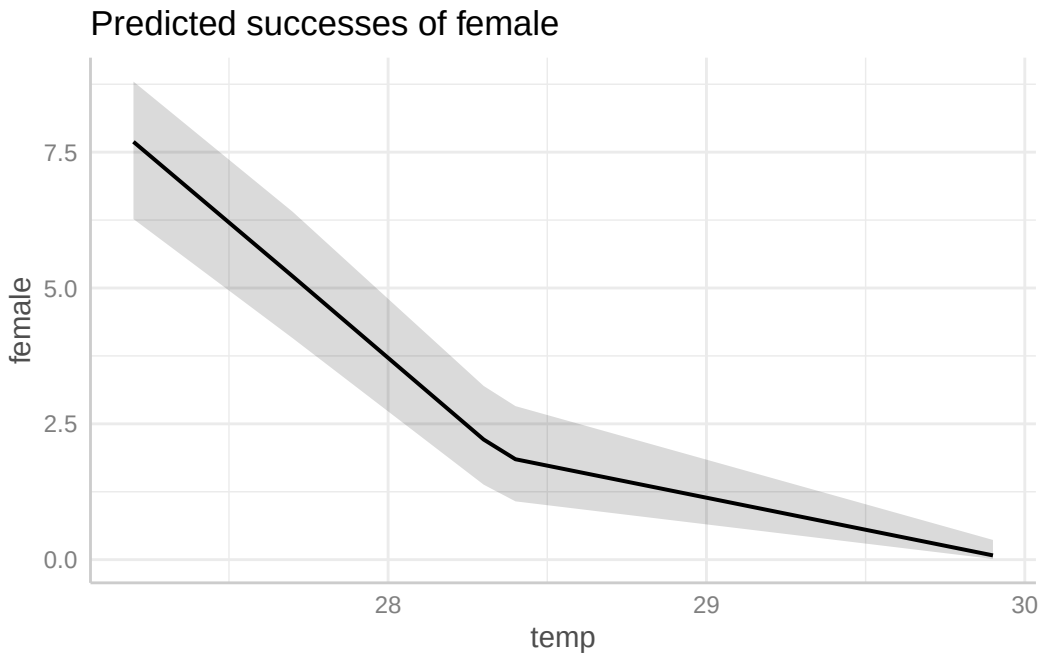
```
summary(m.turt)
```

```
 Family: binomial
  Links: mu = logit
Formula: female | trials(total_turtles) ~ 1 + temp
   Data: turtle (Number of observations: 15)
```

```
  Draws: 4 chains, each with iter = 4000; warmup = 1000; thin = 1;
         total post-warmup draws = 12000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    62.70     11.84    40.81    86.92 1.00     4641     5822
temp         -2.26      0.42    -3.13    -1.48 1.00     4589     5825

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Interpret model and plot predictions

```r
pred <- predict_response(m.turt, condition = c(total_turtles = 10))

plot(pred)
```



2. Multilevel models

Conceptual practice

Q2.1 Fixed effects vs random effects

For the following variables in the model examples below, denote which variables are the fixed effects and which could be accounted for as random effects (some variables could be either, but consider then as being eligible to be random effects):

1. Student high school graduation rates as a function of: parental income, state of residence, and school district

   Fixed effects: parental income, state of residence, school district

   Random Effects:

2. Density of kelp as a function of: latitude, site, transect number, and density of sea urchins

   Fixed: density, latitude

   Random: site, transect number,

3. Probability of whale giving birth as a function of: age, annual temperature, year, individual ID

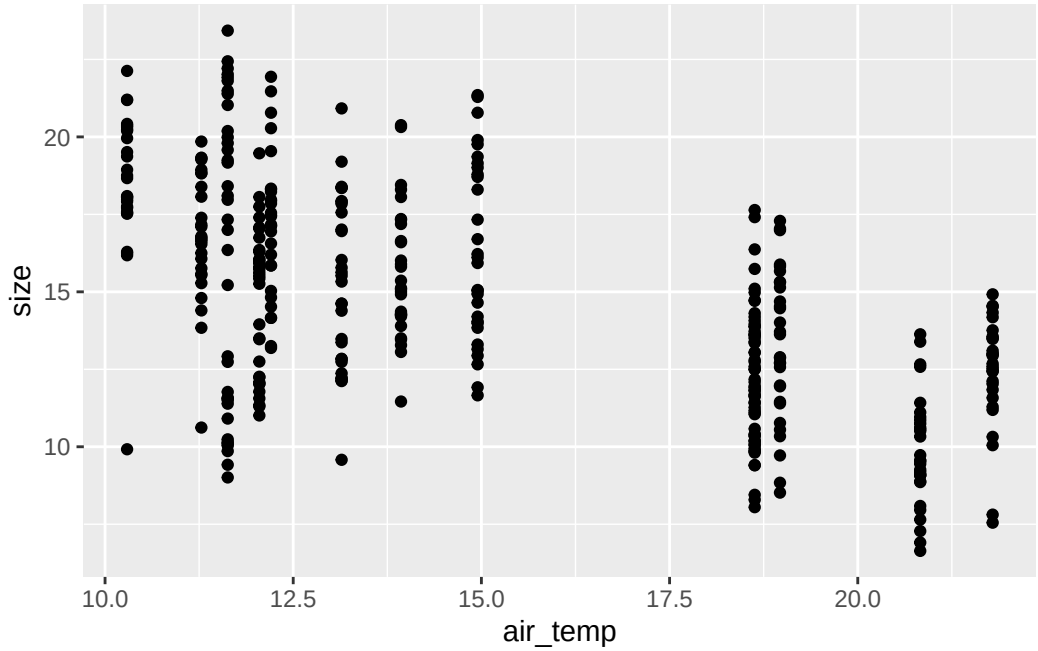
   Fixed: Annual temperature, Age

   Random: Year, Individual ID

Incorporate random effects

```
pie_crab <- lterdatasampler::pie_crab %>%
  mutate(site = as.factor(site))
```

```
pie_crab %>%
  ggplot(aes(x = air_temp, y = size)) +
  geom_point()
```




```
m.watertemp <-
  brm(data = pie_crab, # Give the model the penguins data
      # Use a gamma distribution
      family = Gamma(link = "log"),
      # Specify the model here.
      size ~ 1 + water_temp,
      # Here's where you specify parameters for executing the Markov chains
      # We're using similar to the defaults, except we set cores to 4 so the analysis runs fa
      iter = 2000, warmup = 1000, chains = 4, cores = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.watertemp")

print(m.watertemp, digits = 3)
```

```
 Family: gamma
  Links: mu = log
Formula: size ~ 1 + water_temp
   Data: pie_crab (Number of observations: 392)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
```

```
           Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Intercept     3.355     0.055    3.247    3.463 1.000     4514     3040
water_temp   -0.038     0.003   -0.044   -0.032 1.000     4521     3110

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   23.249     1.659   20.129   26.690 1.000     2569     2683

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
m.watertemp.site <-
  brm(data = pie_crab, # Give the model the penguins data
      # Use a gamma distribution
      family = Gamma(link = "log"),
      # Specify the model here.
      size ~ 1 + water_temp + (1|site),
      # Here's where you specify parameters for executing the Markov chains
      # We're using similar to the defaults, except we set cores to 4 so the analysis runs fa
      iter = 2000, warmup = 1000, chains = 4, cores = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.watertemp.site")

print(m.watertemp.site, digits = 3)
```

```
 Family: gamma
  Links: mu = log
Formula: size ~ 1 + water_temp + (1 | site)
   Data: pie_crab (Number of observations: 392)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~site (Number of levels: 13)
              Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
sd(Intercept)    0.124     0.034    0.075    0.209 1.003      932     1132

Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Intercept     3.366     0.192    2.974    3.749 1.002     1041     1455
water_temp   -0.039     0.011   -0.060   -0.017 1.002     1106     1341
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   30.090     2.184   26.077   34.599 1.000     3107     2923

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Q2.2 What is the effect of water_temp on crab size on the response scale?

1. What is the effect of water_temp on crab size? Report the a) original output on the log scale, b) your backtransformed value, and c) the percent change that this translates to. Describe the effect using the proper units.

   a)

   ```
   -0.039
   ```

   b)

   ```
   exp(-0.039 )
   ```

   ```
   [1] 0.9617507
   ```

   c) For every 1 degree increase in water temperature, there is a 0.96% mm increase in crab carapace size.

2. Does it seem like the slope estimate is different from zero? Why?

   Yes, this is reasonably different from zero since the 95% compatibility intervals are outside of the bounds of zero.

Q2.3 Compare WAIC and PSIS of the two models

```
#PSIS
loo(m.watertemp)
```

```
Computed from 4000 by 392 log-likelihood matrix.

         Estimate   SE
elpd_loo   -984.6 12.1
p_loo         2.7  0.2
```

```
looic      1969.3 24.3
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.0]).

All Pareto k estimates are good (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

```
loo(m.watertemp.site)
```

```
Computed from 4000 by 392 log-likelihood matrix.

         Estimate   SE
elpd_loo   -938.5 13.7
p_loo        12.2  0.8
looic      1877.0 27.4
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.6]).

All Pareto k estimates are good (k < 0.7).
See help('pareto-k-diagnostic') for details.
```

```
#WAIC
waic(m.watertemp)
```

```
Computed from 4000 by 392 log-likelihood matrix.

          Estimate   SE
elpd_waic   -984.6 12.1
p_waic         2.7  0.2
waic        1969.3 24.3
```

```
waic(m.watertemp.site)
```

```
Computed from 4000 by 392 log-likelihood matrix.
```

```
          Estimate   SE
elpd_waic   -938.5 13.7
p_waic        12.1  0.8
waic        1876.9 27.4
```

With side included, the WAIC and PSIS values are lower, so this model is better.

Predict response of random effects

```
preds <- predict_response(m.watertemp.site,
                          interval = "prediction",
                          terms = "site",
                          type = "random")

plot(preds)
```