Stats Reasoning 6

Paige and Peter

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

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
Warning: package 'tibble' was built under R version 4.4.3
```

```
Warning: package 'purrr' was built under R version 4.4.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.1
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
```

```
Warning: package 'brms' was built under R version 4.4.3
```

```
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
```

```
Warning: package 'ggeffects' was built under R version 4.4.3
```

```
# 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
```

```
Warning: package 'modelbased' was built under R version 4.4.3


Attaching package: 'modelbased'

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

    pool_predictions, residualize_over_grid
```

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

```
Warning: package 'faraway' was built under R version 4.4.3


Attaching package: 'faraway'

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

    epilepsy
```

Q1.1

Q1.1a

1. Numerical (0 to infinity, positive, whole numbers)
2. Y/N, 0/1
3. 0-1 or 0-100 , percent
4. Numerical
5. Numerical (positive)

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

1. Poisson
2. Bernoulli
3. Beta
4. Gamma
5. Gamma

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

Now, 1. Write the response variable that you are using for your final project.

Beta diversity

2. What values can your response variable take on? (e.g. real numbers, negative real numbers, positive real numbers, zeroes, integers, fractions, 0's/1's, etc.)

Continuous, real numbers

3. What distribution (or distributions) fits your response variable?

Gaussian / gamma

1.2

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

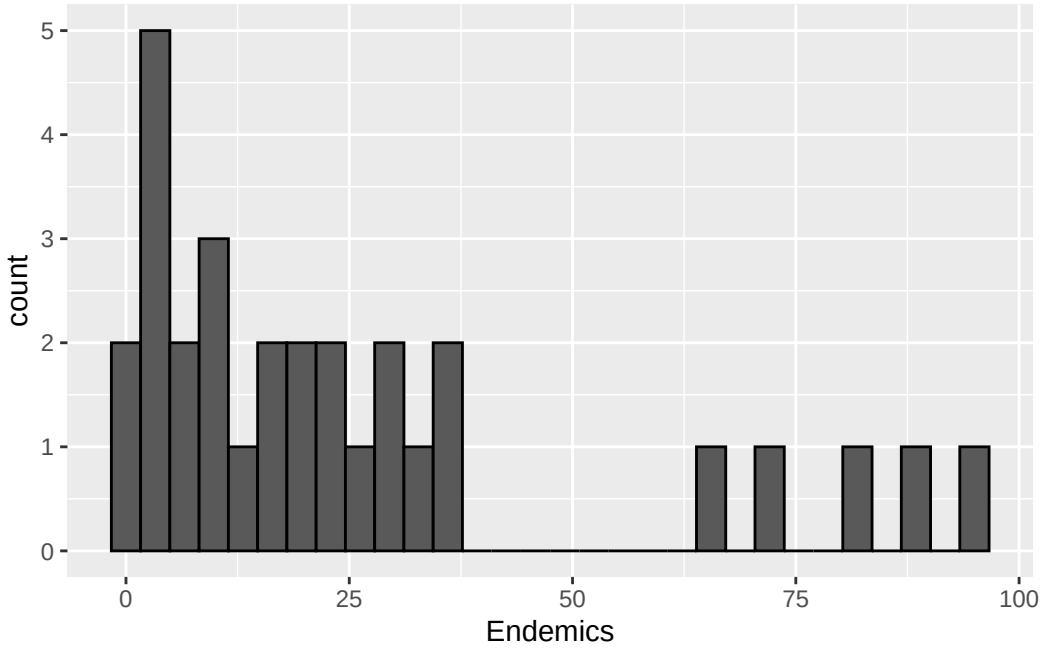
```
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        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
```

```
# Look at the help page too!
```

Q1.3

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

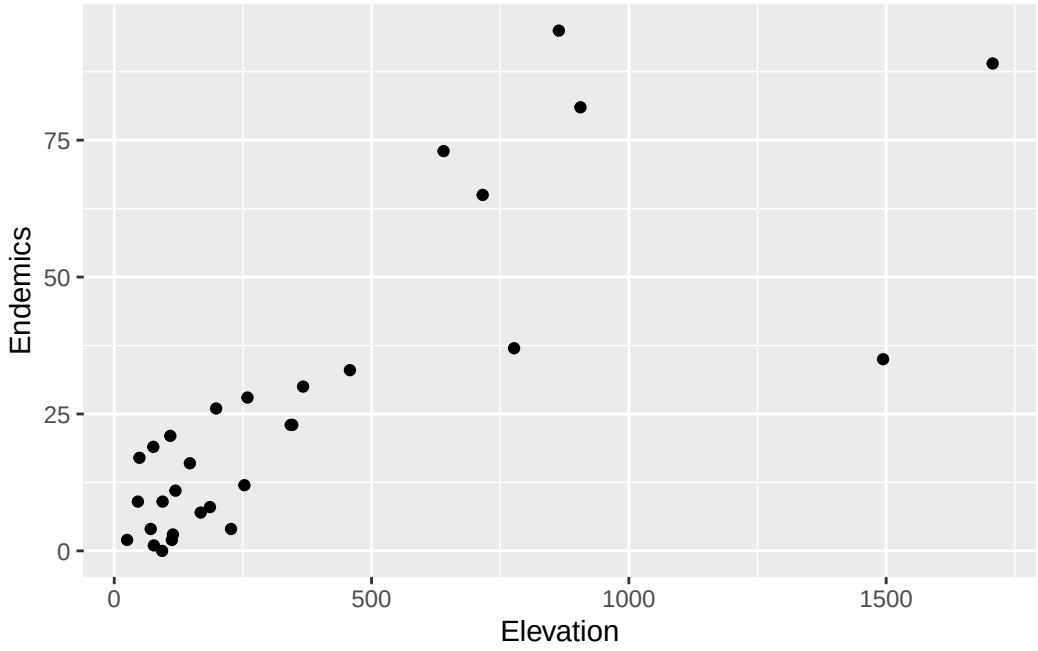
```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



Q1.4 Plot Endemics ~ Elevation

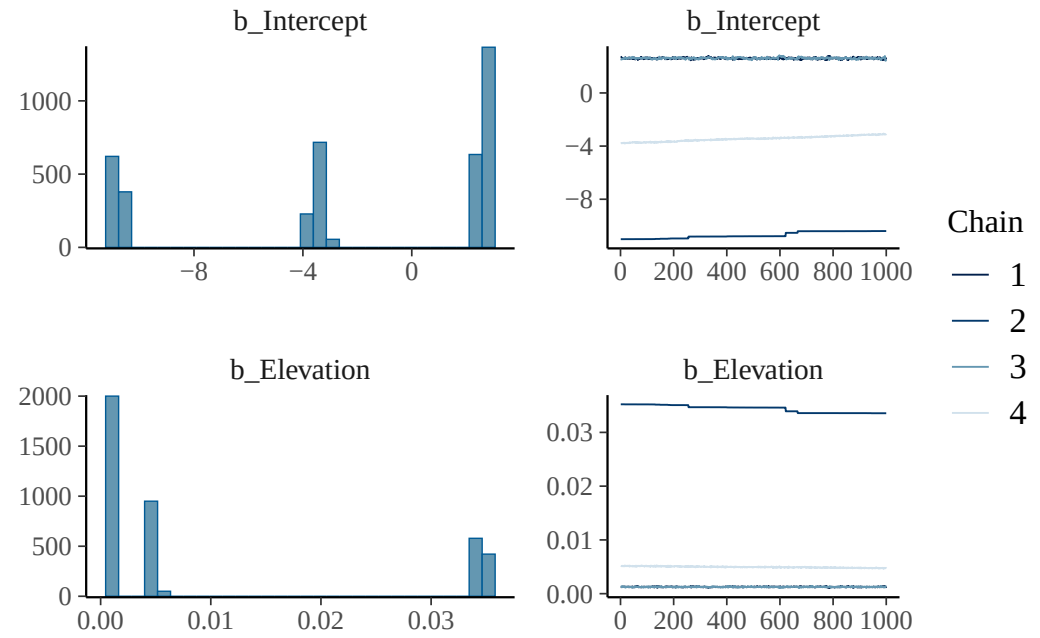
Make a plot to visualize the model we are about to run: Endemics as a function of Elevation.

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

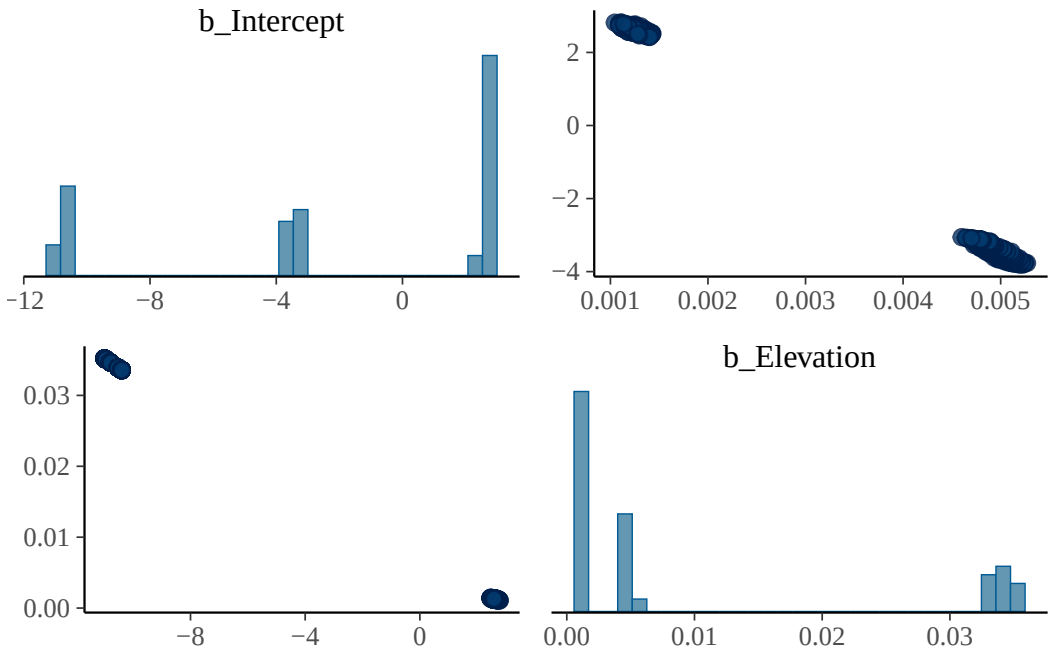



```
# Endemics ~ Elevation
m.elev <-
  brm(data = gala, # Give the model the penguins data
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),
      # Specify the model here.
      Endemics ~ 1 + Elevation,
      # Here's where you specify parameters for 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

```
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    -2.23      5.48   -11.00     2.70 2.31        5       11
Elevation     0.01      0.01     0.00     0.04 2.30        5       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).
```

Q 1.5

Answer: The Rhat values are ~2.3 and very different from 1. The posterior plots are not unimodal and the posterior chains are not overlapping.

Clustering at either extreme of the elevation. Not enough data at either low or high elevations to properly train the model and set the intercept.

Q1.6 Center the predictors

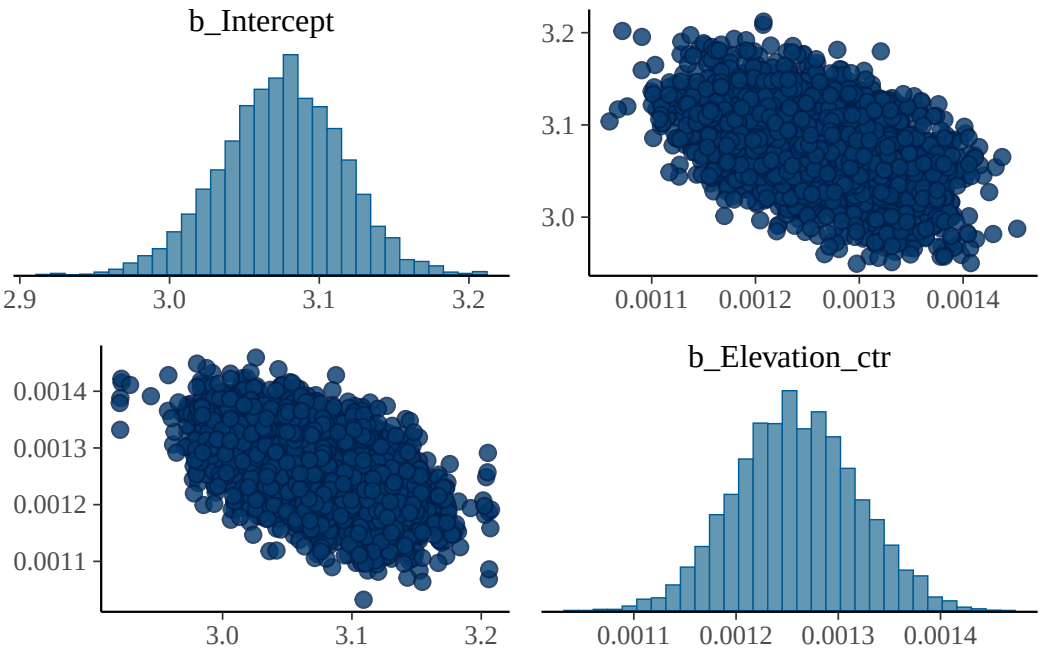
```
print(m.elev2, digits=4)
```

```
 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
```

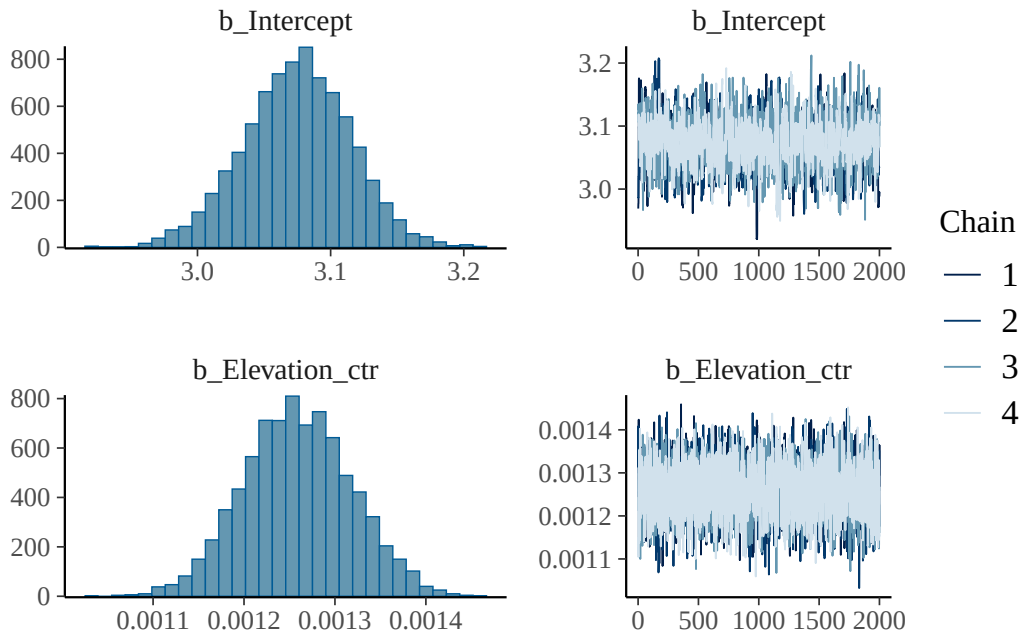
```
Regression Coefficients:
              Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
Intercept       3.0747    0.0402   2.9931   3.1514 1.0016     1397     1298
Elevation_ctr   0.0013    0.0001   0.0011   0.0014 1.0003     3749     4499

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).
```

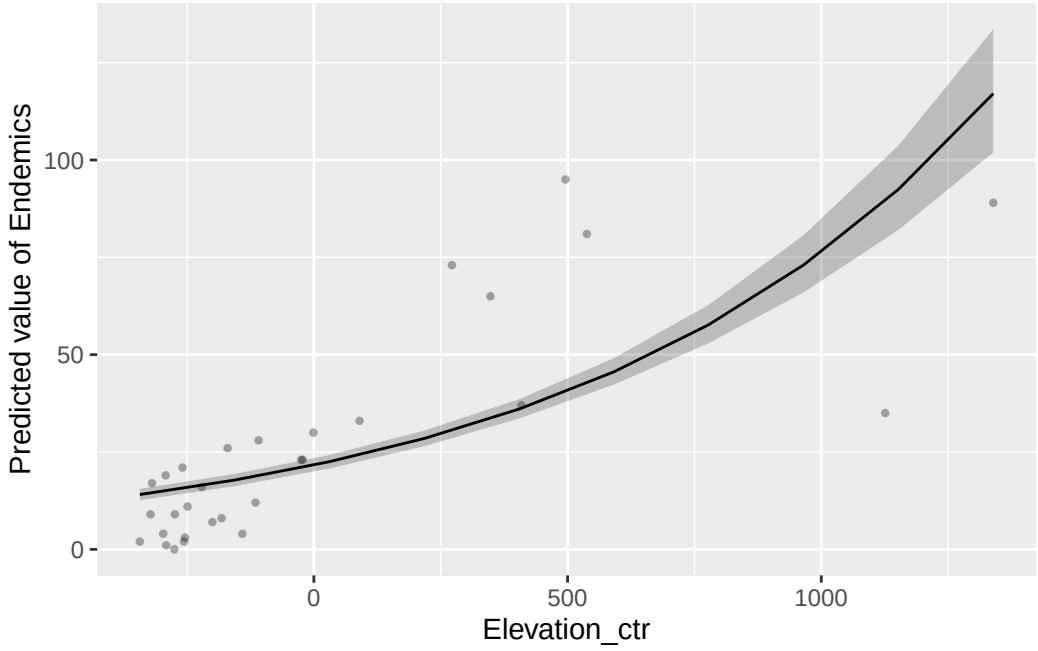
```
pairs(m.elev2)
```



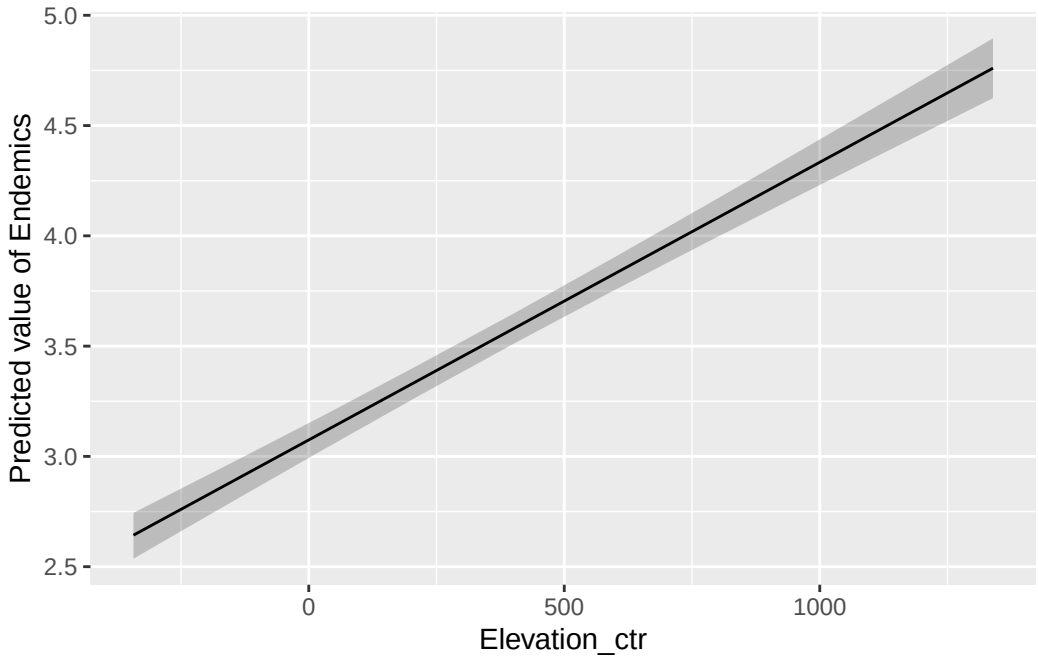

```
plot(m.elev2)
```




```
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)
```




```
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    0.0402   2.9931   3.1514 1.0016     1397     1298
Elevation_ctr   0.0013    0.0001   0.0011   0.0014 1.0003     3749     4499

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).
```

1. Back transform - My slope value is 0.0013, so I will take the `exp()` of that value:

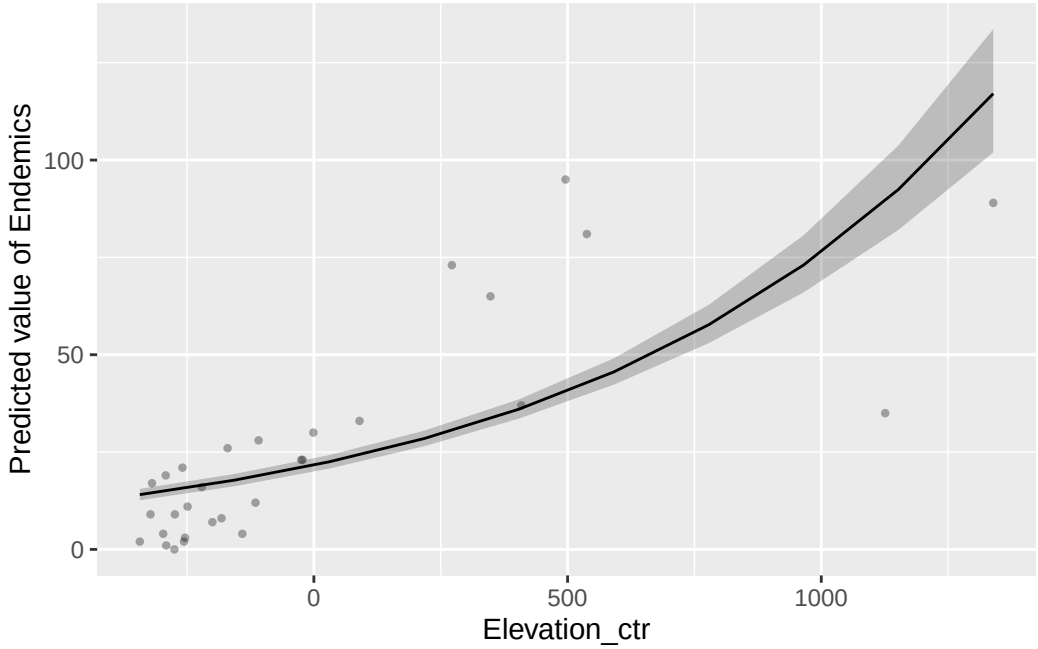
```
exp(0.0013)
```

```
[1] 1.001301
```

This back transforms to 1.0013

2. **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 increase in degree C, there is a 2.9% increase in clarkias blooming.

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 additional sea otter, urchin density decreases by 0.08% per square meter

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

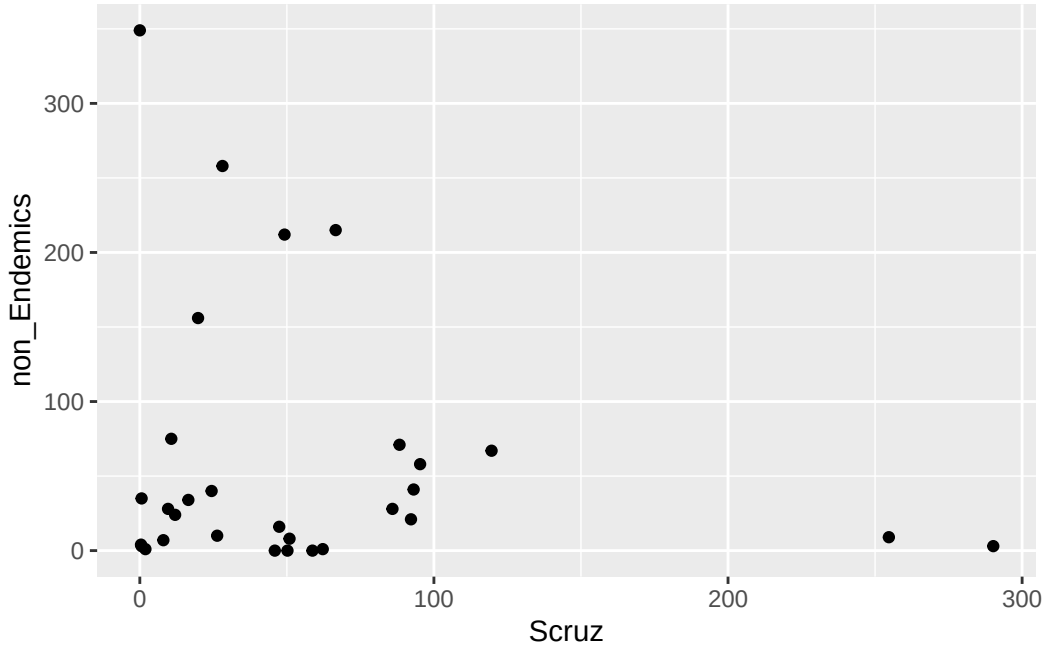
```
exp(6.24)
```

```
[1] 512.8585
```

Q1.8 Create a non-endemic column

```
gala2 <- gala %>%
  mutate(non_Endemics = Species - Endemics)

ggplot(gala2, aes(x= Scruz, y= non_Endemics))+
  geom_point()
```



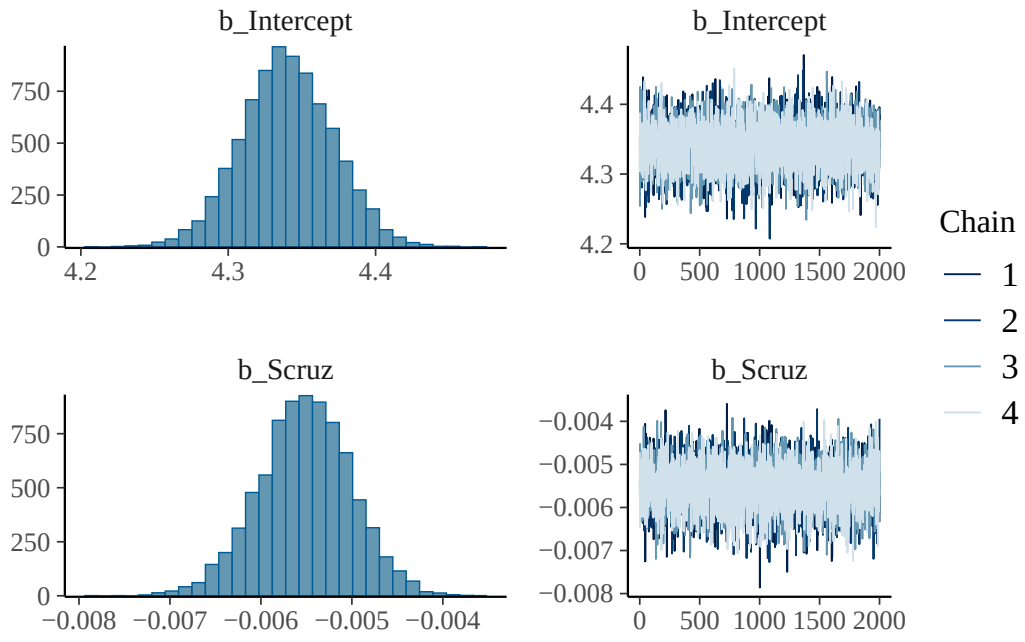
Q1.9 Run a model of non Endemics ~ distance from Santa Cruz Island

Run a model with a Poisson distribution and log-link function that models the number of non-endemic species as a function of distance from Santa Cruz Island.

```
m.nonedemic <-
  brm(data = gala2, # Give the model the penguins data
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),
      # Specify the model here.
      non_Endemics ~ 1 + Scruz,
      # 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 = 6000, warmup = 4000, chains = 4, cores = 4,
      prior = prior(normal(0, 0.1), class = b),
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.nonendemic")
```

Q1.10 Evaluate the output

```
plot(m.nonedemic)
```



```
summary(m.nonedemic)
```

```
 Family: poisson
  Links: mu = log
Formula: non_Endemics ~ 1 + Scruz
   Data: gala2 (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     5472     5174
Scruz        -0.01      0.00    -0.01    -0.00 1.00     5920     5002

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.nonedemic, digits = 4)
```

```
 Family: poisson
  Links: mu = log
Formula: non_Endemics ~ 1 + Scruz
   Data: gala2 (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   4.3400    0.0311   4.2787   4.4009 1.0013     5472     5174
Scruz      -0.0055    0.0005  -0.0066  -0.0045 1.0005     5920     5002

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).
```

rhat values are 1, posterior chains converge, and the posterior plots are unimodal/uniform

Q1.11 Interpret the output

Interpret the model coefficients by writing a 2-3 sentence results paragraph that answers:

a) The original value was -0.1
b)

```
exp(-0.01)
```

```
[1] 0.9900498
```

c) The percent change is then 0.99%

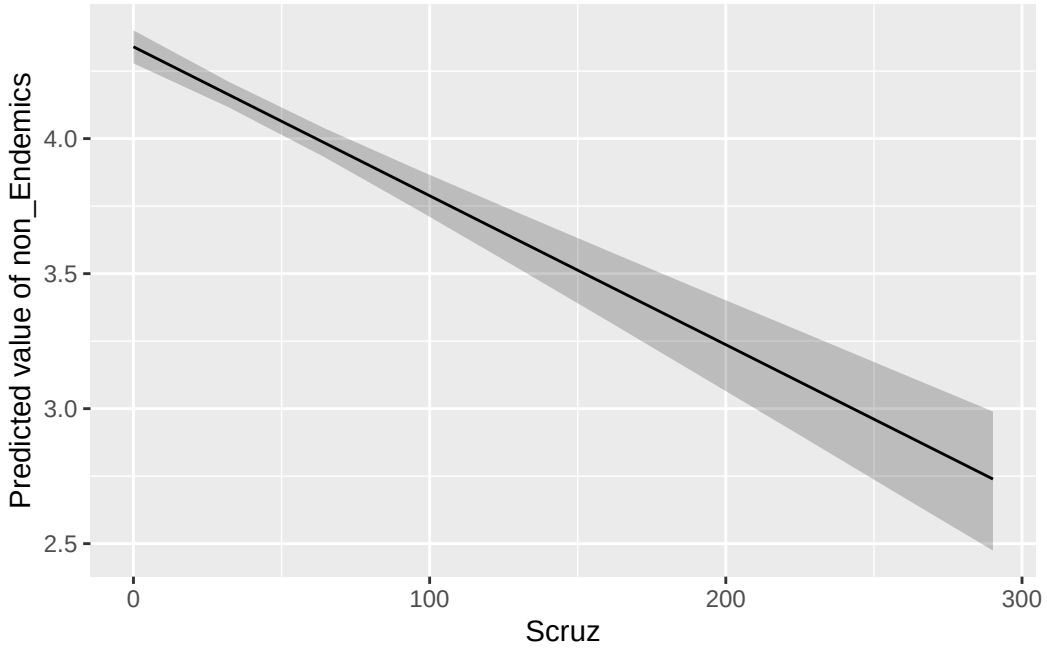
As distance from Santa Cruz Island increases, the frequency of nonendemics decreases. With every unit increase in distance from Santa Cruz island, there is a 1% decrease in non-endemic species.

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

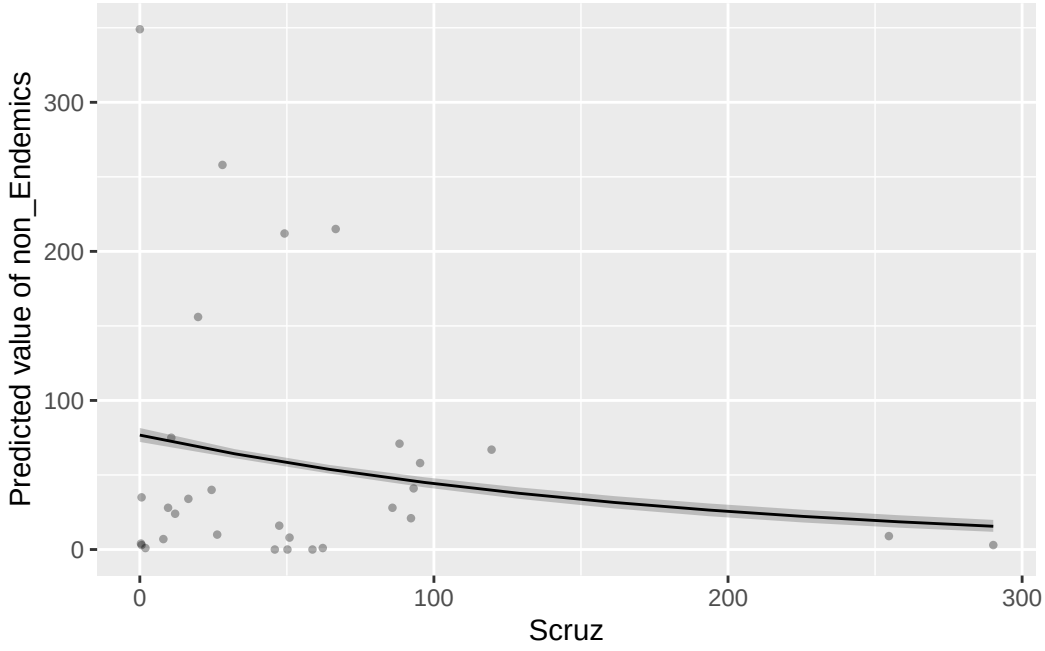
   Yes, the 95% CI's exclude 0

Q1.12 Plot the posterior

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



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

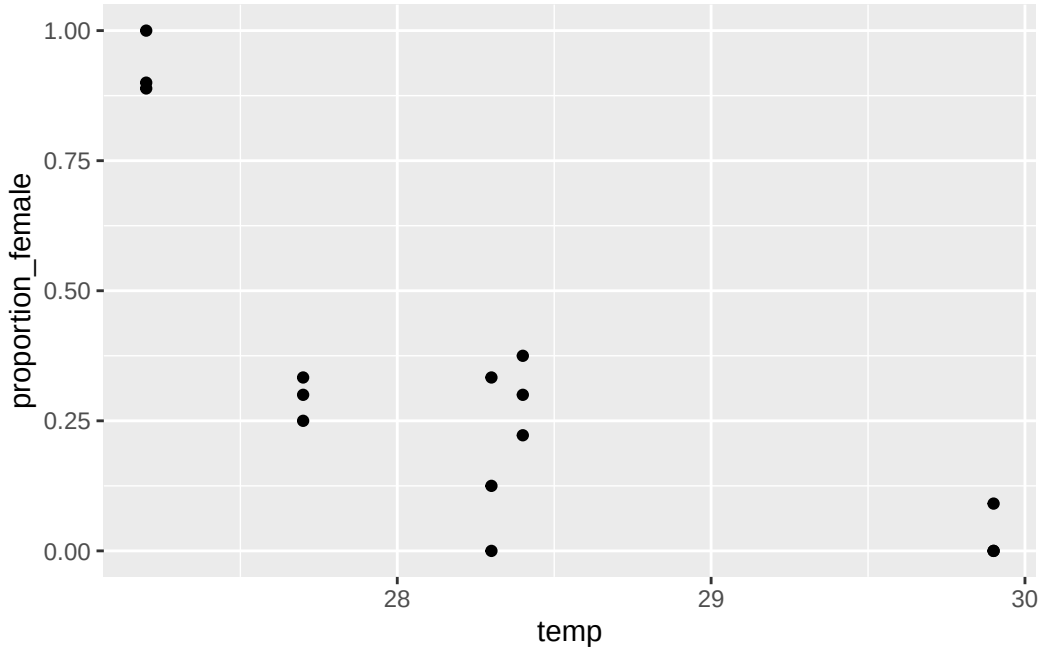


1.3 GLM with a logit link

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()
```




```
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:
```

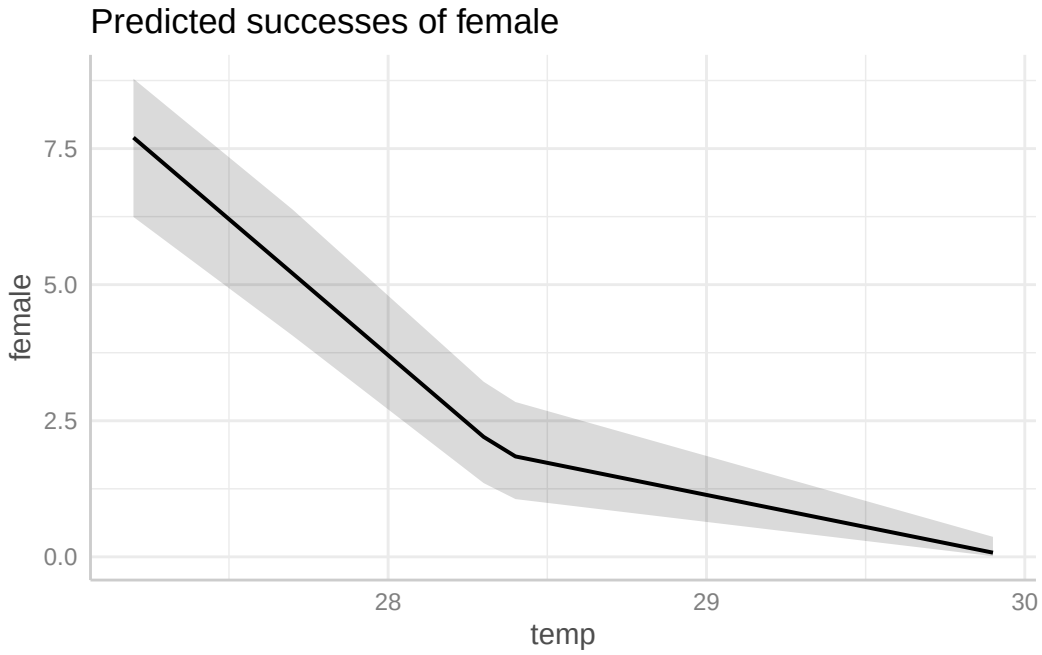
```
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    62.75     11.92    40.58    86.97 1.00     4347     6472
temp         -2.26      0.43    -3.13    -1.47 1.00     4294     6358

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).
```

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

```
Warning in check_dep_version(dep_pkg = "TMB"): package version mismatch:
glmmTMB was built with TMB package version 1.9.18
Current TMB package version is 1.9.16
Please re-install glmmTMB from source or restore original 'TMB' package (see '?reinstalling'
```

```
plot(pred)
```



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: parental income

   Random: State and district

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

   Fixed: latitude, density of sea urchins

   Random: site, transect number

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

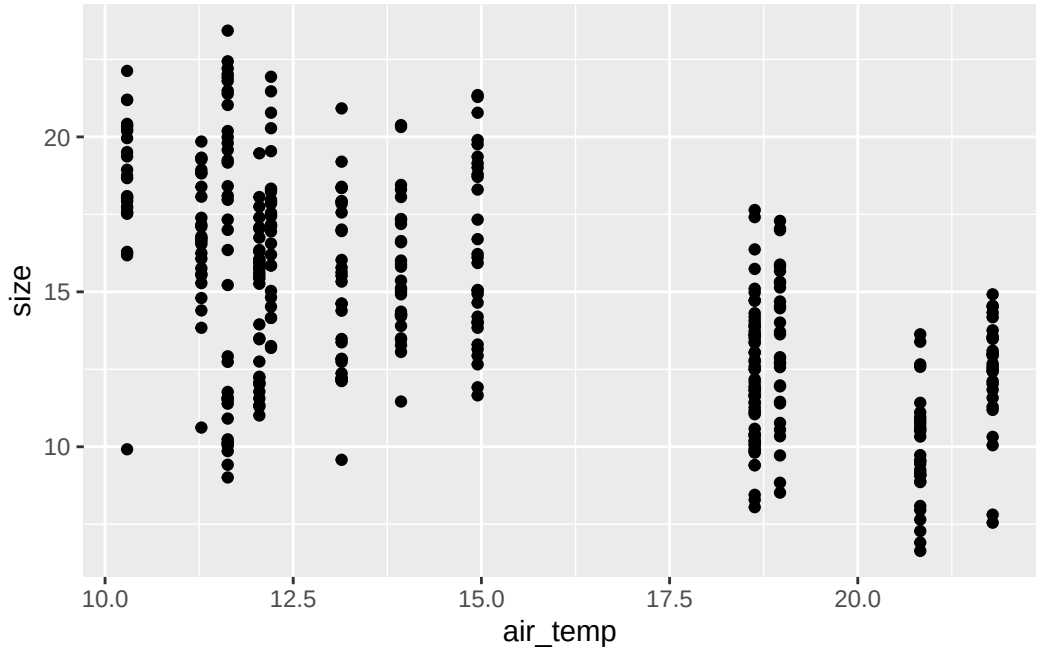
   Fixed: temperature, age

4. Random: individual Id, year

Fiddler Crabs

```
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     0.056    3.246    3.467 1.000     4372     3105
water_temp   -0.038     0.003   -0.045   -0.032 1.000     4570     3015

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   23.259     1.619   20.191   26.516 1.004     2931     2625

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)    0.123     0.033    0.075    0.208 1.002      897     1288

Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Intercept     3.368     0.181    3.014    3.734 1.002     1058     1594
water_temp   -0.039     0.010   -0.059   -0.020 1.001     1158     1744
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   30.075     2.167   25.932   34.506 1.002     2738     2321

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?

Interpret the output by writing a 2-3 sentence results paragraph that answers:

1. What is the effect of water_temp on creab 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.
2. Does it seem like the slope estimate is different from zero? Why?

Q2.3 Compare WAIC and PSIS of the two models

Between the two models, which has the better predictive power?

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

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

         Estimate   SE
elpd_loo   -984.7 12.1
p_loo         2.8  0.2
looic      1969.5 24.3
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.7, 1.1]).

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        12.0  0.8
looic      1876.9 27.3
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 1.6]).

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

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

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

          Estimate   SE
elpd_waic   -984.7 12.1
p_waic         2.8  0.2
waic        1969.4 24.3
```

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

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

          Estimate   SE
elpd_waic   -938.4 13.7
p_waic        12.0  0.8
waic        1876.8 27.3
```

The second model that includes site as a random effect is the better model (lower loo and waic values)

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

```
                          type = "random")

plot(preds)
```