StatisticalReasoning6_generalized_linear_and_multilevel_models

Yeonu&Sam

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

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.0
-- 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 becom
```

```
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
#install.packages('modelbased')
#install.packages('faraway')
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
```

```
?brmsfamily
```

```
starting httpd help server ... done
```

Q1.1a

1. Counts of Clarkia flowers in a meadow
   non-negative integers
2. Whether or not a female elephant seal gives birth
   0's/1's
3. The percent cover of red algae in the intertidal
   fractions
4. Growth of a tree from one year to the next
   positive real numbers

5. The spatial area of a forest in square meters
   positive real numbers

Q1.1b

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

Q1.2

1. Head width and Forelimb length of Kangaroo Rat
2. positive real numbers
3. Gamma distribution

1.2 GLM with a log link

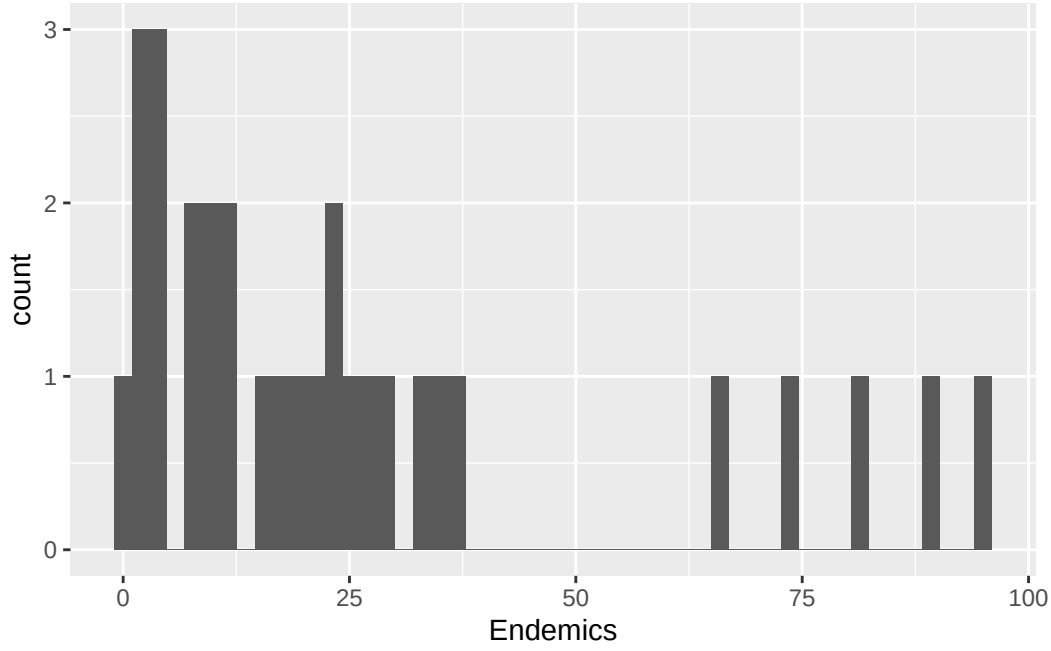
```
# Read in the pre-stored data
data("gala")
# Check out the first 6 rows
head(gala)
```

```
             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

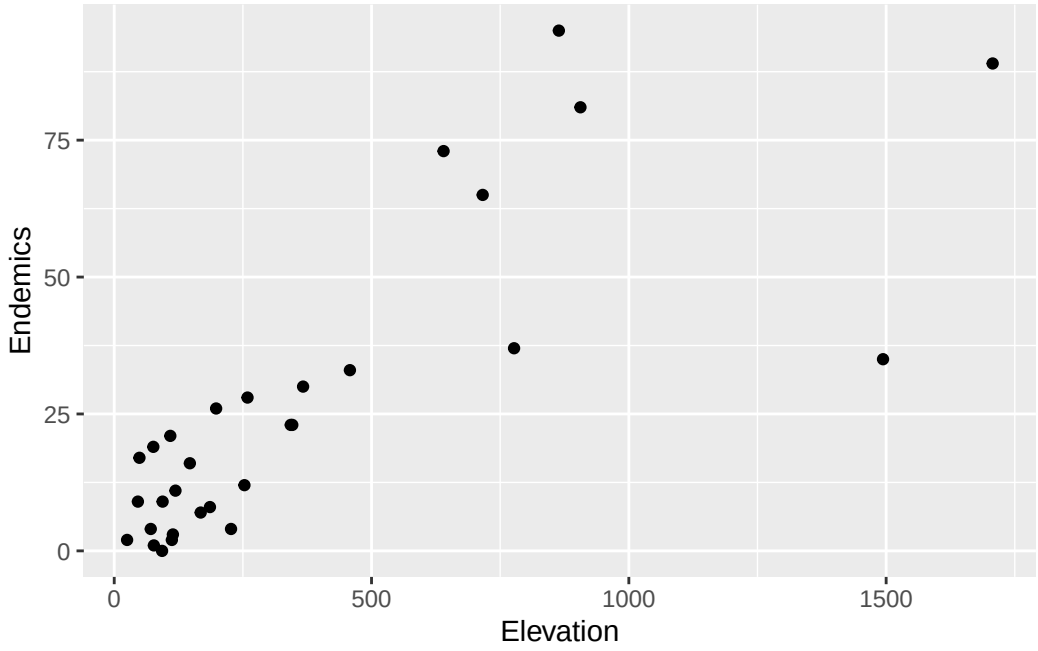
```
ggplot(data=gala,
       aes(x=Endemics))+
  geom_histogram(bins=50)
```



It doesn't seem like a Gaussian distribution, because it is not symmetric. The values of x-axis are non-negative integers and it has long tail.

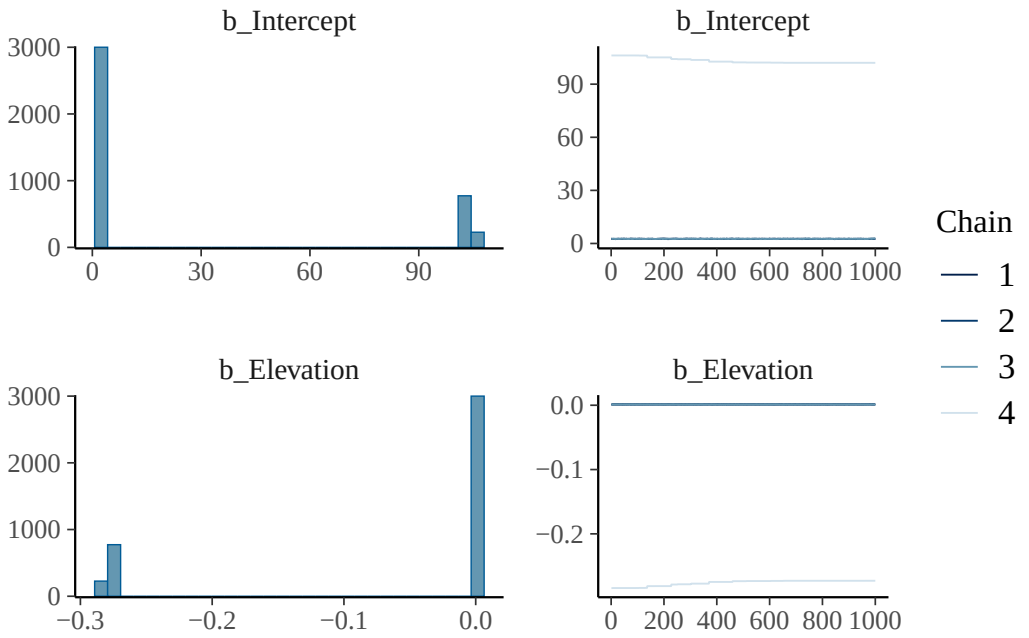
Q1.4

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

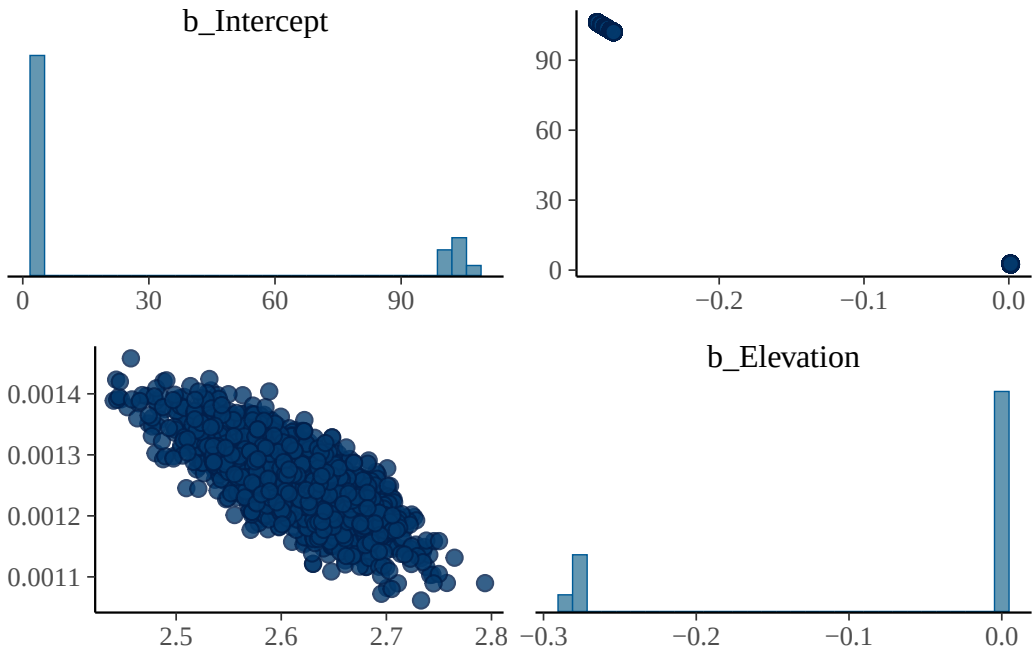



```
# Endemics ~ Elevation
m.elev <-
  brm(data = gala, # Give the model the penguins data
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),
      # Specify the model here.
      Endemics ~ 1 + Elevation,
      # Here's where you specify parameters for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

```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    27.77     43.60     2.50   106.21 1.60        7       11
Elevation    -0.07      0.12    -0.28     0.00 1.60        7       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

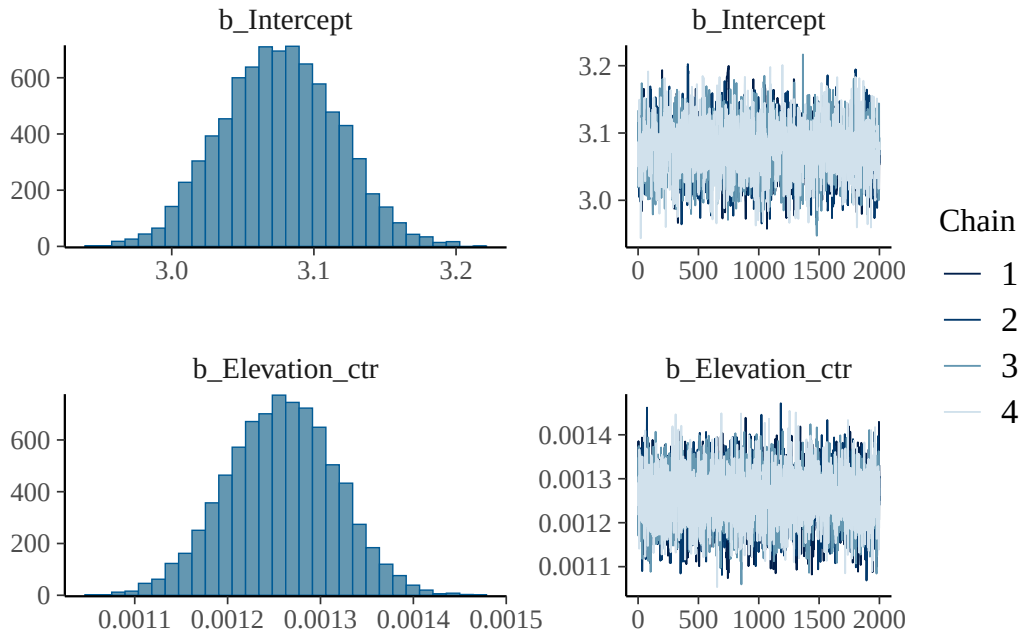
No. The posterior distribution doesn't have one clear peak and the chains don't overlap. Also, the biggest R hat is 1.60 which is far from 1.00

Q1.6

```r
library(dplyr)
# 1) making new Elevation_ctr column
gala_cen_ele<-gala %>%
  mutate(
    Elevation_ctr=Elevation-mean(Elevation)
  )

# 2) change the brm chain argumments

# Endemics ~ Elevation
```

```
m.elev2 <-
  brm(data = gala_cen_ele, # Give the model the penguins data
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),
      ###ChatGPT for writing the prior code
      prior=c(
        prior(normal(100,50),class="Intercept"),
        prior(normal(0,1),class=b)
      ),
      # Specify the model here.
      Endemics ~ 1 + Elevation_ctr,
      # 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,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.elev2")
```

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

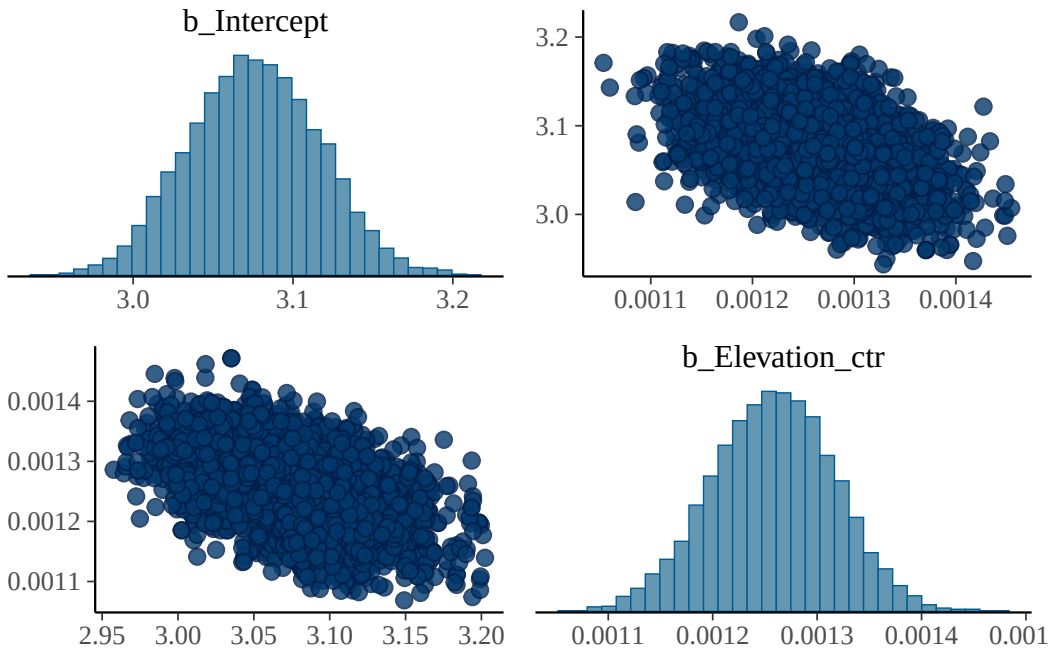



```
summary(m.elev2)
```

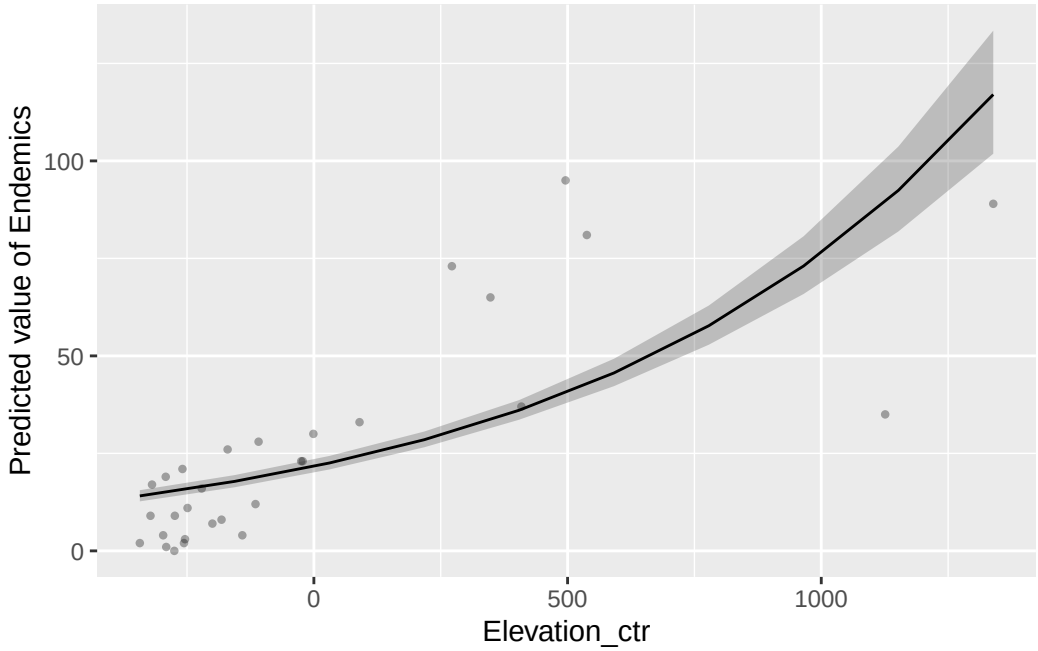
```
 Family: poisson
```

```
  Links: mu = log
Formula: Endemics ~ 1 + Elevation_ctr
   Data: gala_cen_ele (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         3.08      0.04     3.00     3.15 1.00     1710     1841
Elevation_ctr     0.00      0.00     0.00     0.00 1.00     3773     4707

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).
```

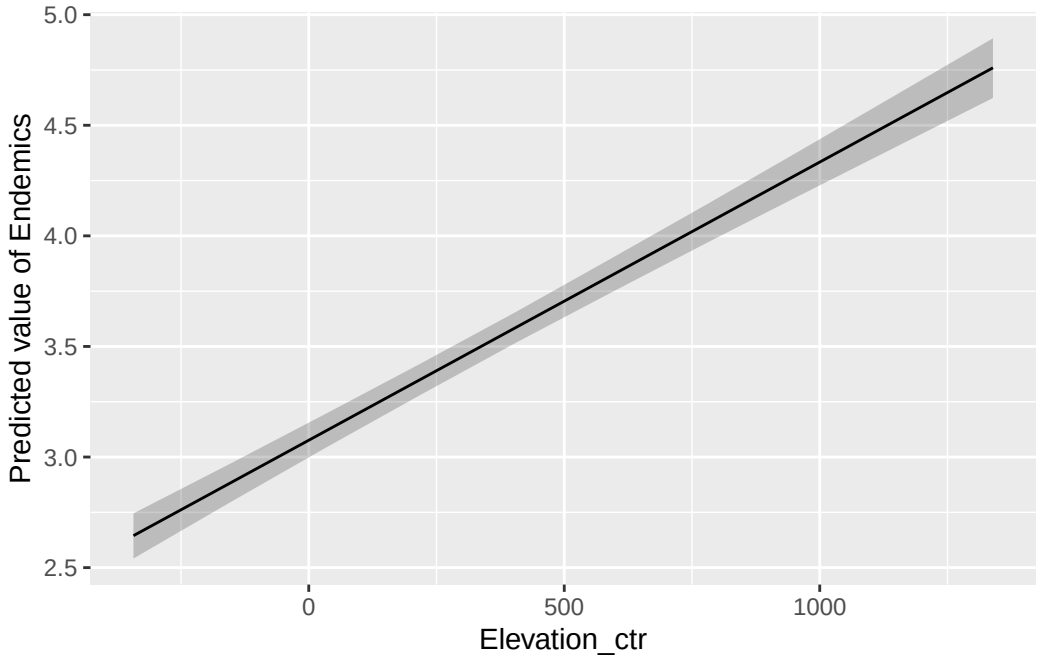
```r
pairs(m.elev2)
```




```r
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)
```



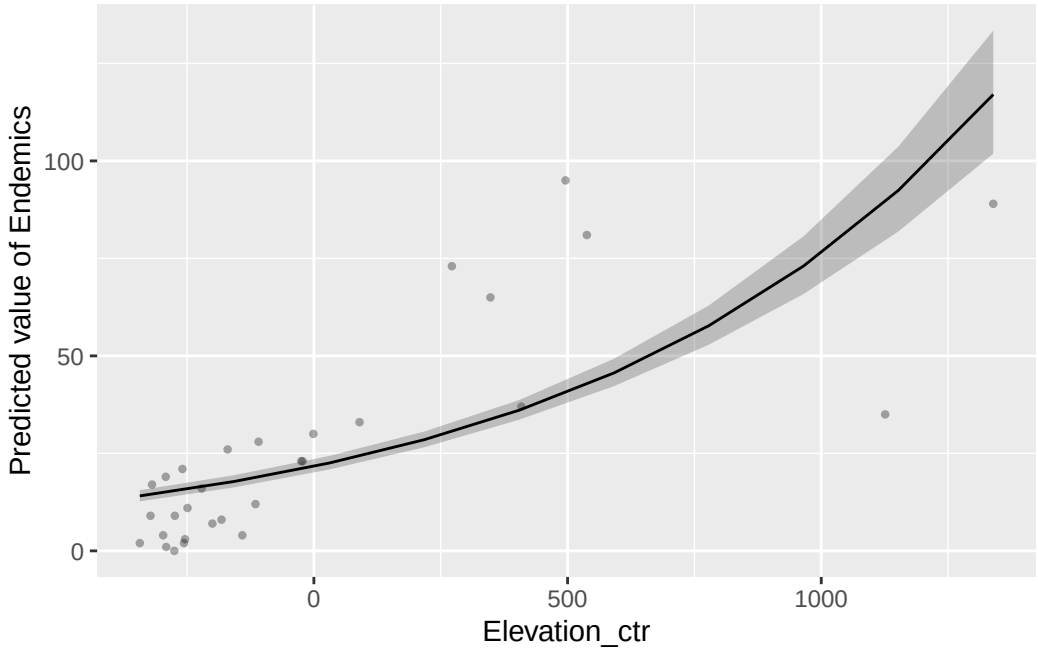
```
print(m.elev2, digits = 4)
```

```
 Family: poisson
  Links: mu = log
Formula: Endemics ~ 1 + Elevation_ctr
   Data: gala_cen_ele (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       3.0759    0.0406   2.9987   3.1548 1.0021     1710     1841
Elevation_ctr   0.0013    0.0001   0.0011   0.0014 1.0011     3773     4707

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).
```

```
exp(0.0013)
```

```
[1] 1.001301
```

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



Q1.7

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

```
exp(1.09)
```

```
[1] 2.974274
```

   For every 1 degree Celsius, number of Clarkias blooming increase by 197.4274%.

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 1 sea otter, density of sea urchins per square meter in a quadrat decreases by -91.7915%.

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

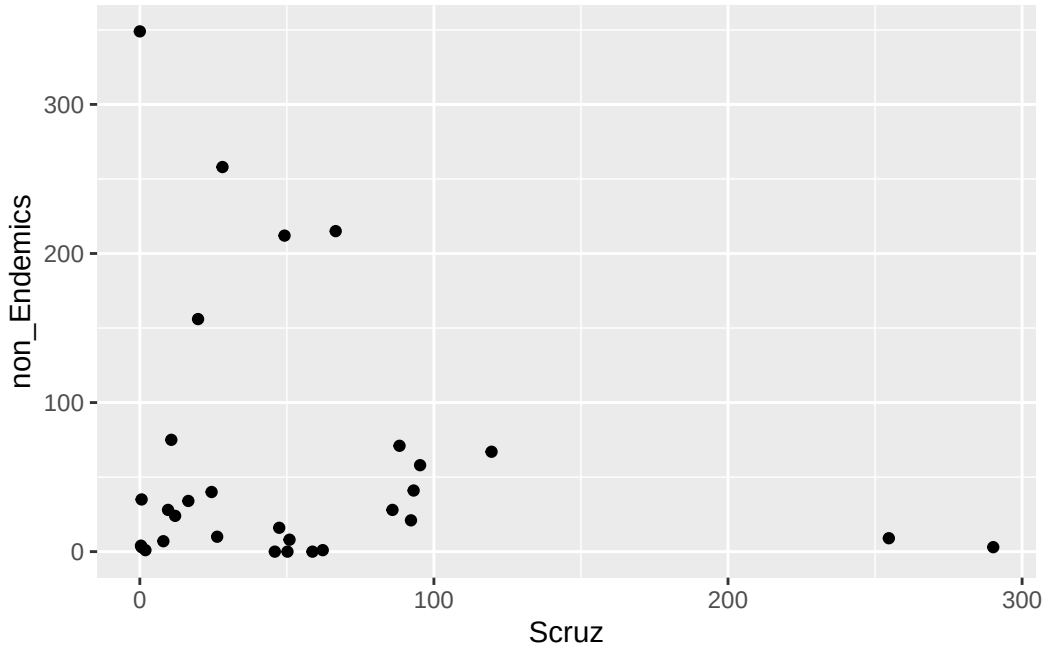
```
exp(6.24)
```

```
[1] 512.8585
```

For every 1 kg of fertilizer, number of tomatoes per plant increases by 51185.85%.

Q1.8

```
# 1) making new non-endemic column
gala_cen_ele<-gala %>%
  mutate(
    non_Endemics=Species-Endemics
  )


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



Q1.9

```
m.dis_end <-
  brm(data = gala_cen_ele, # Give the model the penguins data
```

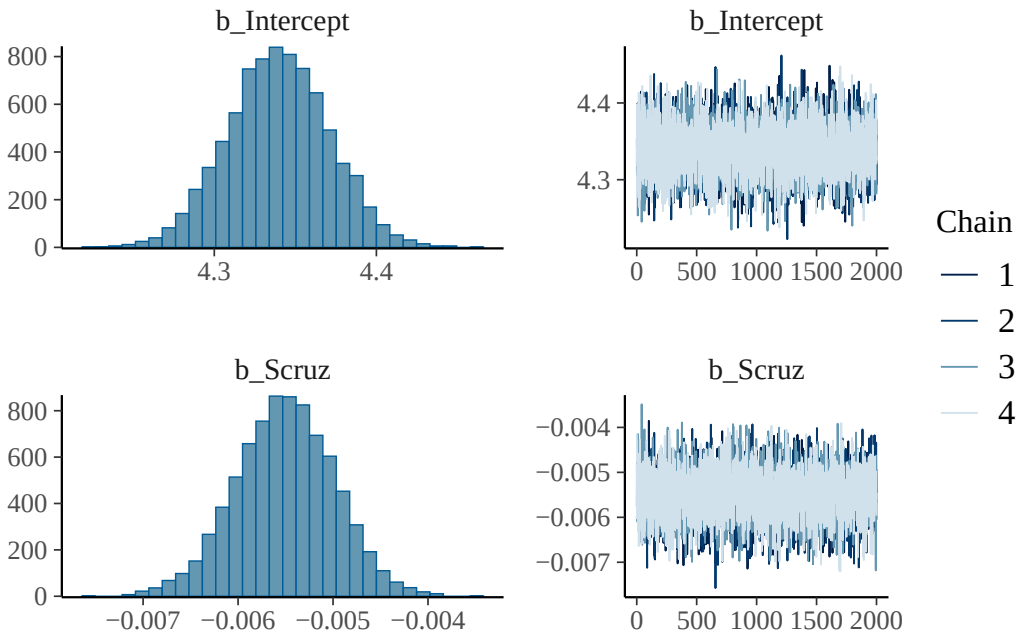
```
      # Choose a poisson distribution - THIS IS THE NEW PART!
      family = poisson(link = "log"),

      # 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,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.dis_end")
```

Q1.10

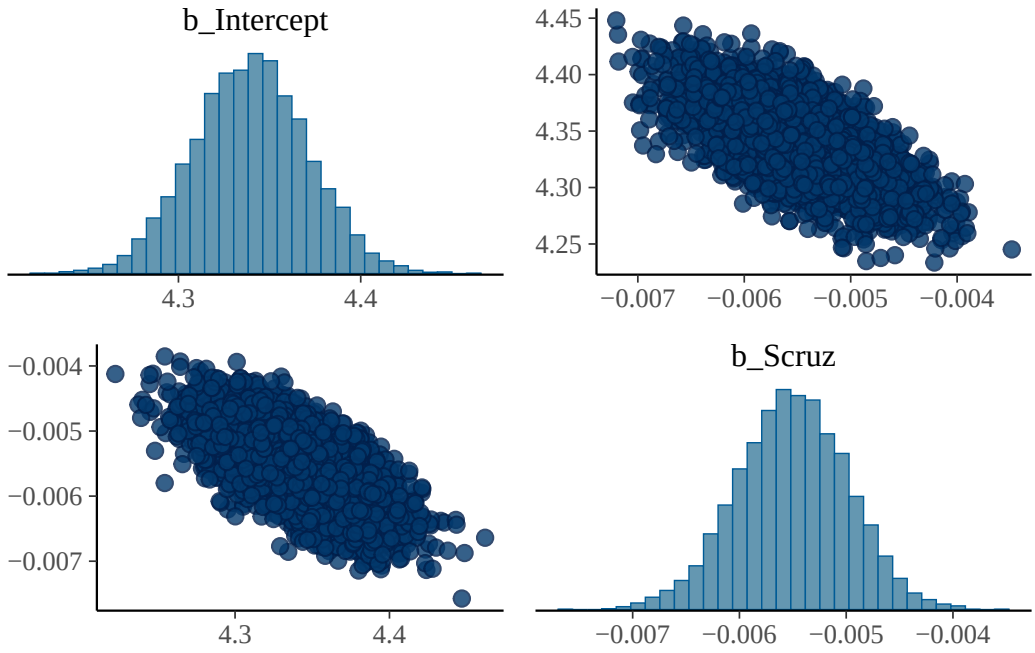
```
print(m.dis_end,digits=4)
```

```
 Family: poisson
  Links: mu = log
Formula: non_Endemics ~ 1 + Scruz
   Data: gala_cen_ele (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   4.3396    0.0314   4.2785   4.4003 1.0006     6180     5972
Scruz      -0.0055    0.0005  -0.0066  -0.0045 1.0002     5697     4783

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.dis_end)
```




```
pairs(m.dis_end)
```



R hat equals 1.00, the chains overlap, and there is one clear peak in posterior distributions. Therefore, the model ran correctly.

Q1.11

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) -0.0055

   b)

   ```
   exp(-0.0055)
   ```

   ```
   [1] 0.9945151
   ```

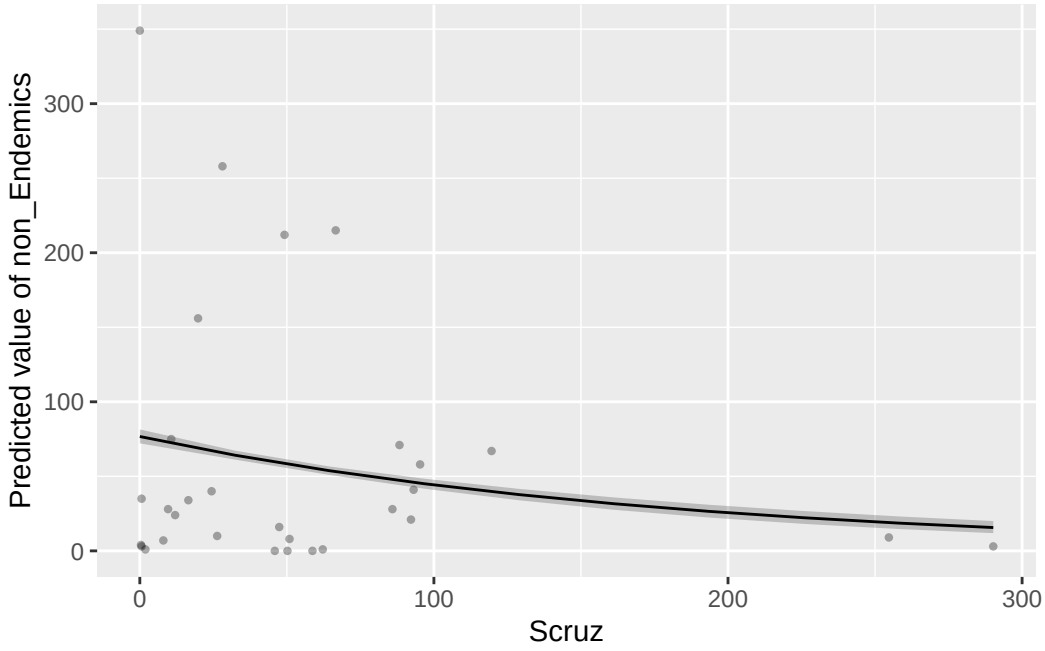
   c)
   For every 1km distance from Santa Cruz, the number of non-endemic species decreases by -0.54849%.

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

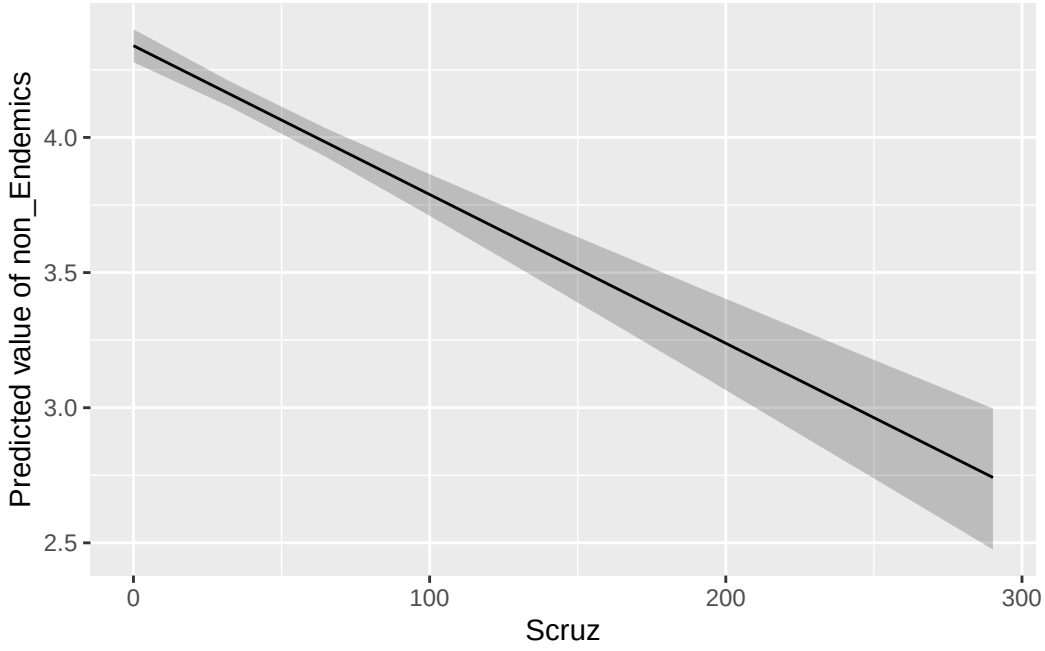
   Yes. Because the 95% CI ranges from -0.0066 to -0.0045 which not include zero.

Q1.12

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



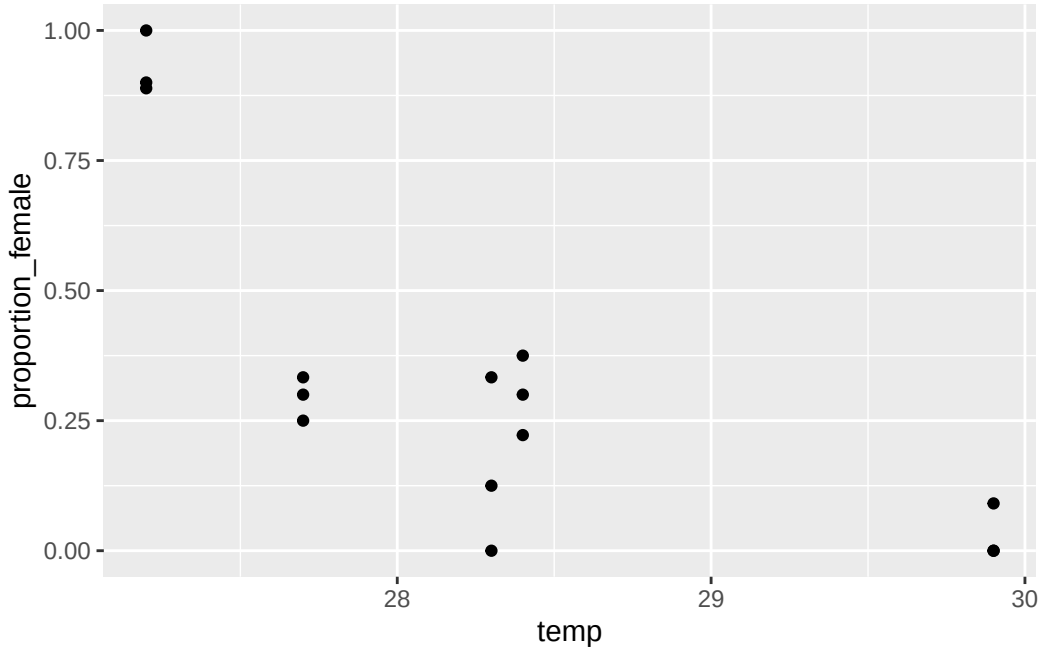

```
predslog <- estimate_expectation(m.dis_end, by = 'Scruz', predict = 'link')
plot(predslog)
```



1.3 GLM with a logit link

```
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")
```

```r
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    62.93     11.99    41.03    87.62 1.00     4187     5511
temp         -2.27      0.43    -3.15    -1.48 1.00     4120     5492

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).
```

2. Multilevel models

Q2.1

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

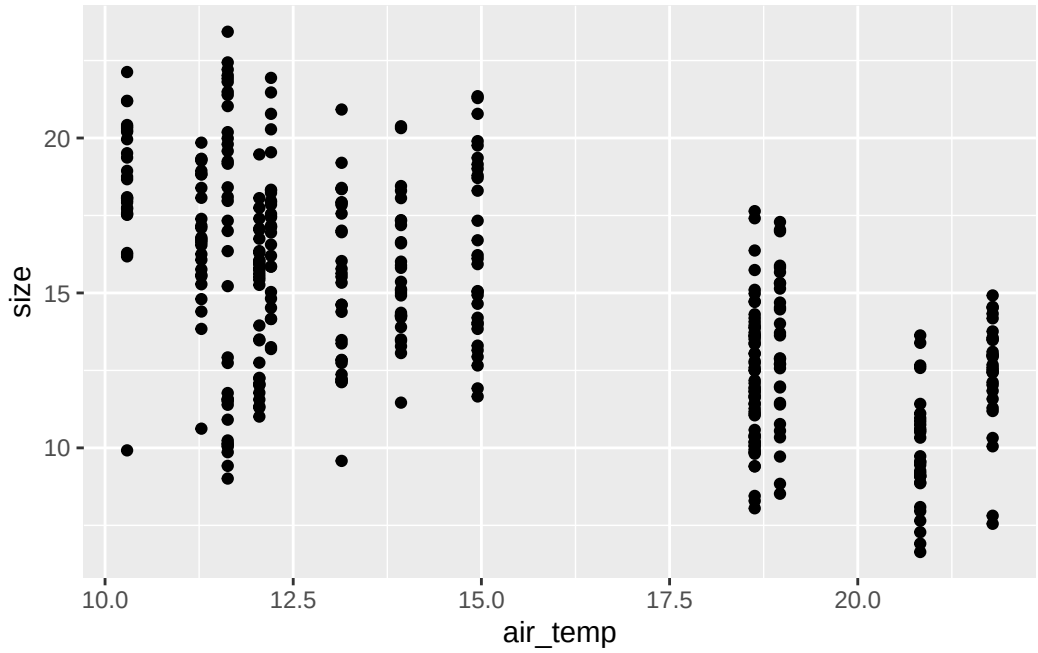
   -fixed effects: parental income
   -random effects: state of residence, school district

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

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

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

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




```r
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     3.354     0.055    3.248    3.463 1.000     4762     3335
water_temp   -0.038     0.003   -0.044   -0.032 1.001     4955     3208

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   23.273     1.668   20.215   26.650 1.002     2885     2540

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     0.035    0.076    0.204 1.006      763     1396
```

```
Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
Intercept     3.361     0.195    2.965    3.757 1.002     1102     1276
water_temp   -0.039     0.011   -0.060   -0.017 1.002     1119     1228

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI  Rhat Bulk_ESS Tail_ESS
shape   30.083     2.178   26.004   34.537 1.001     3159     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).
```

Q2.2

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.

   a) -0.039

   b)

   ```
   exp(-0.039)
   ```

   ```
   [1] 0.9617507
   ```

   c)

   ```
   (0.9617507-1)*100
   ```

   ```
   [1] -3.82493
   ```

As 1 degree Celsius water temperature increases, the carapace width of a crab decreases by -3.82493%.

2. It's different from zero. Because 95% CI ranges from -0.061 to -0.017 which not include zero.

Q2.3

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

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

         Estimate   SE
elpd_loo   -984.7 12.2
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   -938.4 13.7
p_loo        12.1  0.8
looic      1876.8 27.3
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.6, 1.7]).

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 12.2
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
```

Because m.watertemp.site has lower WAIC and PSIS value. m.watertemp.site is better predictive model.