Activity 9: Statistical reasoning 1: intro to models

Joshua Holt and Jacob Chung

1. Fiddler crabs

```
library(brms) # for statistics
```

```
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(tidyverse)
```

```
-- 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   4.0.0     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
```

```
-- 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 becor
```

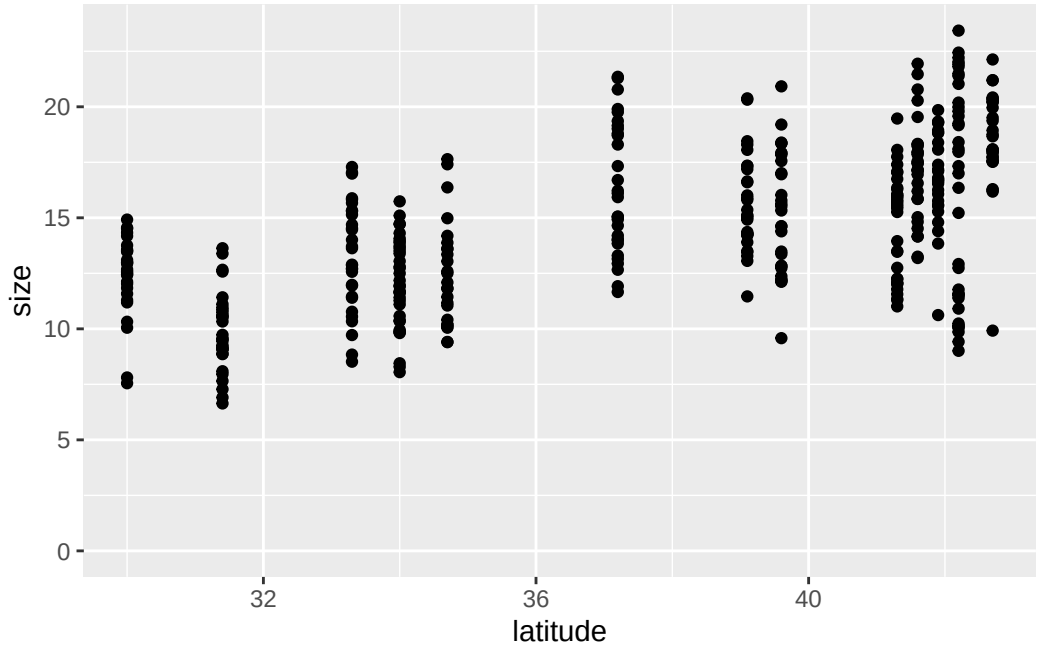
```
library(ggeffects) # for the prediction plot
library(lterdatasampler) # for built-in datasets
```

```
head(pie_crab)
```

```
# A tibble: 6 x 9
  date       latitude site   size air_temp air_temp_sd water_temp water_temp_sd
  <date>        <dbl> <chr> <dbl>    <dbl>       <dbl>      <dbl>         <dbl>
1 2016-07-24       30 GTM    12.4     21.8        6.39       24.5          6.12
2 2016-07-24       30 GTM    14.2     21.8        6.39       24.5          6.12
3 2016-07-24       30 GTM    14.5     21.8        6.39       24.5          6.12
4 2016-07-24       30 GTM    12.9     21.8        6.39       24.5          6.12
5 2016-07-24       30 GTM    12.4     21.8        6.39       24.5          6.12
6 2016-07-24       30 GTM    13.0     21.8        6.39       24.5          6.12
# i 1 more variable: name <chr>
```

1.1 Plot data, pick the model

```
pie_crab %>%
  ggplot(aes(x = latitude, y = size)) +
  geom_point() +
  # Make the y-axis include 0
  ylim(0, NA)
```

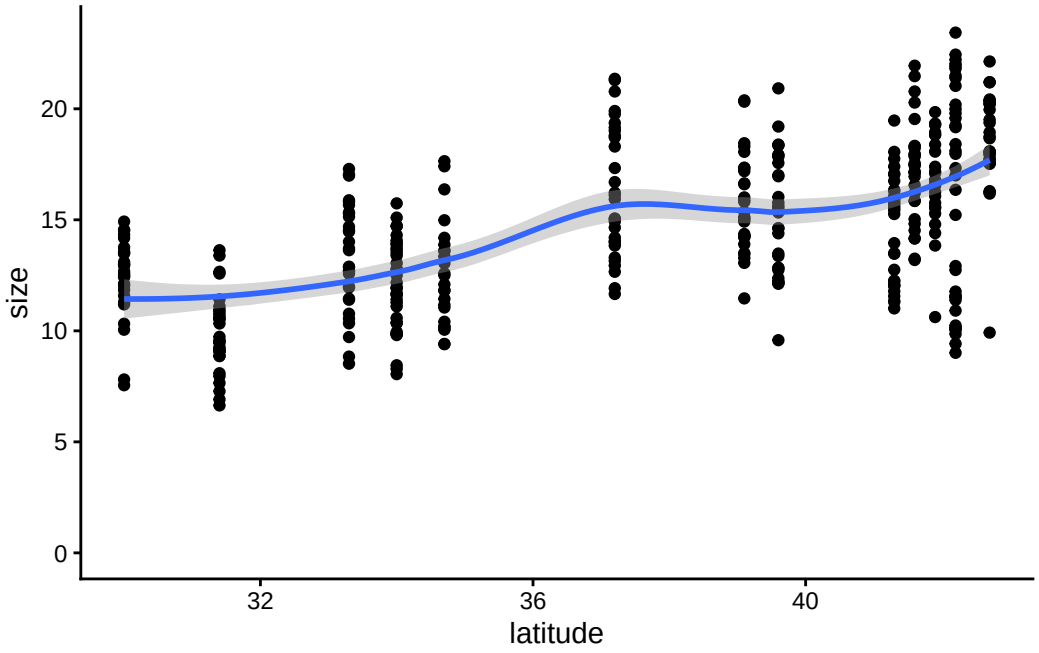


Q1.1) I would say that size does increase with latitude. We are not very confident in this interpretation, due to potential bias in sampling.

Q1.2)

```
pie_crab %>%
  ggplot(aes(x = latitude, y = size)) +
  geom_point() +
  # Make the y-axis include 0
  ylim(0, NA) +
  geom_smooth()+
  theme_classic()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



1.2 Fit linear regression with brms

```
# latitude model
m.crab.lat <-
  brm(data = pie_crab, # Give the model the pie_crab data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      size ~ latitude,
      # 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,
      # Setting the "seed" determines which random numbers will get sampled.
      # In this case, it makes the randomness of the Markov chain runs reproducible
      # (so that both of us get the exact same results when running the model)
      seed = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.crab.lat")
```

Q1.3) The iter tells the computer the number of total iterations per chain (including warmup; defaults to 2000).

1.3 Assess model

```
summary(m.crab.lat)
```

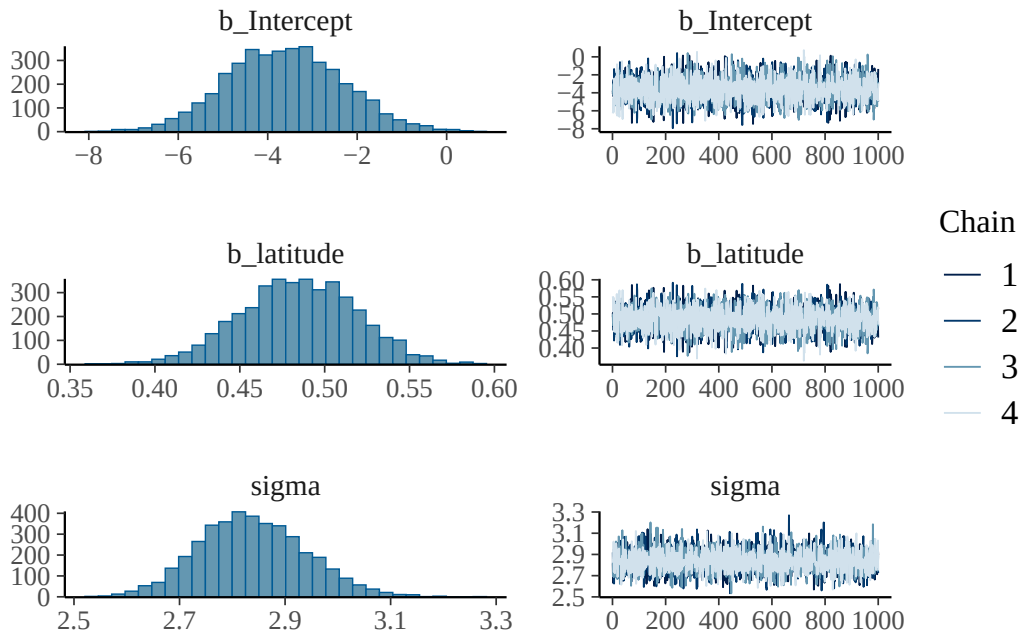
```
 Family: gaussian
  Links: mu = identity
Formula: size ~ latitude
   Data: pie_crab (Number of observations: 392)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -3.61      1.30    -6.09    -1.01 1.00     4116     3192
latitude      0.48      0.03     0.42     0.55 1.00     4108     3140

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     2.84      0.10     2.65     3.04 1.00     3758     2852

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.crab.lat) # show posteriors and chains
```



1.4 Interpret model

```
summary(m.crab.lat)
```

```
 Family: gaussian
  Links: mu = identity
Formula: size ~ latitude
   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    -3.61      1.30    -6.09    -1.01 1.00     4116     3192
latitude      0.48      0.03     0.42     0.55 1.00     4108     3140

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     2.84      0.10     2.65     3.04 1.00     3758     2852

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).
```

```
as_draws_df(m.crab.lat)  %>%  # extract the posterior samples from the model estimate
  select(b_latitude)  %>%  # pull out the latitude samples from all 4 chains. we'll get a wa
  summarize(p_slope_lessthanorequalto_zero = sum(b_latitude <= 0)/length(b_latitude))
```

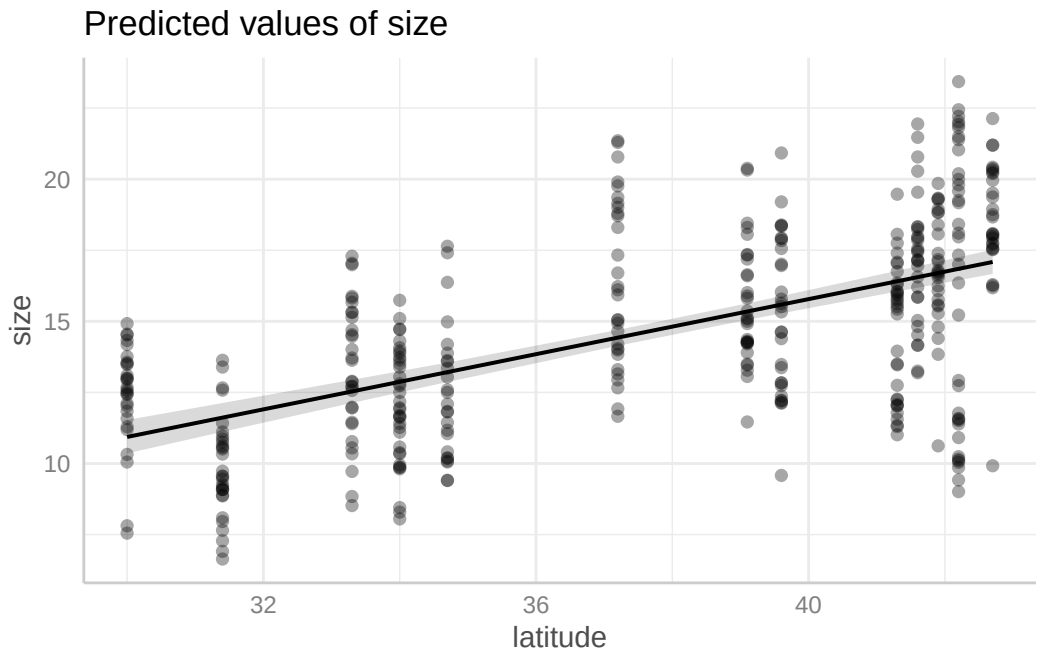
```
Warning: Dropping 'draws_df' class as required metadata was removed.
```

```
# A tibble: 1 x 1
  p_slope_lessthanorequalto_zero
                           <dbl>
1                              0
```

1.5 Plot model on the data

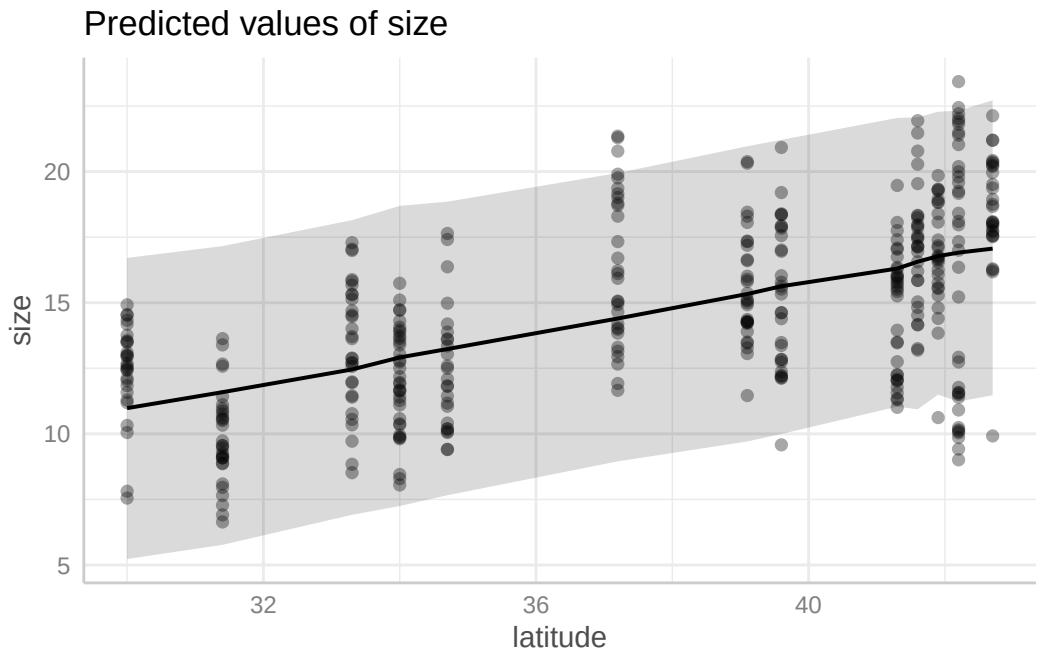
```
# compatibility interval. the shows uncertainty in the average response.
confm.crab.lat <- predict_response(m.crab.lat)
plot(confm.crab.lat, show_data = TRUE)
```

```
Data points may overlap. Use the `jitter` argument to add some amount of
  random variation to the location of data points and avoid overplotting.
```




```
# prediction interval. this shows uncertainty in the data around the average response.
confm.crab.lat <- predict_response(m.crab.lat, interval = 'prediction')
plot(confm.crab.lat, show_data = TRUE)
```

```
Data points may overlap. Use the `jitter` argument to add some amount of
  random variation to the location of data points and avoid overplotting.
```



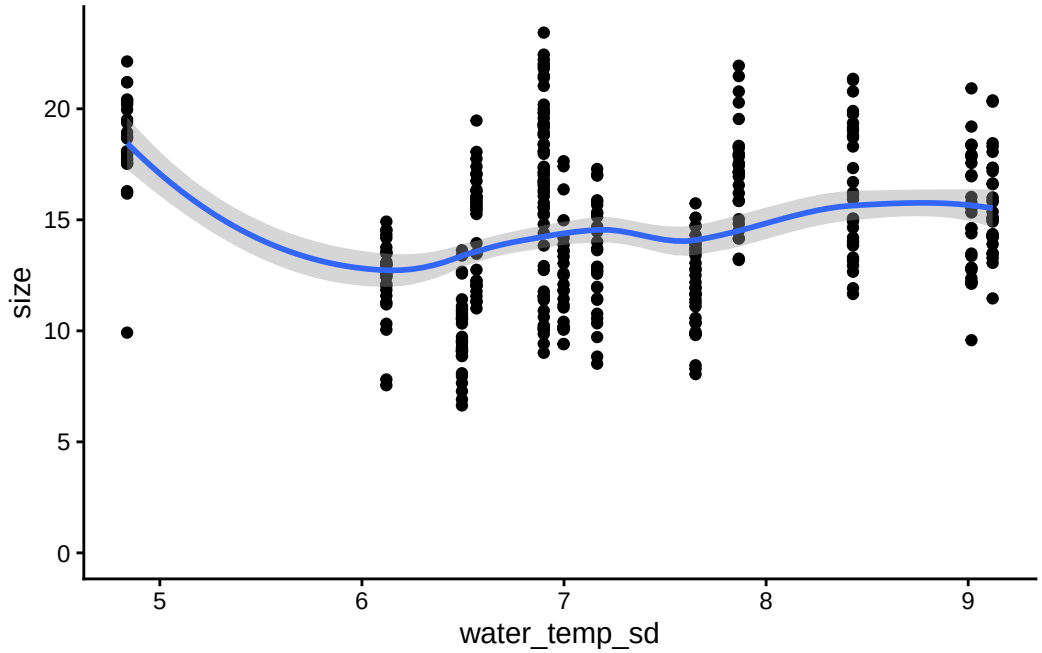
1.6 Repeat with new variable: water temp sd

Q1.4) We predict that higher variability will be observed in smaller crabs, with the larger crabs showing lower variability. This is due to the expectation that larger crabs will handle water temperature variability the best.

Q1.5)

```
pie_crab %>%
  ggplot(aes(x = water_temp_sd, y = size)) +
  geom_point() +
  # Make the y-axis include 0
  ylim(0, NA) +
  geom_smooth()+
  theme_classic()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



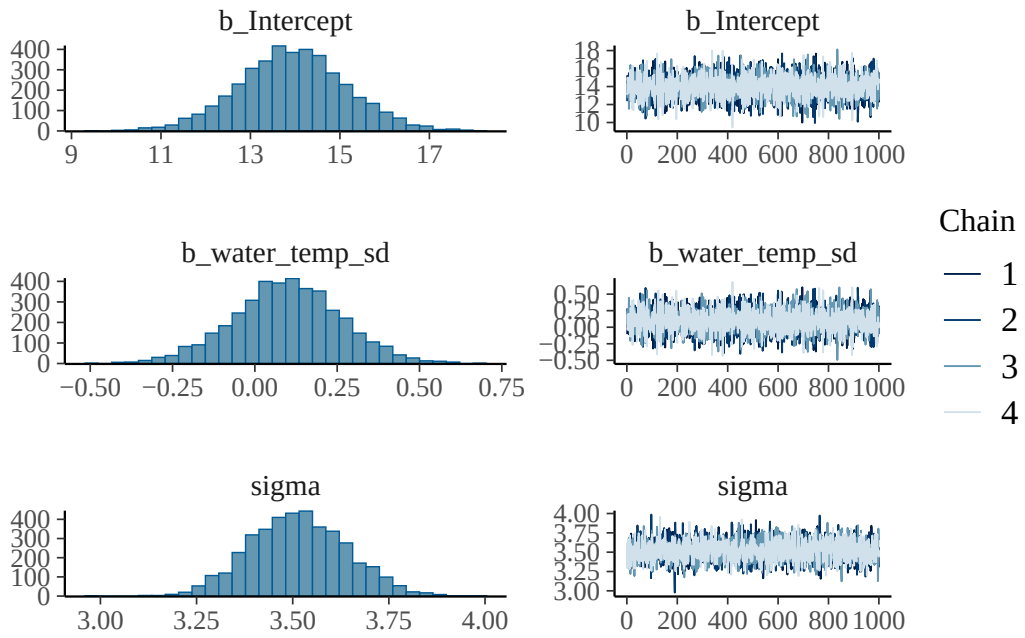
Q1.6) It does not look like size changes with the sd of water temperature. We are moderately confident with this assessment due to the lack of slope in the regression line.

Q1.7)

```
# water temp sd model
m.crab.watersd <-
  brm(data = pie_crab, # Give the model the pie_crab data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      size ~ water_temp_sd,
      # 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,
      # Setting the "seed" determines which random numbers will get sampled.
      # In this case, it makes the randomness of the Markov chain runs reproducible
      # (so that both of us get the exact same results when running the model)
      seed = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/m.crab.watersd")
```

Q1.8) Our model ran correctly because our Rhat is 1 and based on our MCMC graphs, our models all came to the same conclusion.

```
# show posteriors and chains
plot(m.crab.watersd)
```




```
# show summary, including rhat
summary(m.crab.watersd)
```

```
 Family: gaussian
  Links: mu = identity
Formula: size ~ water_temp_sd
   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        13.95      1.21    11.54    16.36 1.00     4569     2971
water_temp_sd     0.10      0.16    -0.23     0.42 1.00     4615     2950
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     3.52      0.13     3.28     3.76 1.00     3960     2894

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).
```

Q1.9)

1. We found that standard deviation has a miniscule effect on body size. With every 1 standard deviation in water temperature, we see an increase of 0.1 mm of carapace width.

2. The effect is not reasonably differentthan 0 because our confidence interval includes 0 (-0.23 to 0.42).

2. Back to Pikas!

```
head(nwt_pikas)
```

```
# A tibble: 6 x 8
  date       site      station utm_easting utm_northing sex   concentration_pg_g
  <date>     <fct>     <fct>         <dbl>        <dbl> <fct>              <dbl>
1 2018-06-08 Cable Ga~ Cable ~      451373      4432963 male               11563.
2 2018-06-08 Cable Ga~ Cable ~      451411      4432985 male               10629.
3 2018-06-08 Cable Ga~ Cable ~      451462      4432991 male               10924.
4 2018-06-13 West Kno~ West K~      449317      4434093 male               10414.
5 2018-06-13 West Kno~ West K~      449342      4434141 male               13531.
6 2018-06-13 West Kno~ West K~      449323      4434273 <NA>                7799.
# i 1 more variable: elev_m <dbl>
```

```
nwt_pikas_doy <- nwt_pikas %>%
  # Add a new column called day_of_year
  # yday extracts the day of year from the date column
  mutate(day_of_year = yday(date)) %>%
  # relocate the day_of_year column after the date column
  relocate(day_of_year, .after = date)
```

```
head(nwt_pikas_doy)
```

```
# A tibble: 6 x 9
  date       day_of_year site       station      utm_easting utm_northing sex
  <date>           <dbl> <fct>      <fct>              <dbl>        <dbl> <fct>
1 2018-06-08         159 Cable Gate Cable Gate 1      451373      4432963 male
2 2018-06-08         159 Cable Gate Cable Gate 2      451411      4432985 male
3 2018-06-08         159 Cable Gate Cable Gate 3      451462      4432991 male
4 2018-06-13         164 West Knoll West Knoll 3      449317      4434093 male
5 2018-06-13         164 West Knoll West Knoll 4      449342      4434141 male
6 2018-06-13         164 West Knoll West Knoll 5      449323      4434273 <NA>
# i 2 more variables: concentration_pg_g <dbl>, elev_m <dbl>
```

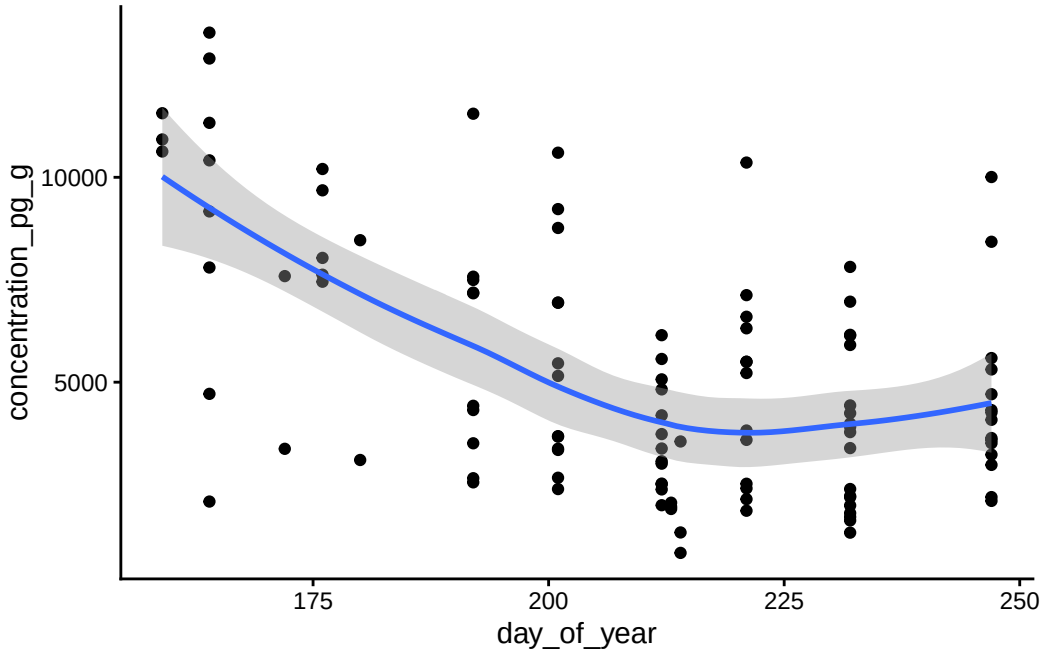
Q2.1) How does stress vary based on the day of the year?

Q2.2) We expect pika stress to increase as the year progresses. We think think this because as the year progresses, temperature increases making the pikas more stressed.

Q2.3)

```
nwt_pikas_doy |>
  ggplot(aes(x = day_of_year, y = concentration_pg_g))+
  geom_point()+
  theme_classic()+
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

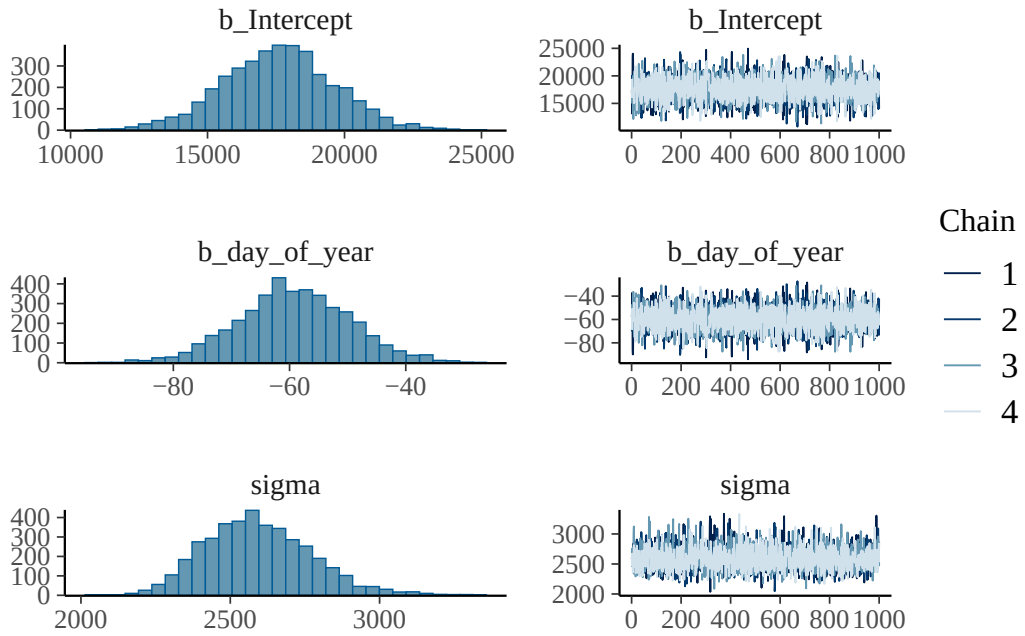


Q2.4)

```
# Pika stress over year model
pika.s.day <-
  brm(data = nwt_pikas_doy,
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      concentration_pg_g ~ day_of_year,
      # 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,
      # Setting the "seed" determines which random numbers will get sampled.
      # In this case, it makes the randomness of the Markov chain runs reproducible
      # (so that both of us get the exact same results when running the model)
      seed = 4,
      # Save the fitted model object as output - helpful for reloading in the output later
      file = "output/pika.s.day")
```

Q2.5) We believe the model ran correctly because our Rhat is equal to 1 and our MCMC all overlap.

```
# show posteriors and chains
plot(pika.s.day)
```




```
# show summary, including rhat
summary(pika.s.day)
```

```
 Family: gaussian
  Links: mu = identity
Formula: concentration_pg_g ~ day_of_year
   Data: nwt_pikas_doy (Number of observations: 109)
  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   17607.56   2056.61 13452.87 21661.70 1.00     3755     2585
day_of_year   -59.17      9.73   -78.41   -39.66 1.00     3708     2696
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma  2594.85    180.48  2282.99  2985.71 1.00     3668     2754

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).
```

Q2.6)

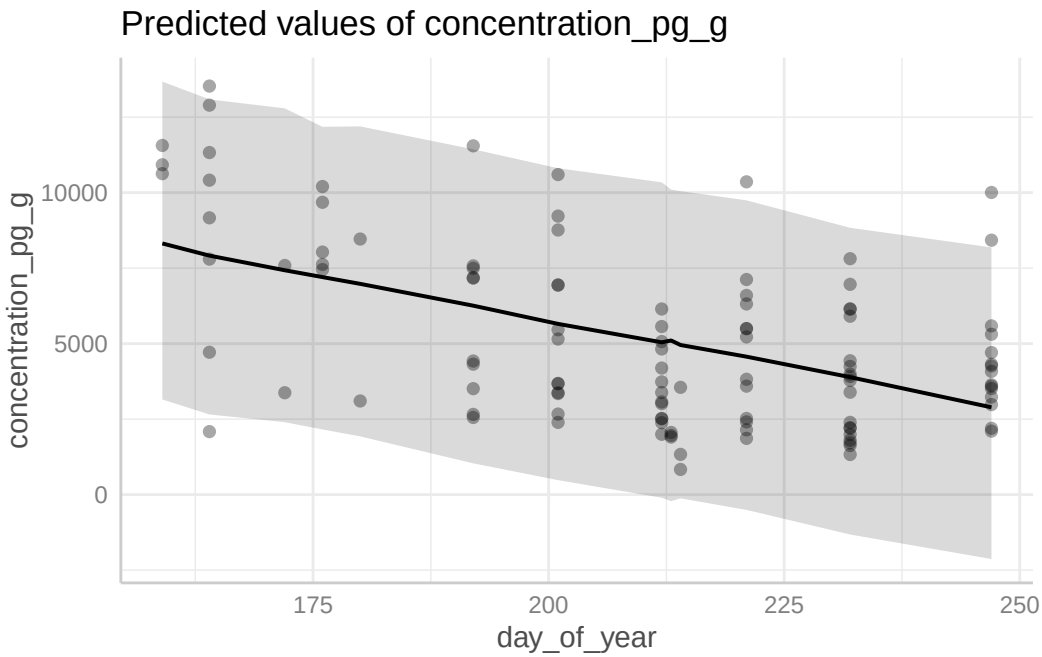
1. For every 1 day increase, there is a -59.17 picogram per gram (parts per trillion) concentration change in Pika stress.

2. Our effect is reasonably different than 0 because our confidence interval doesn't include 0 (-78.41, -39.66).

Q2.7) Prediction Interval

```r
# prediction interval. this shows uncertainty in the data around the average response.
pred.pika.day <- predict_response(pika.s.day, interval = 'prediction')
plot(pred.pika.day, show_data = TRUE)
```

```
Data points may overlap. Use the `jitter` argument to add some amount of
  random variation to the location of data points and avoid overplotting.
```



Q2.8) We found a decrease of (-)59.17 picogram per gram of stress concentration per one additional day of the year. Since our 95% credible intervals does not include 0 (-78.41, -39.66), this suggests that given our model, Pika stress decreases as the year progresses.