Statistical reasoning 1: intro to models

Alex Stadlinger & Emily Halim

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

```
Loading required package: Rcpp
```

```
Loading 'brms' package (version 2.22.0). 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   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
```

```
-- 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 beco
```

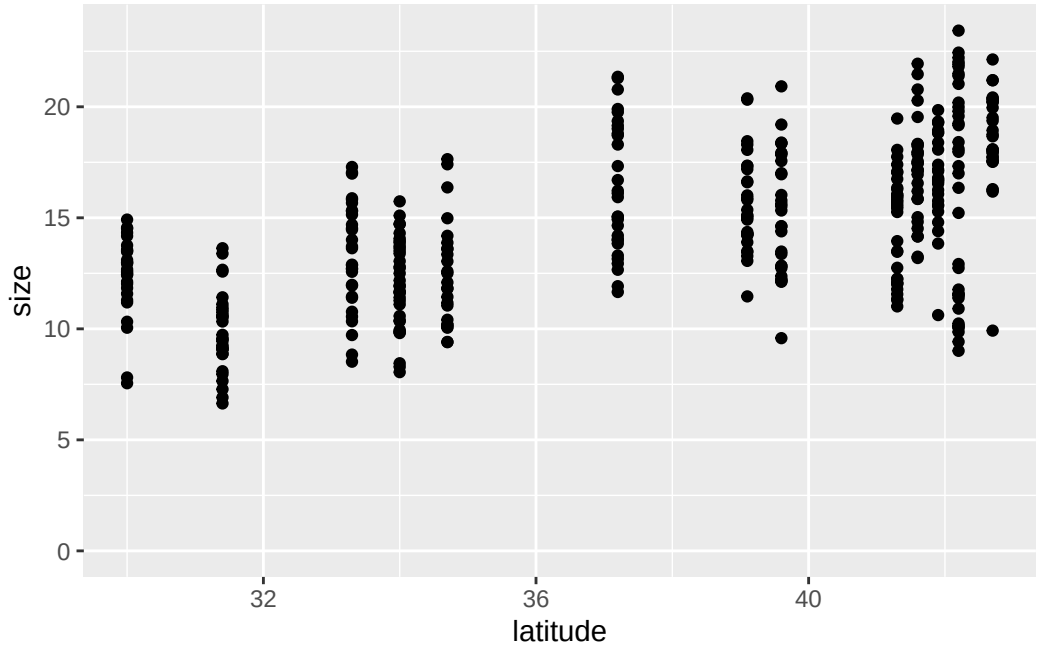
```
library(ggeffects) # for the prediction plot
library(lterdatasampler) # for built-in datasets
```

Fiddler crab data

```
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>
```

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



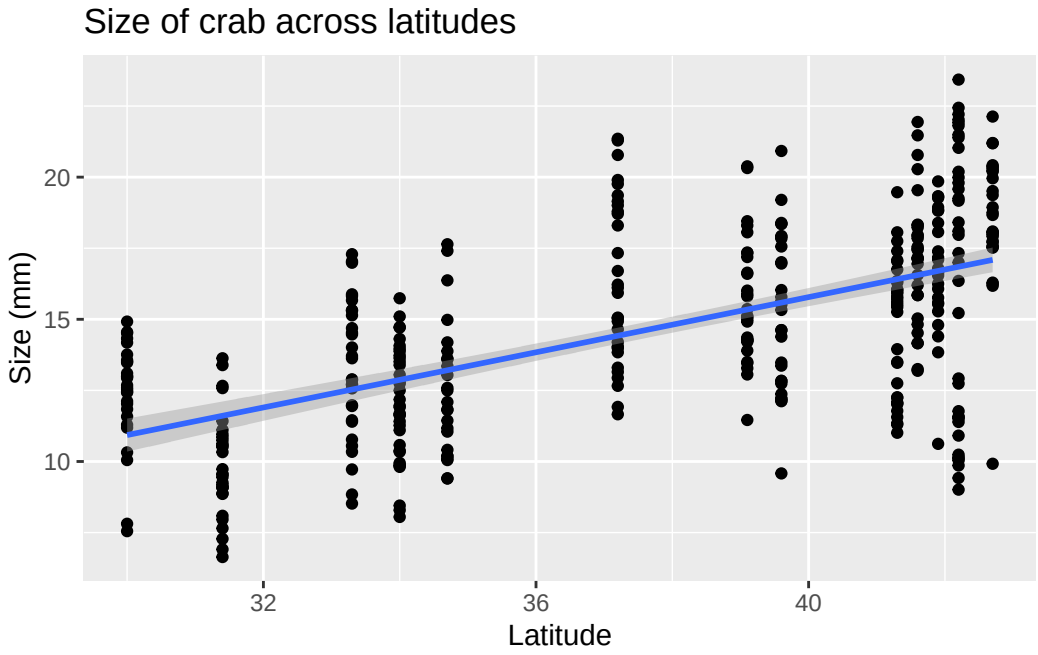
Q1.1 Interpret the graph

As latitude increases, size, on average, also increases. We are farily confident in this assertion because we can see a general increase.

Q1.2 Beautify this graph

```
pie_crab %>%
  ggplot(aes(x = latitude, y = size)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ggtitle("Size of crab across latitudes") +
  labs(x = "Latitude",
       y = "Size (mm)")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
  # Make the y-axis include 0
  ylim(0, NA)
```

```
<ScaleContinuousPosition>
 Range:
 Limits:    0 --    1
```

```
# 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 What does the "iter" argument do?

The iter argument defines how many iterations the Markov chains will run through.

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

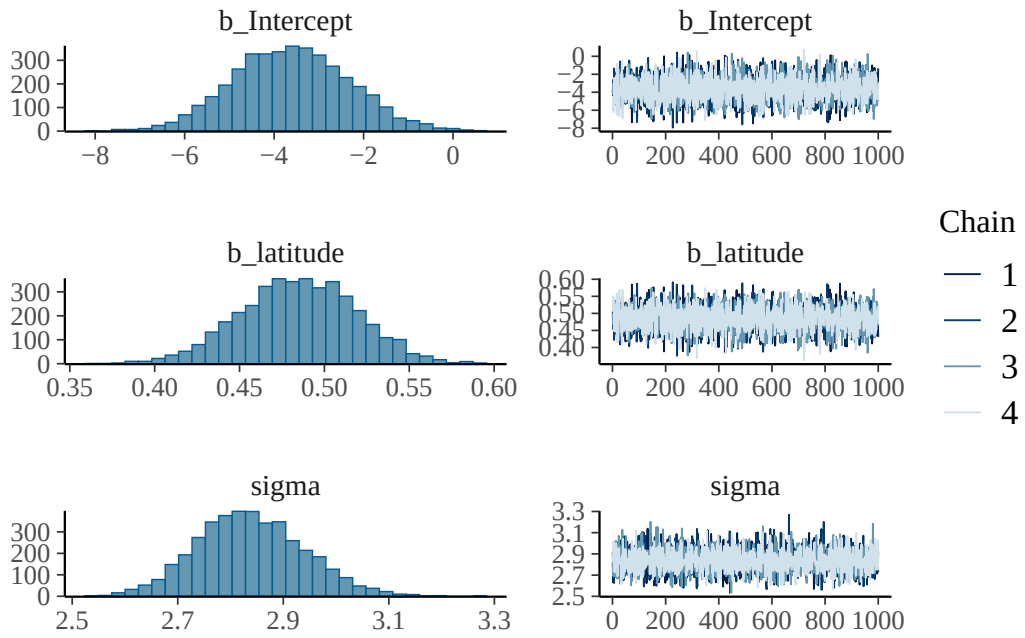
```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: size ~ latitude
   Data: pie_crab (Number of observations: 392)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

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
```




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

```
 Family: gaussian
  Links: mu = identity; sigma = 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).

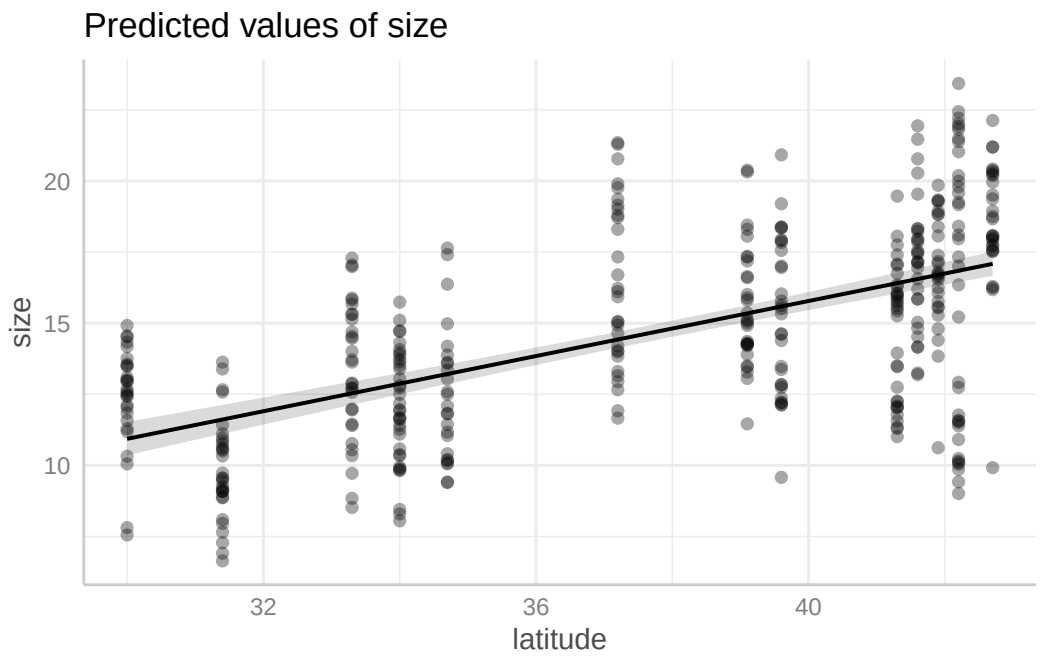
```
as_draws_df(m.crab.lat)  %>%  # extract the posterior samples from the model estimate
  select(b_latitude)  %>%  # pull out the latitude samples from all 4 chains. we'll get a wa
  summarize(p_slope_lessthanorequalto_zero = sum(b_latitude <= 0)/length(b_latitude))
```

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

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

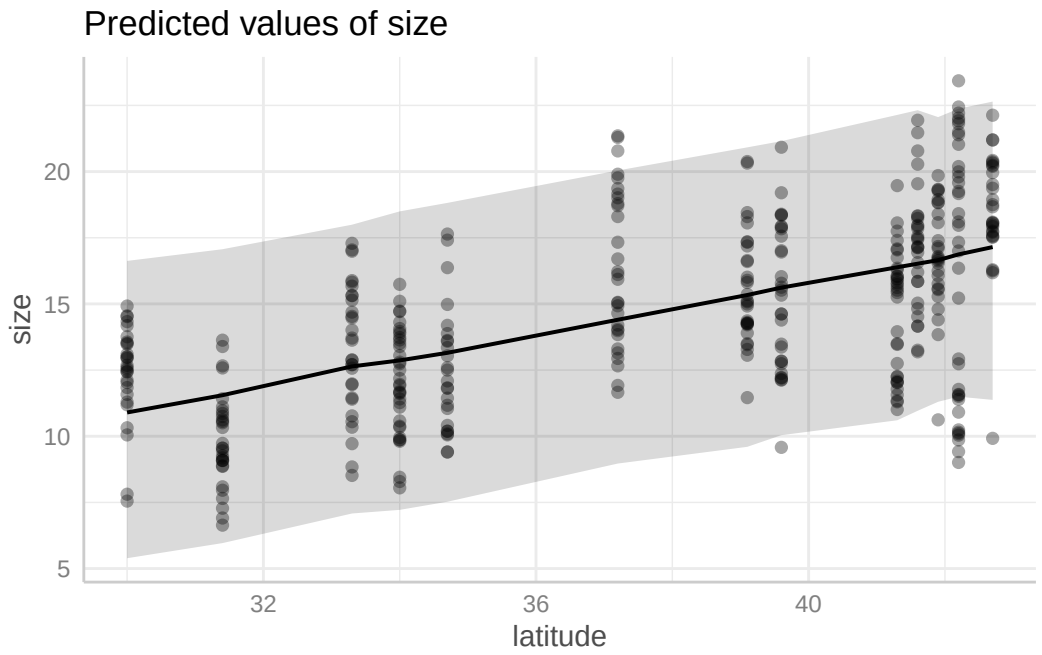
```
# compatibility interval. the shows uncertainty in the average response.
confm.crab.lat <- predict_response(m.crab.lat)
plot(confm.crab.lat, show_data = TRUE)
```

```
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.
```



Q1.4 Make a hypothesis

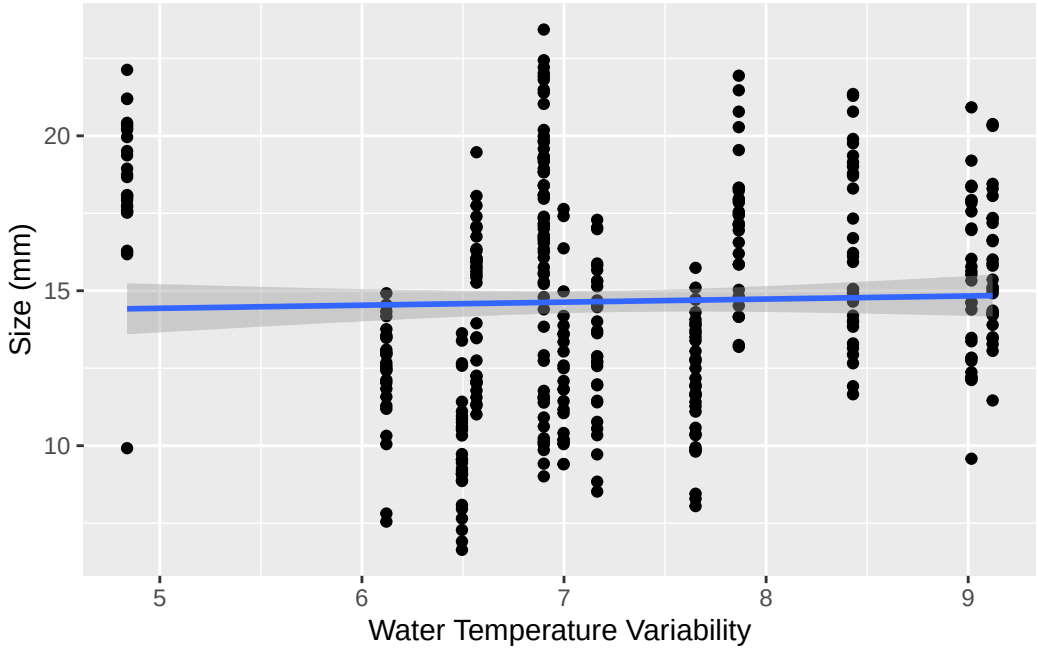
Higher variability would be associated with smaller crabs since we are operating on the assumption that we are going to consistently have larger crabs with a higher latitude, then with more variability, then we will have more smaller crabs.

Q1.5 Graph the data

```
pie_crab %>%
  ggplot(aes(x = water_temp_sd, y = size)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```
  labs(x = "Water Temperature Variability",
       y = "Size (mm)")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
  # Make the y-axis include 0
  ylim(0, NA)
```

```
<ScaleContinuousPosition>
 Range:
 Limits:    0 --    1
```

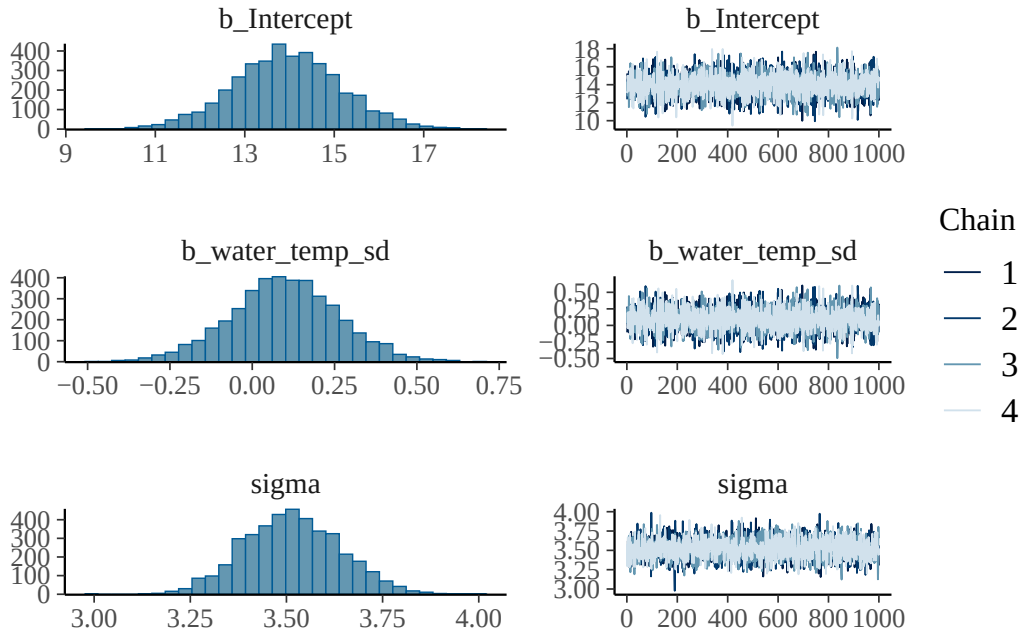
Q1.6 Interpret the graph

Water temperature variability has very little affect on crab size. We are confident in this assertion because there is no relationship.

Q1.7 Set up and run this new model

```
# 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")

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




```
summary(m.crab.watersd)
```

```
 Family: gaussian
```

```
  Links: mu = identity; sigma = 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).
```

Q1.8 Assess the model

The model ran correctly because the posterior samples have a smooth distribution with one clean peak, the chains are overlapping, and the chains are flat.

Q1.9 Interpret the model

For every one standard deviation increase in water temperature, carapace size increases 0.10 millimeters. The confidence interval is -0.23 to 0.42, so it includes 0, and is not reasonably different from 0.

```
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>
```

```r
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 Make a question

Does stress increase as a function of elevation?

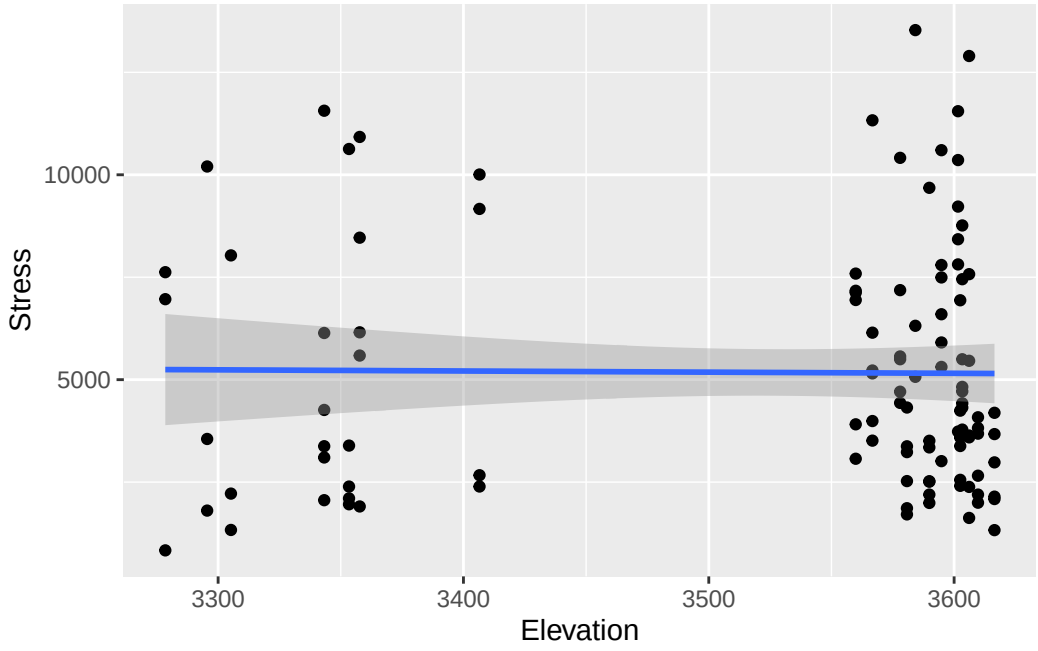
Q2.2 Make a hypothesis

We hypothesize that higher elevation increases stress in pika because there may be less food food availabale, as well as less oxygen.

Q2.3 Graph the data

```r
nwt_pikas_doy %>%
  ggplot(aes(x = elev_m, y = concentration_pg_g)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Elevation",
       y = "Stress")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
  # Make the y-axis include 0
  ylim(0, NA)
```

```
<ScaleContinuousPosition>
 Range:
 Limits:    0 --    1
```

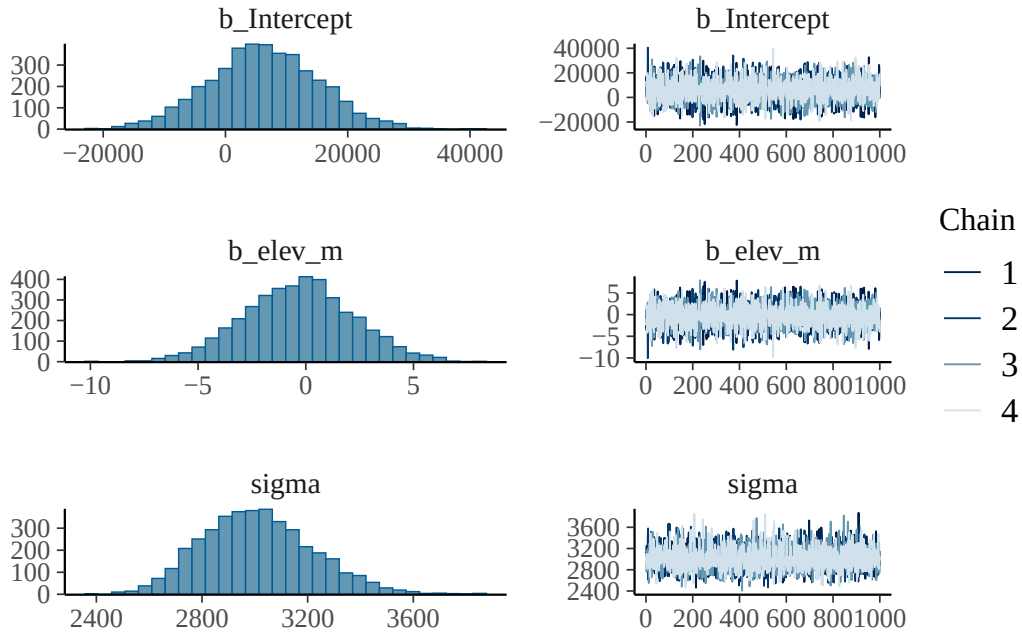
Q2.4 Set up and run a model

```
# pika elevation model
m.pika.elevation <-
  brm(data = nwt_pikas_doy, # Give the model the pika data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      concentration_pg_g ~ elev_m,
      # 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.pika.elevation")
```

Q2.5 Assess the model

```
plot(m.pika.elevation) # show posteriors and chains
```




```
summary(m.pika.elevation)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: concentration_pg_g ~ elev_m
   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  6274.92   8927.98 -11361.19 23720.60 1.00     3832     2607
elev_m       -0.31      2.53     -5.25     4.69 1.00     3830     2635

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma  3011.79    209.67  2641.09  3445.75 1.00     4000     2896

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).
```

The model ran correctly because the posterior samples have a smooth distribution with one clean peak, the chains are overlapping, and the chains are flat.

Q2.6 Interpret the model

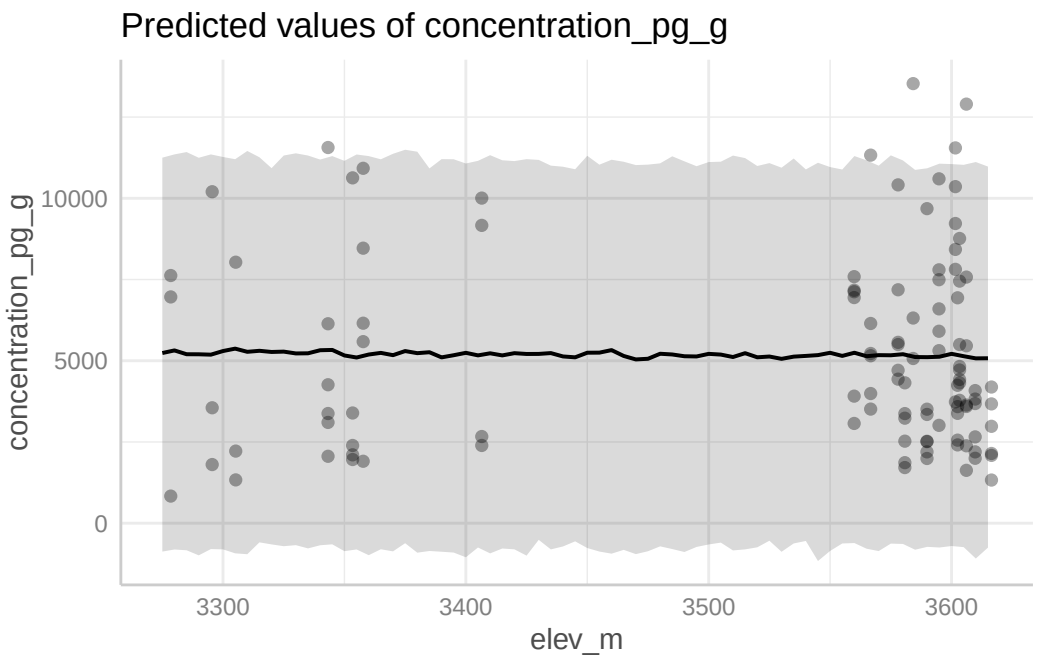
For every one meter increase in elevation, stress level in pikas, measured using glucocorticoid metabolite (GCM) concentration, decreases by 0.31 picogram GCM/gram. The confidence interval is -5.25 to 4.69, so it includes 0, and is not reasonably different from 0.

Q2.7 Plot the model on the data

We are using the prediction interval.

```r
# prediction interval. this shows uncertainty in the data around the average response.
confm.pika.elev <- predict_response(m.pika.elevation, interval = 'prediction')
plot(confm.pika.elev, 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 Write a small results paragraph

We found an decrease of 0.31 picogram GCM/gram of glucocorticoid metabolite (GCM) concentration per 1 meter of elevation, but our 95% credible intervals included zero (-5.25 to 4.69), suggesting that given our model, the effect of elevation on GCM concentration is not different from zero.