https://github.com/jtholt-cntrl/StatisticalReasoning5-handling_complexity.git

Joshua Holt and Luis Rouzaud

1. Interactions

Q1.1) The relationship shown in the comic is not a simple linear regression. Therefor we would need to use a GLM or a multilevel model to describe the different trends.

Q1.2)

Bread dough rises because of yeast - The impact of yeast on the rate of bread rising would increase with time.

Education leads to higher income - Job opportunity can increase due to higher education which leads to higher income. Without a strong education, competitive job opportunities will not be achievable.

Gasoline makes a car go - Engines convert gas to energy which powers the drive train and makes the car go. If the engine is faulty, the gas can't make the car go.

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

```
-- 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   4.0.2     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 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(palmerpenguins) # For the data
library(ggeffects) # for plotting model predictions

# Store the data as penguins
penguins <- palmerpenguins::penguins

# Look at the column names
penguins %>% colnames()
```

```
[1] "species"           "island"            "bill_length_mm"
[4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
[7] "sex"               "year"
```

Q1.3)

```
#Assign new dataframe
penguins.AC <- penguins |>
  #Filter by species
  filter(species == "Adelie"|
         species == "Chinstrap")

penguins.AC
```

```
# A tibble: 220 x 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
 1 Adelie  Torgersen           39.1          18.7               181        3750
 2 Adelie  Torgersen           39.5          17.4               186        3800
 3 Adelie  Torgersen           40.3          18                 195        3250
 4 Adelie  Torgersen           NA            NA                  NA          NA
 5 Adelie  Torgersen           36.7          19.3               193        3450
 6 Adelie  Torgersen           39.3          20.6               190        3650
 7 Adelie  Torgersen           38.9          17.8               181        3625
 8 Adelie  Torgersen           39.2          19.6               195        4675
 9 Adelie  Torgersen           34.1          18.1               193        3475
10 Adelie  Torgersen           42            20.2               190        4250
# i 210 more rows
# i 2 more variables: sex <fct>, year <int>
```

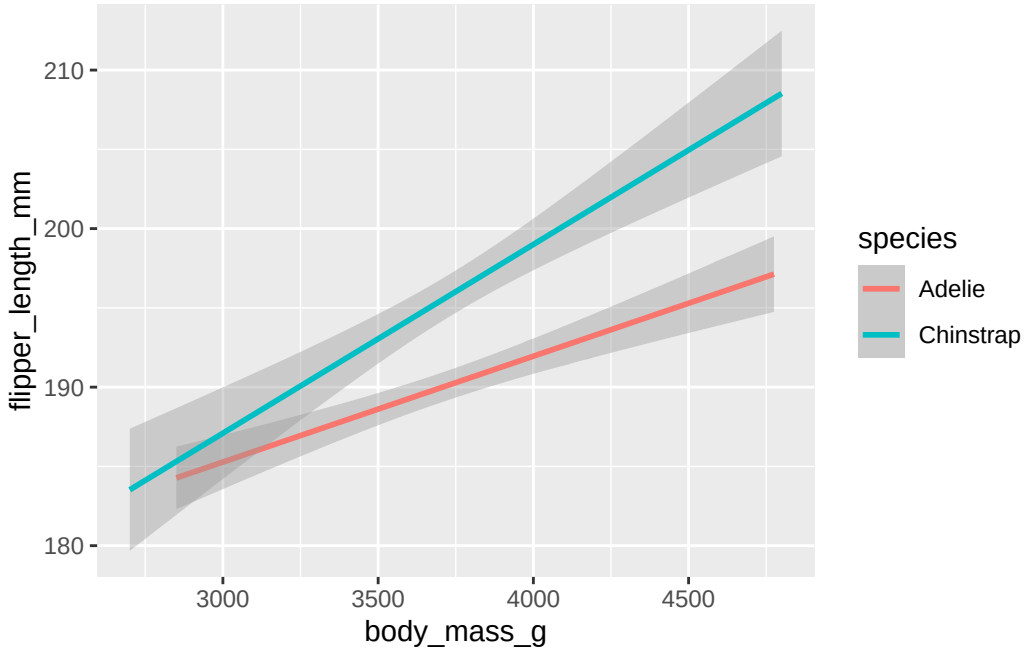
Q1.4)

```
fl.bms <- penguins.AC|>
  ggplot(aes(y = flipper_length_mm, x = body_mass_g, color = species))+
  geom_smooth(method = "lm")


fl.bms
```

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

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```



Q1.5) I think that the effect of body mass on flipper length is conditional on species. The slopes of the best fit lines are distinctly different.

Q1.6) Yes our model ran correctly, all the posterior distributions are normal, the markov chains all overlap, and the Rhat is around 1 for all of them.

```
m.flip.mass.spp.additive <-
  brm(data = penguins.AC, #Give the model the tibble
      family = gaussian, #Tell model what type of distribution
      flipper_length_mm ~ 0 + species + body_mass_g, #Specifying the model
      iter = 2000, warmup = 1000, chains = 4, cores = 4, #Markov chain parameters
      seed = 4, #Determines which random numbers
      file = 'output/m.flip.mass.spp.additive') #saves it as an output

print(m.flip.mass.spp.additive, digits = 4)
```

```
 Family: gaussian
  Links: mu = identity
Formula: flipper_length_mm ~ 0 + species + body_mass_g
   Data: penguins.AC (Number of observations: 219)
```

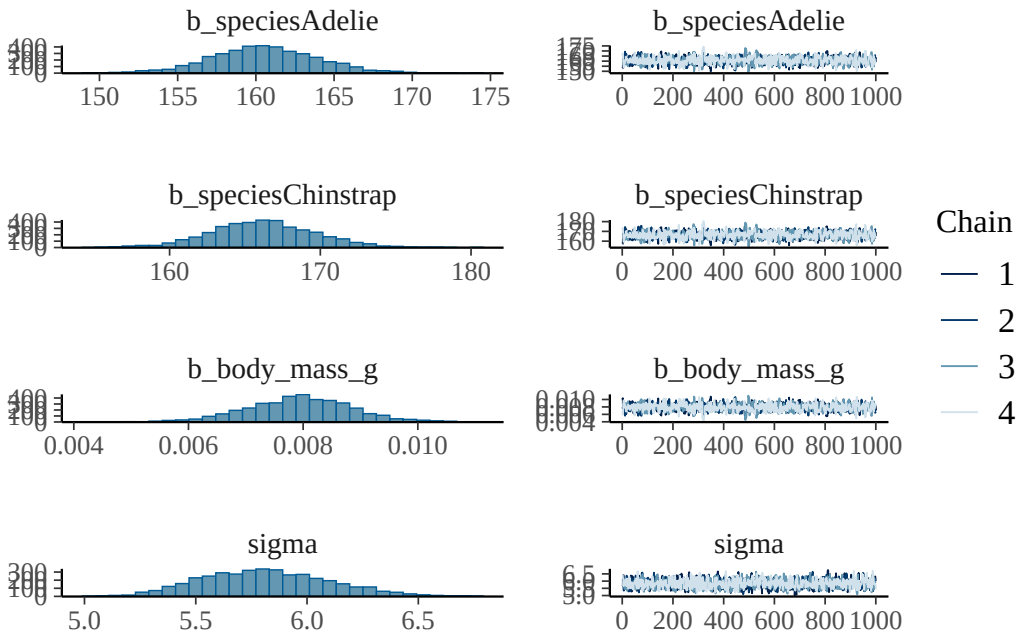
```
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
                 Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
speciesAdelie    160.5778    3.4416 153.7570 167.5652 1.0073     1164     1320
speciesChinstrap 166.2003    3.5091 159.2169 173.1608 1.0087     1191     1326
body_mass_g        0.0079    0.0009   0.0061   0.0098 1.0081     1183     1332

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
sigma   5.8177    0.2857   5.3071   6.4033 1.0028     1538     1633

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).
```

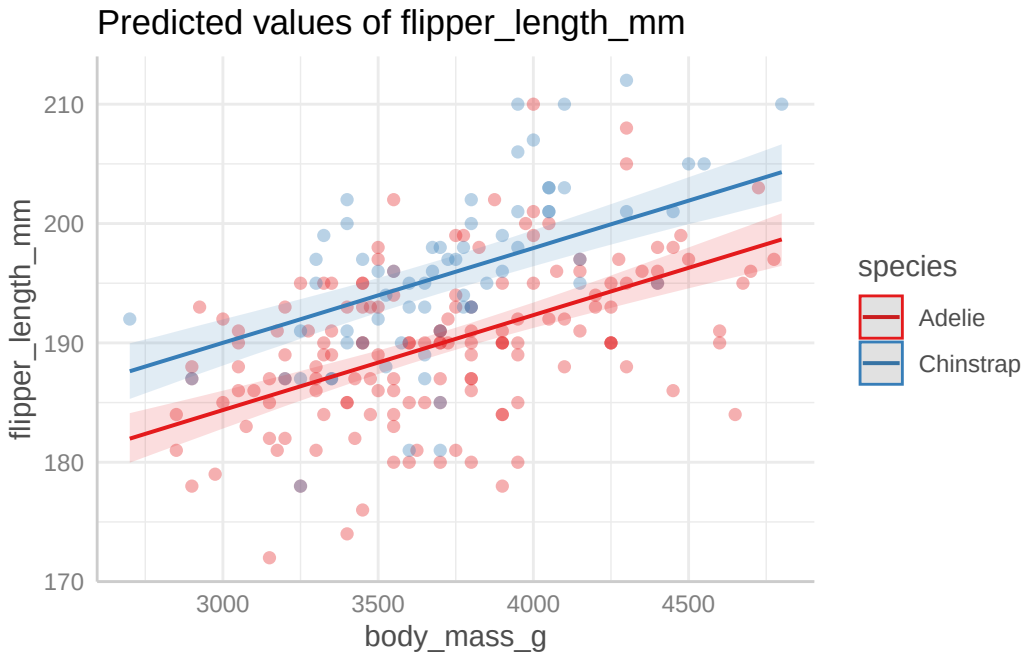
```
plot(m.flip.mass.spp.additive)
```




```
preds.add <- predict_response(m.flip.mass.spp.additive,
                              terms = c("body_mass_g", "species"))

plot(preds.add, 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.
```



Q1.7)

1. Chinstraps have longer flippers by about 6 mm

2. Body mass has a positive effect on flipper length, more specifically as the penguin increase 1 gram of body mass, flipper length is expected to increase by 0.0079 mm. The effect is distinctly different from 0 with the 95% CI being positive (0.0061, 0.0098).

3. No, the effect foes not vary per species in the additive model. We tell this because the slopes are the same.

```
# flipper length by body mass and species - INTERACTIVE model
m.flip.mass.spp.interactive <-
  brm(data = penguins.AC, # Give the model the penguins data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      # First, we write the equation with a as an intercept and b as a slope next to body mas
```

```
      bf(flipper_length_mm ~ 0 + a + b*body_mass_g,
         # Then, we specify that we want our intercept, a, to vary with species
         a ~ 0 + species,
         # Next, we specify that we want our slope, b, to ALSO vary with species
         # (this is the interaction part!!)
         b ~ 0 + species,
         # Lastly, we tell it that we are writing in a particular notation
         nl = TRUE),
      # 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.flip.mass.spp.interactive")

# Print out the model output with 4 digits to avoid the display rounding to zero
print(m.flip.mass.spp.interactive, digits = 4)
```

```
 Family: gaussian
  Links: mu = identity
Formula: flipper_length_mm ~ 0 + a + b * body_mass_g
         a ~ 0 + species
         b ~ 0 + species
   Data: penguins.AC (Number of observations: 219)
  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
a_speciesAdelie    165.1809    3.8183 157.5669 172.5782 1.0001     2447
a_speciesChinstrap 151.1712    6.5569 138.2504 164.1750 1.0019     1698
b_speciesAdelie      0.0067    0.0010   0.0047   0.0087 1.0001     2465
b_speciesChinstrap   0.0120    0.0018   0.0085   0.0155 1.0018     1697
                  Tail_ESS
a_speciesAdelie       2131
a_speciesChinstrap    1906
b_speciesAdelie       2124
b_speciesChinstrap    1916
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
sigma   5.7306    0.2723   5.2303   6.3094 1.0016     2133     2075

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).
```

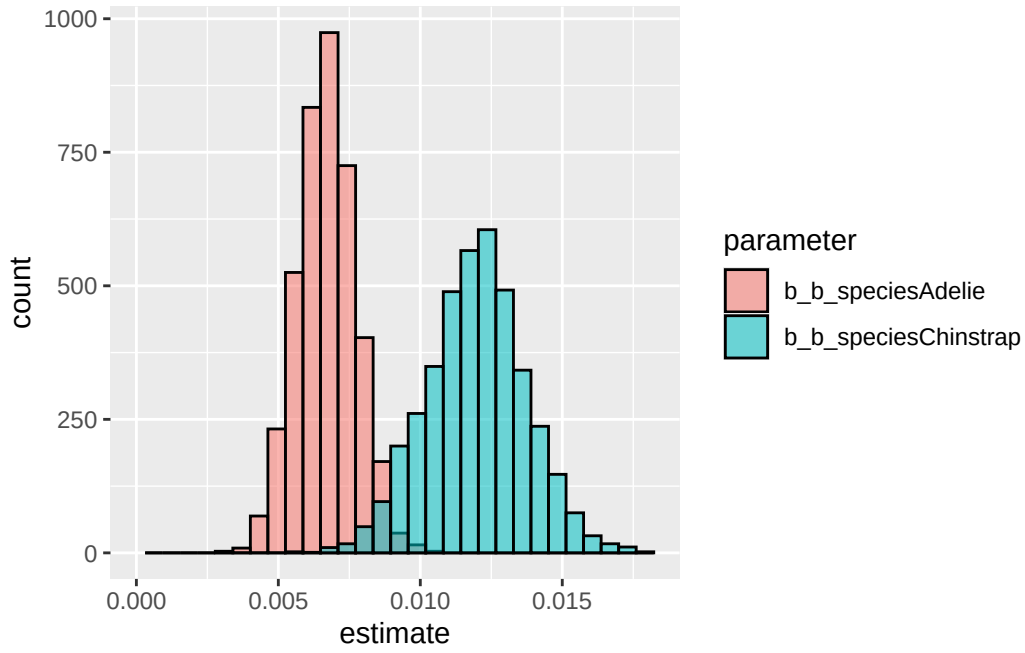
```r
# Store posterior parameter estimates
model_posterior_samples <- as_draws_df(m.flip.mass.spp.interactive)

# Wrangle and then plot samples for the slopes (b) specifically
model_posterior_samples %>%
  # Select the columns we care about (exclude the posterior estimates for the intercepts and
  dplyr::select(b_b_speciesAdelie, b_b_speciesChinstrap) %>%
  # Pivot to long format for easy ggplotting
  pivot_longer(cols = everything(),
               names_to = "parameter",
               values_to = "estimate") %>%
  # Plot as a two overlapping histograms, with the fill colored by parameter
  ggplot(aes(x = estimate, fill = parameter)) +
  geom_histogram(alpha = 0.55, color = "black",
                 # position = "identity" makes the histograms overlay each other instead of s
                 position = "identity") +
  xlim(0, NA)
```

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

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

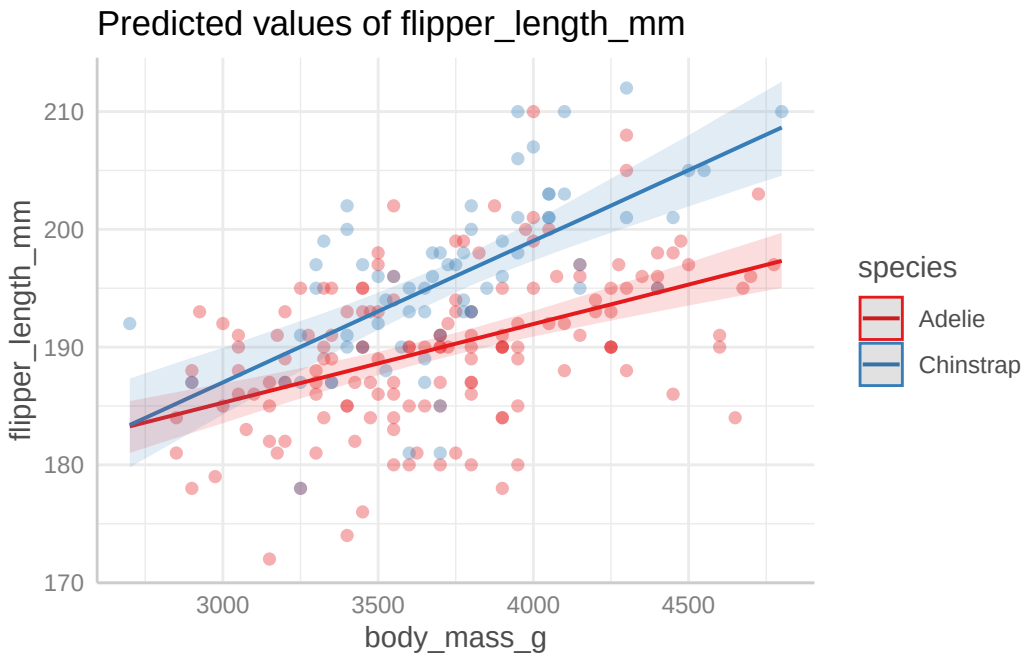
```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_bar()`).
```




```
preds.int <- predict_response(m.flip.mass.spp.interactive,
                              terms = c("body_mass_g", "species"))

plot(preds.int, 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.
```




```
waic(m.flip.mass.spp.interactive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

          Estimate   SE
elpd_waic   -695.3 11.2
p_waic         4.8  0.6
waic        1390.5 22.4
```

```
waic(m.flip.mass.spp.additive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

          Estimate   SE
elpd_waic   -697.6 11.1
p_waic         4.0  0.5
waic        1395.3 22.1
```

```
loo(m.flip.mass.spp.interactive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

         Estimate   SE
elpd_loo   -695.3 11.2
p_loo         4.8  0.6
looic      1390.6 22.4
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 1.0]).

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

```
loo(m.flip.mass.spp.additive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

         Estimate   SE
elpd_loo   -697.6 11.1
p_loo         4.0  0.5
looic      1395.3 22.1
------
MCSE of elpd_loo is 0.0.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.3, 1.0]).

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

Q1.8)The interactive model has better predictive power since both waic (1390.5<1395.3) and looic (1390.6 < 1395.3) values were lower compared to the additive model.

2. Do it yourself

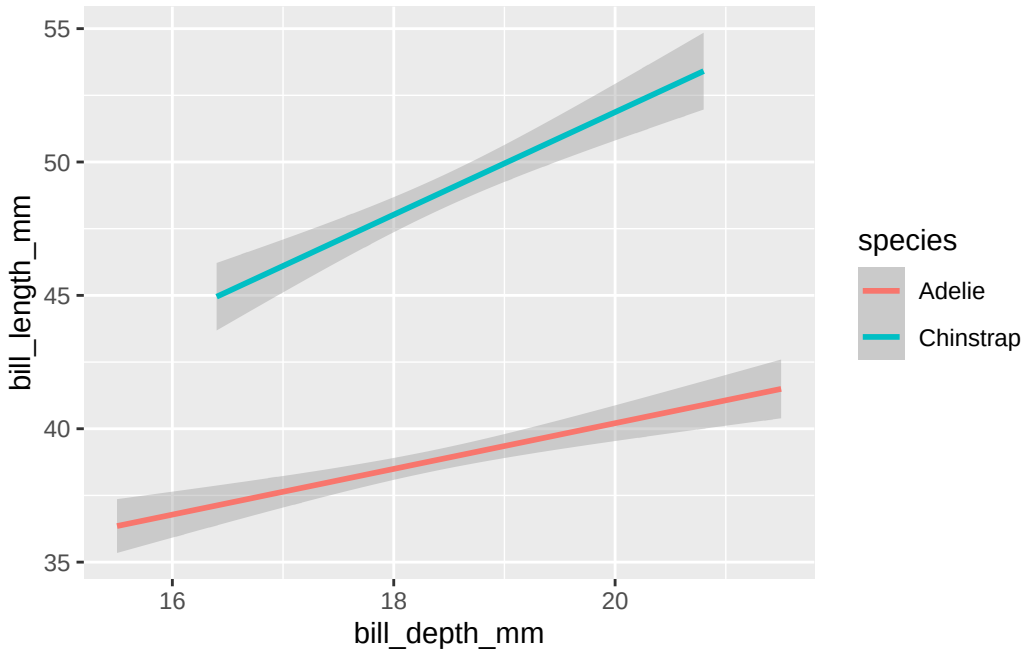
Q2.1)

```
bl.bds <- penguins.AC|>
  ggplot(aes(y = bill_length_mm, x = bill_depth_mm, color = species))+
  geom_smooth(method = "lm")


bl.bds
```

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

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```



Q2.2)

```
# bill length by bill depth and species - INTERACTIVE model
m.bl.bd.spp.interactive <-
  brm(data = penguins.AC, # Give the model the penguins data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      # First, we write the equation with a as an intercept and b as a slope next to body mas
      bf(bill_length_mm ~ 0 + a + b*bill_depth_mm,
         # Then, we specify that we want our intercept, a, to vary with species
         a ~ 0 + species,
         # Next, we specify that we want our slope, b, to ALSO vary with species
         # (this is the interaction part!!)
         b ~ 0 + species,
         # Lastly, we tell it that we are writing in a particular notation
         nl = TRUE),
      # 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.bl.bd.spp.interactive")

# Print out the model output with 4 digits to avoid the display rounding to zero
print(m.bl.bd.spp.interactive, digits = 4)
```

```
 Family: gaussian
  Links: mu = identity
Formula: bill_length_mm ~ 0 + a + b * bill_depth_mm
         a ~ 0 + species
         b ~ 0 + species
   Data: penguins.AC (Number of observations: 219)
  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
```

```
a_speciesAdelie     23.0085    3.1230  16.8320  28.9194 1.0016     1670
a_speciesChinstrap  13.4744    4.9014   4.0860  22.9206 1.0006     1785
b_speciesAdelie      0.8601    0.1697   0.5368   1.1963 1.0017     1675
b_speciesChinstrap   1.9197    0.2659   1.4063   2.4347 1.0006     1790
                   Tail_ESS
a_speciesAdelie        1811
a_speciesChinstrap     1446
b_speciesAdelie        1692
b_speciesChinstrap     1465

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
sigma   2.4958    0.1181   2.2762   2.7433 1.0021     2318     2241

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).
```

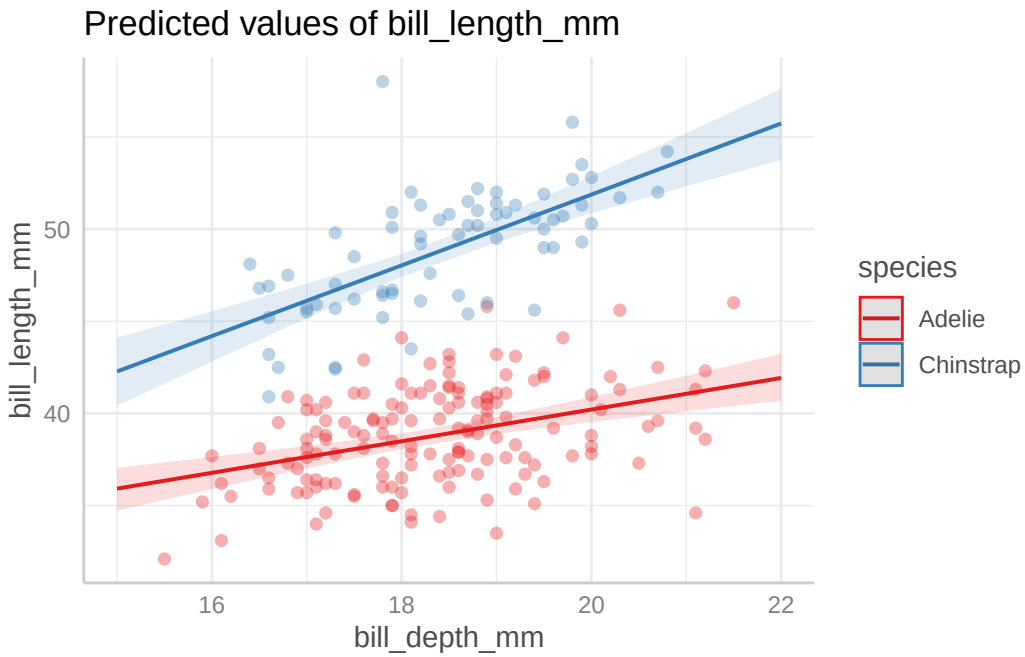
Q2.3) The Rhat for all predictors is close to 1, the posterior distributions are normal, and the chains all overlap, so we determined that our model ran correctly.

Q2.4)

```r
preds.int2 <- predict_response(m.bl.bd.spp.interactive,
                          terms = c("bill_depth_mm", "species"))

plot(preds.int2, 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.
```



Q2.5)

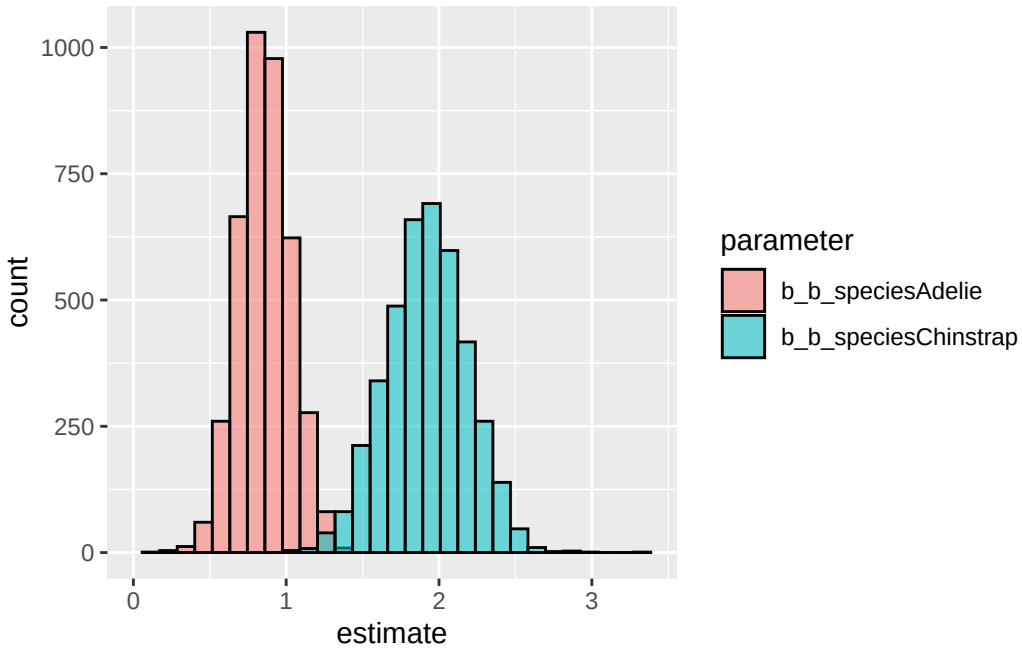
```
model_posterior_samples2 <- as_draws_df(m.bl.bd.spp.interactive)

# Wrangle and then plot samples for the slopes (b) specifically
model_posterior_samples2 %>%
  # Select the columns we care about (exclude the posterior estimates for the intercepts and
  dplyr::select(b_b_speciesAdelie, b_b_speciesChinstrap) %>%
  # Pivot to long format for easy ggplotting
  pivot_longer(cols = everything(),
               names_to = "parameter",
               values_to = "estimate") %>%
  # Plot as a two overlapping histograms, with the fill colored by parameter
  ggplot(aes(x = estimate, fill = parameter)) +
  geom_histogram(alpha = 0.55, color = "black",
                 # position = "identity" makes the histograms overlay each other instead of s
                 position = "identity") +
  xlim(0, NA)
```

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

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

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



Q2.6)

1. Bill Depth (mm) positively influences Bill Length for both species. For Adelie, an increase in 1 mm bill depth would lead to a 0.8601 mm increase in bill length and for Chinstraps, an increase of 1 mm in bill depth would lead to an increase of 1.9197 mm in bill length.

2. The slope estimates are all different from zero since teh 95% CI for all of them are positive and don't include 0. a_speciesAdelie (16.8320, 28.9194), a_speciesChinstrap (4.9014, 22.9206), b_speciesAdelie(0.5368, 1.1963), b_speciesChinstrap(1.4063, 2.4347).

3. There is some overlap between the two predictor histograms, but overall they are distinct in their shape and placement on the x-axis. The lines of the predicted values also don't overlap, showing that the values are different.

Q2.7) From interactive model waic = 1026.7, looic = 1026.7. From additive model waic = 1036.3, looic = 1036.3. The interactive model has better predictor power, having a lower value for both waic and psis.

```
m.bl.bd.spp.additive <-
  brm(data = penguins.AC, #Give the model the tibble
      family = gaussian, #Tell model what type of distribution
      bill_length_mm ~ 0 + species + bill_depth_mm, #Specifying the model
      iter = 2000, warmup = 1000, chains = 4, cores = 4, #Markov chain parameters
      seed = 4, #Determines which random numbers
      file = 'output/m.bl.bd.spp.additive') #saves it as an output

print(m.bl.bd.spp.additive, digits = 4)
```

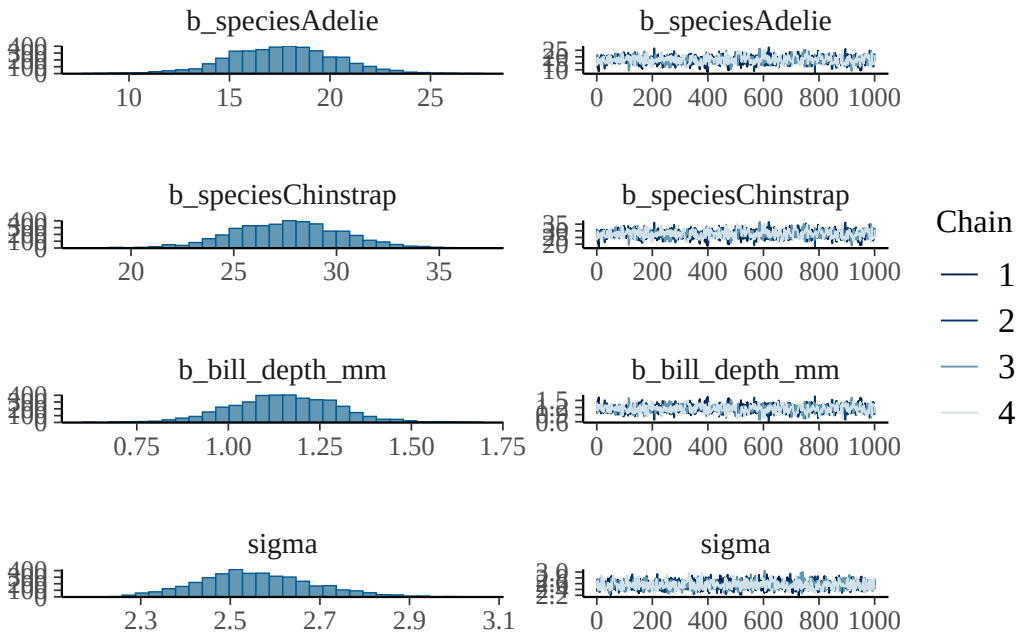
```
 Family: gaussian
  Links: mu = identity
Formula: bill_length_mm ~ 0 + species + bill_depth_mm
   Data: penguins.AC (Number of observations: 219)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
                 Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
speciesAdelie     17.6613    2.6567  12.3249  22.8738 1.0055      750      944
speciesChinstrap  27.6128    2.6798  22.2198  32.9197 1.0048      754      948
bill_depth_mm      1.1519    0.1446   0.8685   1.4442 1.0054      739      934
```

```
Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI   Rhat Bulk_ESS Tail_ESS
sigma   2.5580    0.1277   2.3206   2.8236 1.0031     1229     1479

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.bl.bd.spp.additive)
```



s

```
waic(m.bl.bd.spp.interactive)
```

```
Warning:
1 (0.5%) p_waic estimates greater than 0.4. We recommend trying loo instead.


Computed from 4000 by 219 log-likelihood matrix.

          Estimate   SE
```

```
elpd_waic   -513.4 12.4
p_waic         5.2  1.0
waic        1026.7 24.9

1 (0.5%) p_waic estimates greater than 0.4. We recommend trying loo instead.
```

```
waic(m.bl.bd.spp.additive)
```

```
Warning:
1 (0.5%) p_waic estimates greater than 0.4. We recommend trying loo instead.


Computed from 4000 by 219 log-likelihood matrix.

          Estimate   SE
elpd_waic   -518.2 11.9
p_waic         4.4  0.9
waic        1036.3 23.7

1 (0.5%) p_waic estimates greater than 0.4. We recommend trying loo instead.
```

```
loo(m.bl.bd.spp.interactive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

         Estimate   SE
elpd_loo   -513.4 12.5
p_loo         5.2  1.0
looic      1026.7 24.9
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 1.1]).

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

```
loo(m.bl.bd.spp.additive)
```

```
Computed from 4000 by 219 log-likelihood matrix.

         Estimate   SE
elpd_loo   -518.2 11.9
p_loo         4.4  0.9
looic      1036.3 23.7
------
MCSE of elpd_loo is 0.1.
MCSE and ESS estimates assume MCMC draws (r_eff in [0.2, 1.1]).

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