Stats reasoning 2

Sam and Paige

Load Libraries

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

```
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(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(ggeffects) # for  prediction plots
```

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

```
library(palmerpenguins) # data we'll be using
```

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

1.1 Refresh on coefficients

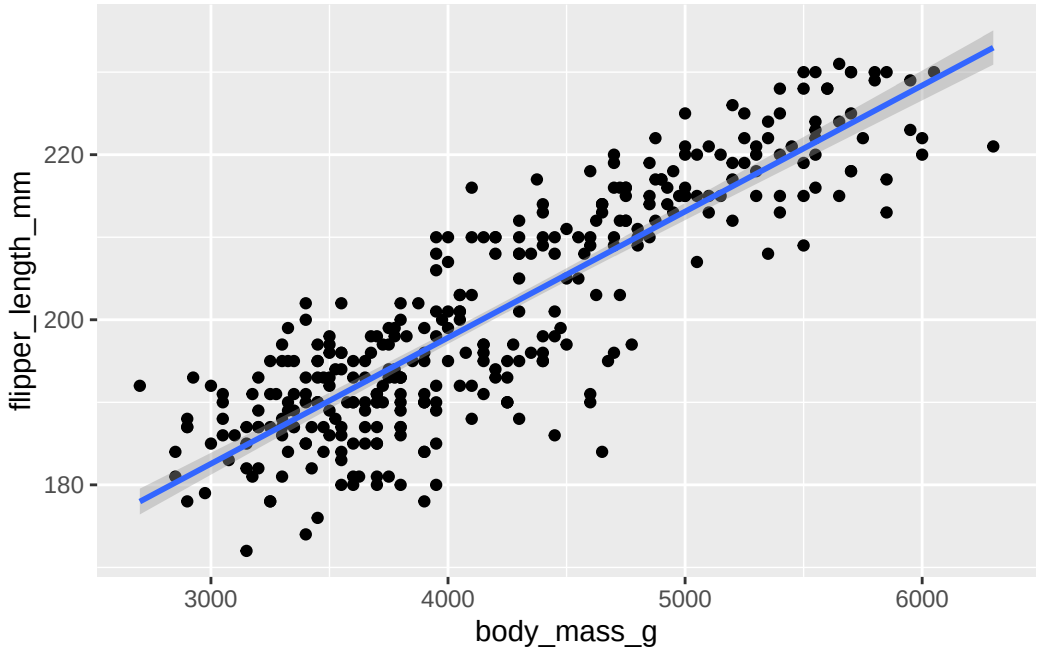
```
penguins <- palmerpenguins::penguins
```

```
penguins %>%
  ggplot(aes(x = body_mass_g,
             y = flipper_length_mm)) +
  geom_point() +
  # Add a geom_smooth with an "lm" line for visualization
  geom_smooth(method = "lm")
```

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

Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).

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




```
m.flip.mass <-
  brm(data = penguins, # Give the model the penguins data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      flipper_length_mm ~ 1 + body_mass_g,
      # 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")
```

Q 1.1

```
?palmerpenguins
```

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

The magnitude of the relationship is 0.02. Therefore, the flipper length increases by 0.02 for every 1 gram of body mass.

Q1.1b

We estimate that the slope is ~0.015 mm/g. This calculated slope is fairly close to the modeled slope.

The probability the slope was greater than 0 is 1.

Q1.2

```
summary(m.flip.mass)
```

```
 Family: gaussian
  Links: mu = identity
Formula: flipper_length_mm ~ 1 + body_mass_g
   Data: penguins (Number of observations: 342)
  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     136.73      2.02   132.78   140.67 1.00     4555     3039
body_mass_g     0.02      0.00     0.01     0.02 1.00     4719     3135

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     6.94      0.26     6.43     7.46 1.00     1740     1500

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).
```

Yes, we can conclude that the slope has a high probability of being different from 0. The 95% CI's exclude 0 (0.01-0.02).

Q1.3

```
as_draws_df(m.flip.mass) |> # extract the posterior samples from the model estimate
  select(b_body_mass_g) |> # pull out the latitude samples from all 4 chains. we'll get a wa
  summarize(p_slope_greaterthan_zero = sum(b_body_mass_g > 0)/length(b_body_mass_g))
```

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

```
# A tibble: 1 x 1
  p_slope_greaterthan_zero
                     <dbl>
1                        1
```

The probability that the slope is greater than 0 is 1.

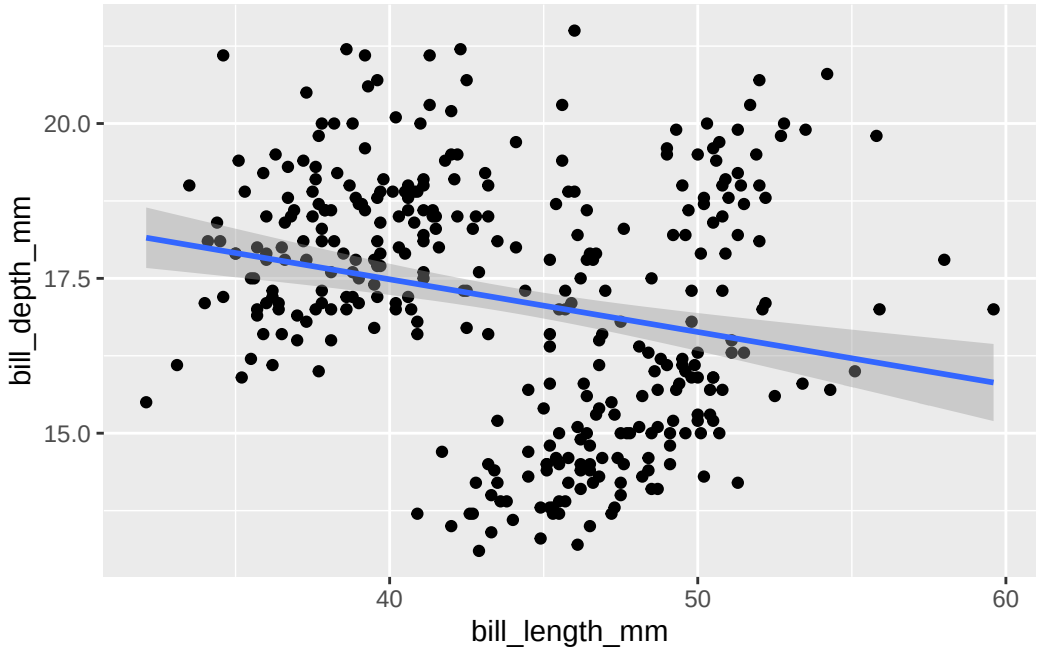
Q1.4

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
             y = bill_depth_mm)) +
  geom_point() +
  # Let's add in a basic lm just to visualize
  geom_smooth(method = "lm")
```

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

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

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



Q1.4

Bill length has a negative effect on bill depth.

Q1.5 Write a couple "results section" sentences with your conclusions about whether bill length is associated with bill depth given this model.

There is a negative relationship between bill length and depth. For every 1 mm increase in bill length, there is a 0.09 mm decrease in bill depth (95% CI -0.12- -0.05).

Q1.6

```
# flipper length by body mass model
m.depth.length.species <-
  brm(data = penguins, # Give the model the penguins data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      bill_depth_mm ~ 0 + bill_length_mm + species,
      # 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.depth.length.species")
```

```
summary(m.depth.length.species)
```

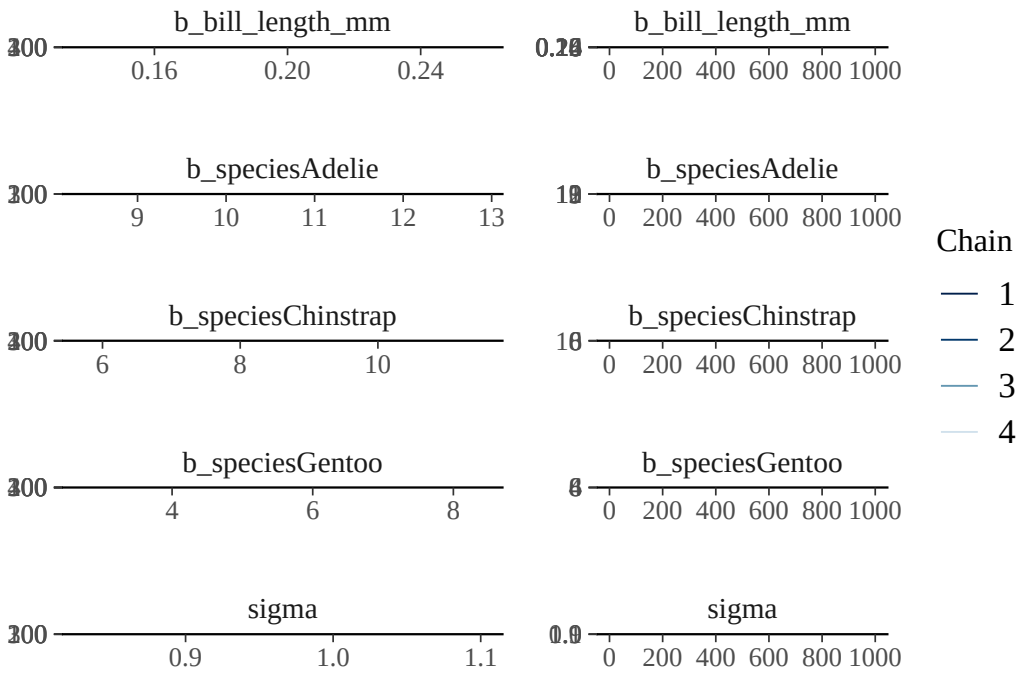
```
 Family: gaussian
  Links: mu = identity
Formula: bill_depth_mm ~ 0 + bill_length_mm + species
   Data: penguins (Number of observations: 342)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
bill_length_mm      0.20      0.02     0.17     0.23 1.00      709      896
speciesAdelie      10.60      0.68     9.23    11.90 1.00      712      817
speciesChinstrap    8.67      0.85     6.93    10.31 1.00      719      942
speciesGentoo       5.50      0.83     3.84     7.07 1.00      713      927

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.96      0.04     0.88     1.03 1.00     1288     1529

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.depth.length.species)
```

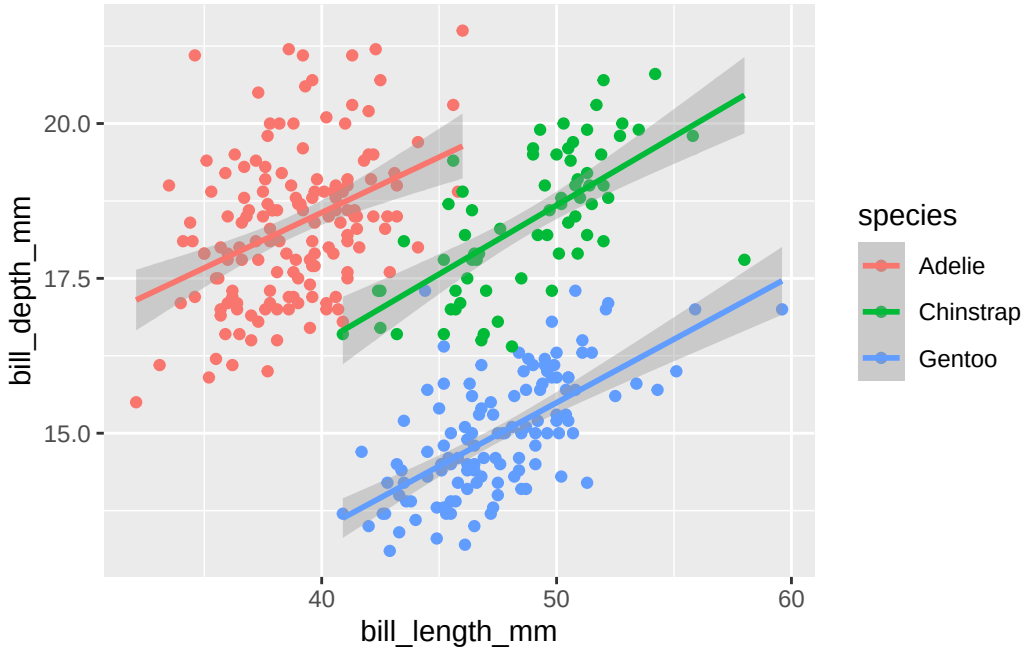



```
penguins %>%
  ggplot(aes(x = bill_length_mm,
             y = bill_depth_mm,
             color = species)) +
  geom_point() +
  geom_smooth(method = "lm")
```

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

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

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



Q1.7 Visually measure differences in bill depth between species

Visually measure differences in bill depth between species; choose a region of the x-axis which has data for all three species (e.g. around 45mm). Estimate the difference in bill depth (the y-axis) between a) Adelie and Chinstrap and b) Adelie and Gentoo. Are those differences consistent with the differences that we calculated from the model?

Adelie to Chinstrap: 19 - 17.5 = 1.5

Adelie to Gentoo : 19 - 14 = 5

These values are very close to the differences calculated by the model.

Q 1.8

The results from the univariate regression of `bill depth ~ bill length` *indicate that the effect of bill length on bill depth was a change of* ***0.02 mm*** *of bill depth for every 1mm of bill length. When we add in* `species`*, the results from the multivariate regression of* `bill depth ~ bill length + species` *indicate that the effect of bill length on bill depth was a change of* ***0.2*** *mm of bill depth for every 1mm of bill length.*

By adding in the effect of species on bill depth, the relationship switched from negative to positive. This reveals that the assumption that bill depth decreases with bill length, as shown in the univariate model, was false, and it was necessary to account for the effect of length.

Part 2

Load in the data

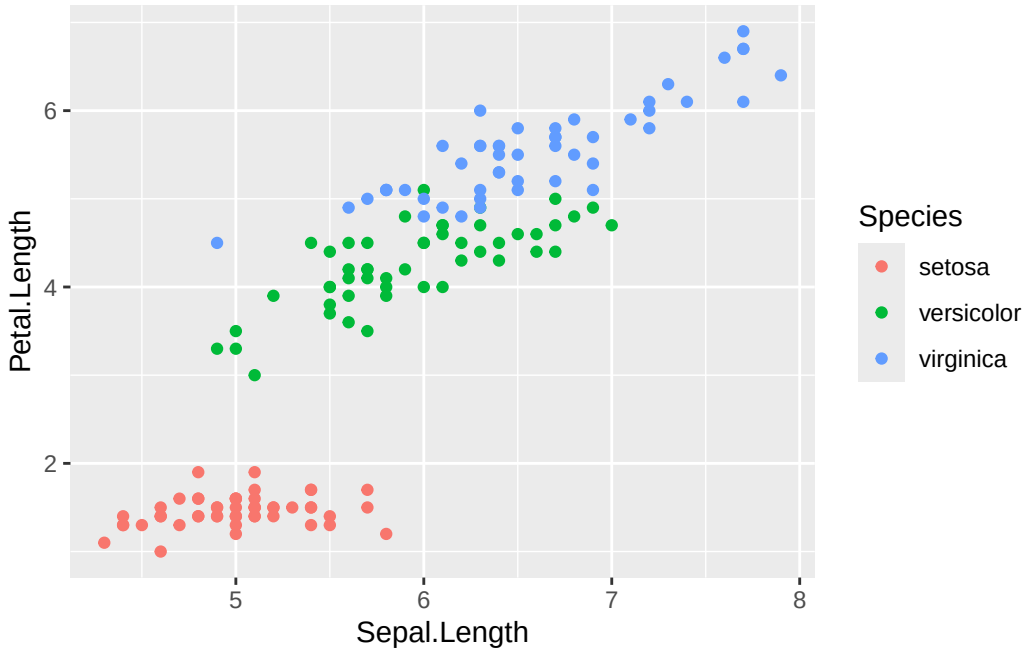
```
iris <- datasets::iris
```

Q2.1

We hypothesize that as flower petal length has a positive correlation with sepal length, but that this effect varies by species.

Q2.2

```
ggplot(iris,aes(x=Sepal.Length, y=Petal.Length,color=Species))+
  geom_point()
```



Q2.3

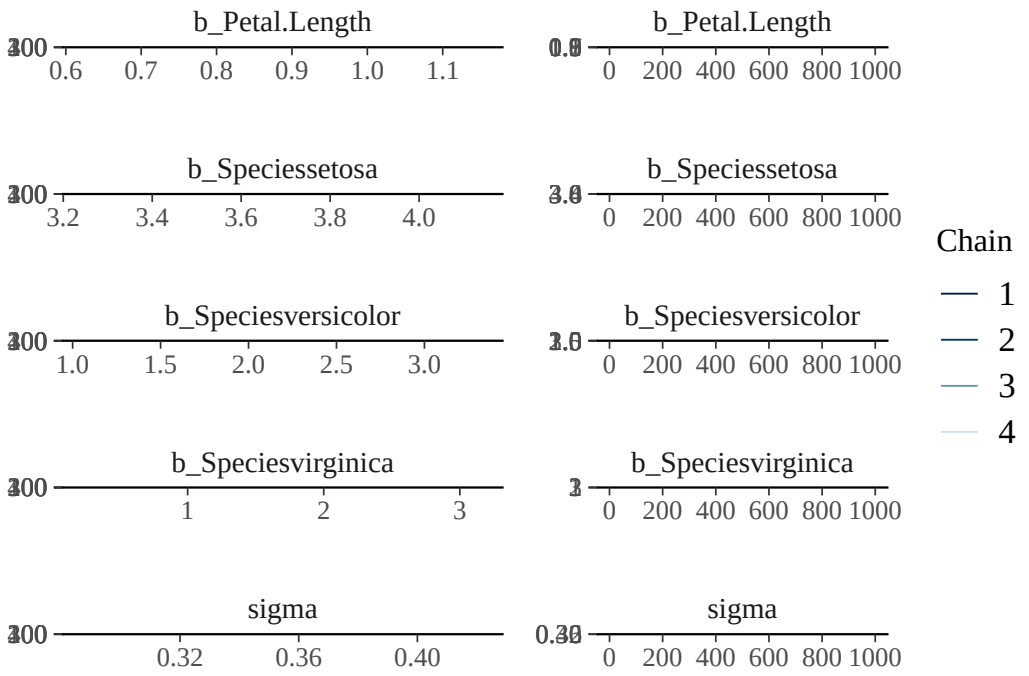
```
iris_sepal_petal <-
  brm(data = iris, # Give the model the penguins data
      # Choose a gaussian (normal) distribution
      family = gaussian,
      # Specify the model here.
      Sepal.Length ~ 0 + Petal.Length + Species,
      # Here's where you specify parameters for executing the Markov 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/iris_sepal_petal")
```

Q2.4

The rhat values for the different species are all equal to 1, which means we can be confident in the model.

Additionally, the plots have a single, clear peak and a good variance distribution.

```
plot(iris_sepal_petal)
```




```
summary(iris_sepal_petal)
```

```
 Family: gaussian
  Links: mu = identity
Formula: Sepal.Length ~ 0 + Petal.Length + Species
   Data: iris (Number of observations: 150)
  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
Petal.Length         0.90      0.07     0.78     1.04 1.00      885      823
Speciessetosa        3.69      0.11     3.47     3.90 1.00      921      926
Speciesversicolor    2.09      0.29     1.51     2.65 1.00      901      788
Speciesvirginica     1.57      0.37     0.80     2.30 1.00      881      813

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.34      0.02     0.31     0.39 1.00     1274     1618

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).
```

Q2.5

The effect of sepal length on petal length is 0.9: for every 1 cm increase in sepal length, there is a 0.9 cm increase in petal width. This is reasonable different from 0, because the CI's exclude 0.

Setosa has the longest petal length, then versicolor, then virginica has the shortest.

Q2.6

With our analysis, we can determine that for every 1 cm increase in sepal length there is a 0.9 cm increase in petal length (95% CI 0.79-1.04). This effect is consistent across species of iris flower, though with variance in petal length across species.