# Intro to Multiple Regression

Nate Wells

Math 141, 3/3/21

# Outline

In this lecture, we will. . .

# Outline

In this lecture, we will...

- Investigate linear regression models with 1 quantitative and 1 categorical explanatory variable
- Compare parallel slopes and interaction models for multiple regression

Section 1

# Introduction to Multiple Linear Regression

# Penguins!

Figure 1: Three Gentoo penguins. Photo: Barend (Barry) Becker

The `palmerpenguins` package contains size information for three species of penguins on islands near the Palmer Archipelago.

# A glimpse

```
library(palmerpenguins)
penguins <- penguins %>% drop_na()
glimpse(penguins)
```

```
## Rows: 333
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 3...
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 17.6, 2...
## $ flipper_length_mm <int> 181, 186, 195, 193, 190, 181, 195, 182, 191, 198,...
## $ body_mass_g       <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3...
## $ sex               <fct> male, female, female, female, male, female, male,...
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

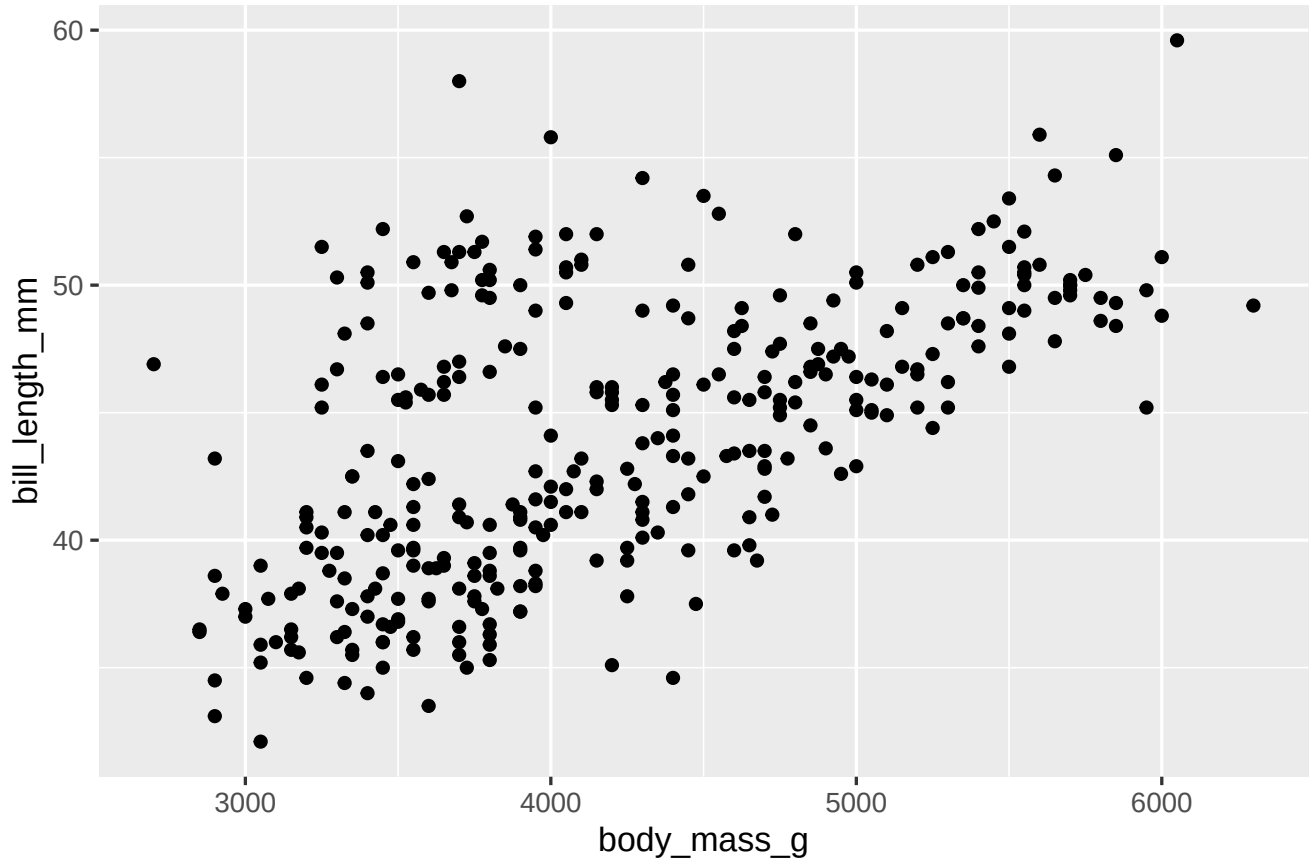
# A (simple) model

How well can we predict `bill_length_mm` based on `body_mass_g`?

# A (simple) model

How well can we predict `bill_length_mm` based on `body_mass_g`?

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm))+
  geom_point()
```



# A (simple) model

How well can we predict `bill_length_mm` based on `body_mass_g`?

# A (simple) model

How well can we predict `bill_length_mm` based on `body_mass_g`?

```
slr_penguins <-lm(bill_length_mm ~ body_mass_g, data = penguins)
get_regression_table(slr_penguins)
```

```
## # A tibble: 2 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     27.2        1.29      21.0       0   24.6     29.7
## 2 body_mass_g    0.004      0         13.3       0    0.003    0.005
```

```
get_correlation(penguins, bill_length_mm ~ body_mass_g)
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.589
```

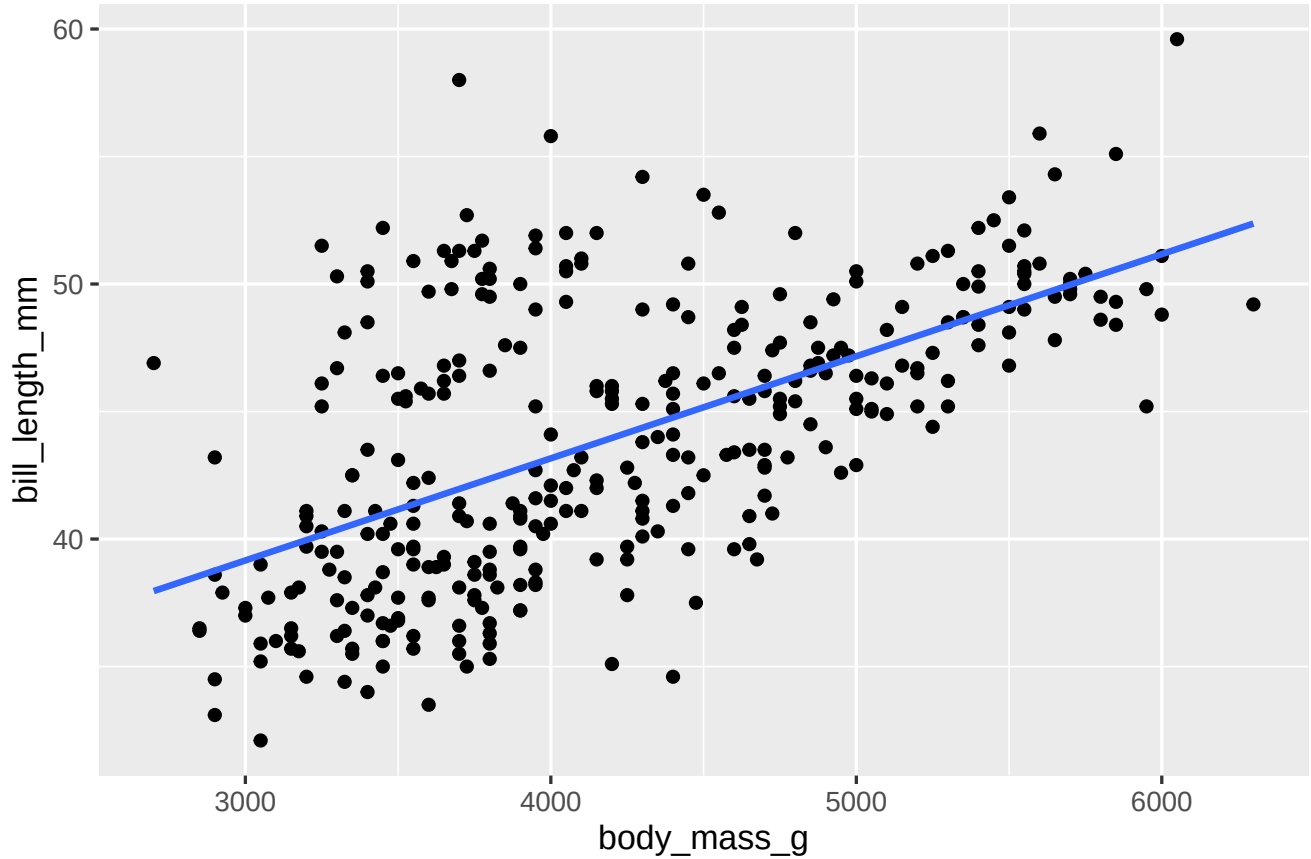
# A (simple) model

What are some explanations for the strength of linear trend? (Moderate, positive)

# A (simple) model

What are some explanations for the strength of linear trend? (Moderate, positive)

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm))+
  geom_point()+
  geom_smooth(method="lm", se = F)
```



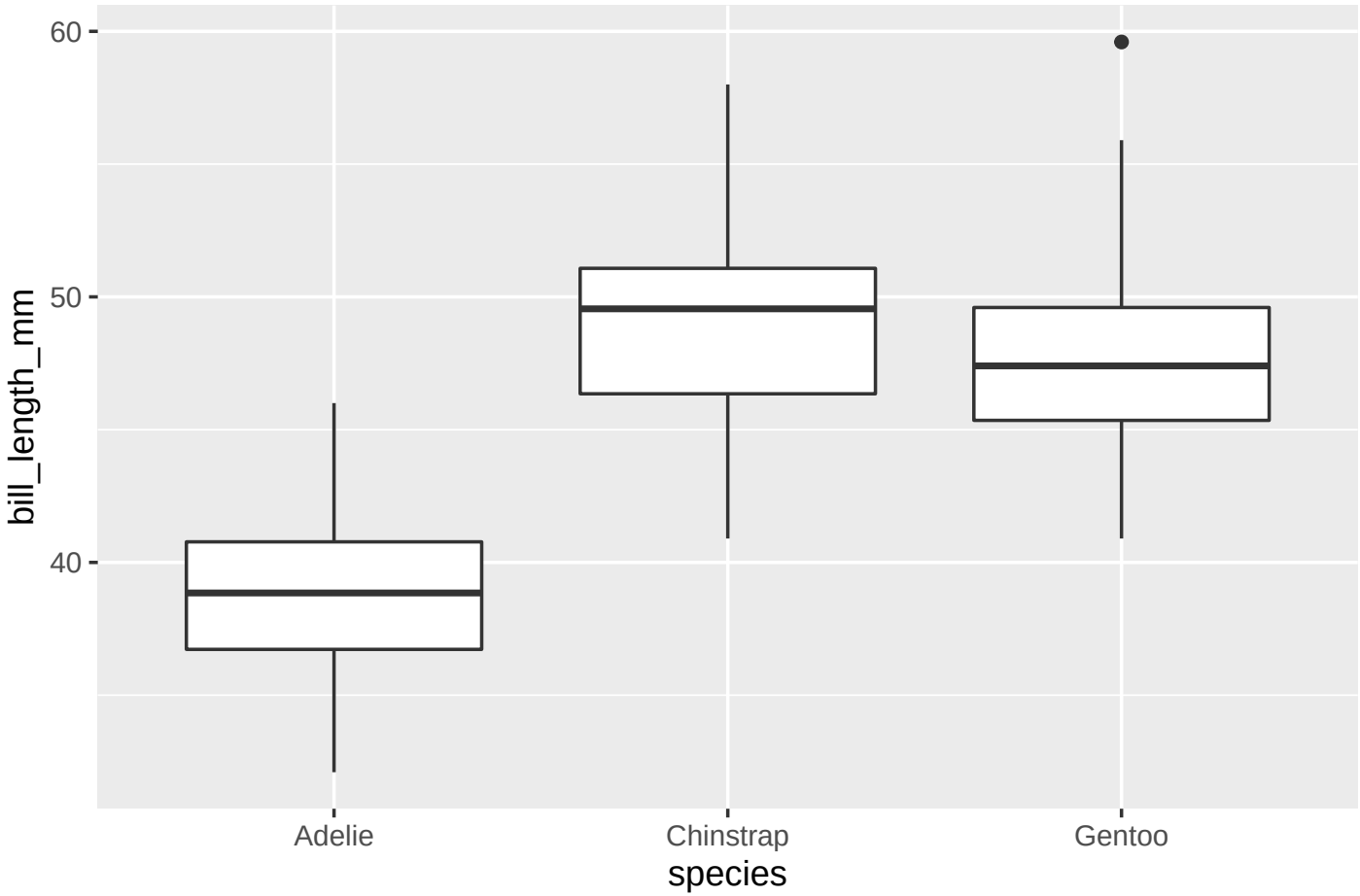
# Regression with 1 Categorical Variable

Does `bill_length_mm` depend on species?

# Regression with 1 Categorical Variable

Does `bill_length_mm` depend on species?

```
ggplot(penguins, aes(x = species, y = bill_length_mm))+geom_boxplot()
```



# Regression with 1 Categorical Variable

Does `bill_length_mm` depend on species?

# Regression with 1 Categorical Variable

Does `bill_length_mm` depend on species?

```
cat_penguins <- lm(bill_length_mm ~ species, data = penguins)
get_regression_table(cat_penguins)
```

```
## # A tibble: 3 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             38.8     0.246     158.        0    38.3     39.3
## 2 speciesChinstrap      10.0     0.436      22.9       0     9.15    10.9
## 3 speciesGentoo          8.74    0.367      23.8       0     8.02     9.47
```

# Regression with 1 Categorical Variable

Does `bill_length_mm` depend on species?

```
cat_penguins <- lm(bill_length_mm ~ species, data = penguins)
get_regression_table(cat_penguins)
```

```
## # A tibble: 3 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           38.8      0.246     158.        0    38.3     39.3
## 2 speciesChinstrap    10.0      0.436      22.9       0     9.15    10.9
## 3 speciesGentoo        8.74     0.367      23.8       0     8.02     9.47
```

Regression Equation for `bill_length_mm` $Y$ as a function of species $s$

$$\hat{Y} = 38.824 + 10.01 I_{\text{Chinstrap}}(s) + 8.744 I_{\text{Gentoo}}(s)$$

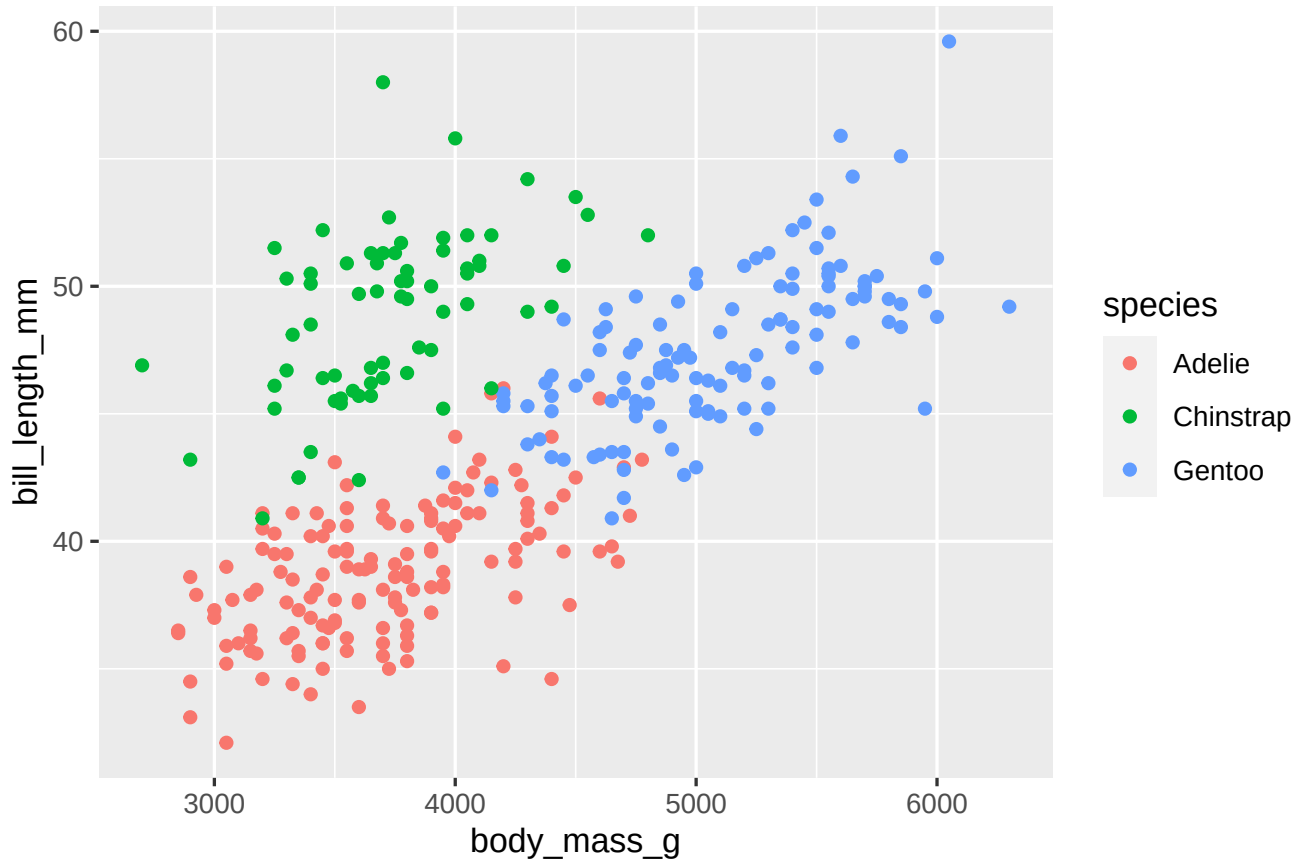
# Scatterplot with Color!

Let's compare `bill_length_mm` with `body_mass_g` and `species`

# Scatterplot with Color!

Let's compare `bill_length_mm` with `body_mass_g` and `species`

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm, color = species))+
  geom_point()
```

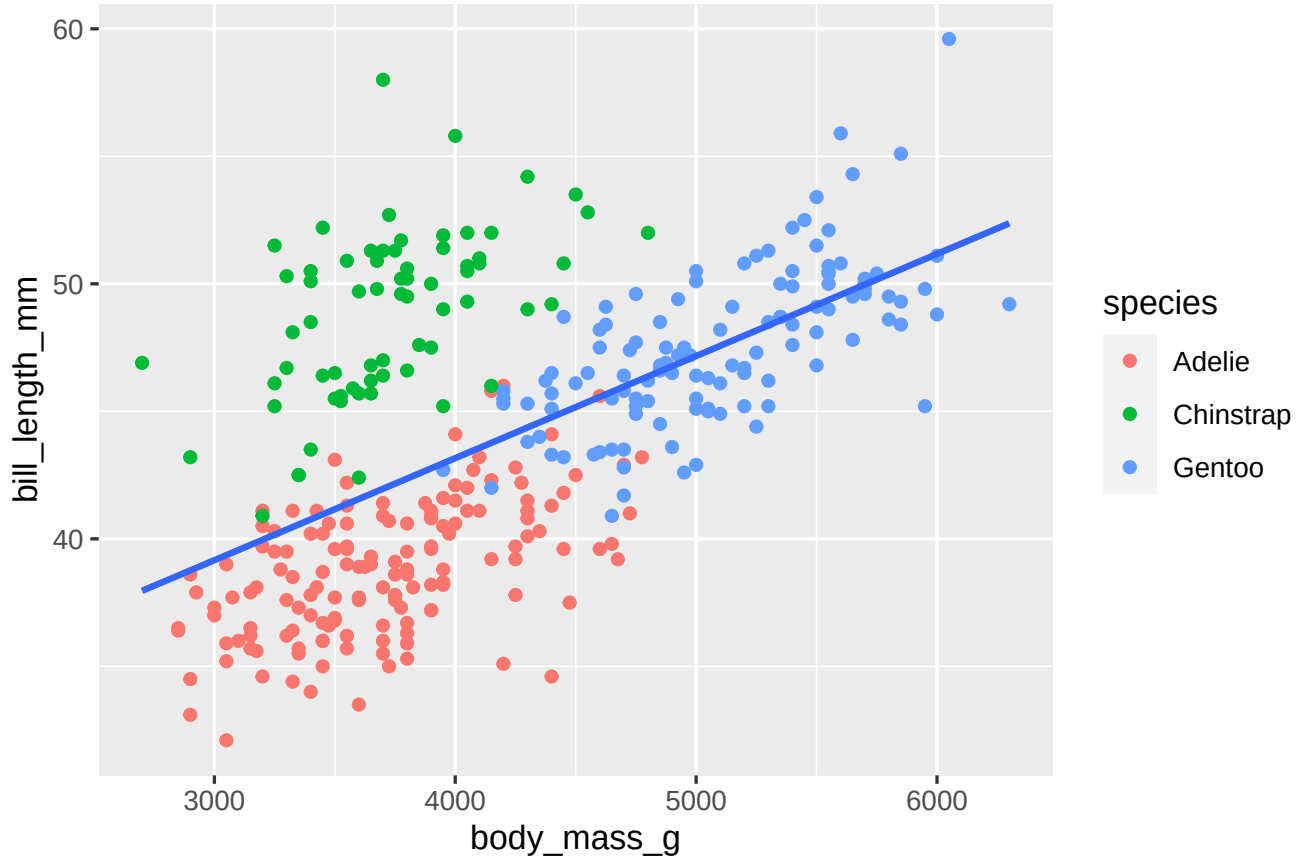


# Scatterplot with Color!

How does our linear model do for each species?

# Scatterplot with Color!

How does our linear model do for each species?



# Multiple Regression

Can we build a model that takes the best features of the linear model with `body_mass_g` $x$ and the model with `species` $s$?

$$\hat{Y} = 27.151 + .004x$$

$$\hat{Y} = 38.824 + 10.01 \cdot I_{\text{Chinstrap}}(s) + 8.744 \cdot I_{\text{Gentoo}}(s)$$

# Multiple Regression

Can we build a model that takes the best features of the linear model with `body_mass_g` $x$ and the model with `species` $s$?

$$\hat{Y} = 27.151 + .004x$$

$$\hat{Y} = 38.824 + 10.01 \cdot I_{\text{Chinstrap}}(s) + 8.744 \cdot I_{\text{Gentoo}}(s)$$

Yes!

$$\hat{Y} = \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_3 \cdot I_{\text{Gentoo}}(s)$$

We just need to refit the coefficients.

# Building the Multiple Regression Model

- To create a linear regression model with multiple variables, we still use the `lm()` and `get_regression_table()` functions.

# Building the Multiple Regression Model

- To create a linear regression model with multiple variables, we still use the `lm()` and `get_regression_table()` functions.
- To incorporate multiple explanatory variables, we use + in the model formula:

```
mlr_penguins <-lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(mlr_penguins)
```

```
## # A tibble: 4 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             24.9      1.09      22.9       0     22.8     27.1
## 2 body_mass_g            0.004    0         13.0       0      0.003    0.004
## 3 speciesChinstrap       9.91     0.355     27.9       0      9.21    10.6
## 4 speciesGentoo          3.54     0.5        7.08      0      2.56     4.52
```

# Building the Multiple Regression Model

- To create a linear regression model with multiple variables, we still use the `lm()` and `get_regression_table()` functions.
- To incorporate multiple explanatory variables, we use + in the model formula:

```
mlr_penguins <-lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(mlr_penguins)
```

```
## # A tibble: 4 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept          24.9       1.09      22.9        0   22.8     27.1
## 2 body_mass_g         0.004     0         13.0        0    0.003    0.004
## 3 speciesChinstrap    9.91      0.355     27.9        0    9.21    10.6
## 4 speciesGentoo       3.54      0.5        7.08       0    2.56     4.52
```

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

# Using the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

# Using the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The predicted `bill_length_mm` for a Chinstrap penguin with `body_mass_g` of 3500 is

$$\hat{Y} = 24.9 + 0.004 \cdot 3500 + 9.91 = 48.81$$

# Using the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The predicted `bill_length_mm` for a Chinstrap penguin with `body_mass_g` of 3500 is

$$\hat{Y} = 24.9 + 0.004 \cdot 3500 + 9.91 = 48.81$$

- The predicted `bill_length_mm` for a Gentoo penguin with `body_mass_g` of 4000 is

$$\hat{Y} = 24.9 + 0.004 \cdot 4000 + 3.54 = 44.44$$

# Using the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The predicted `bill_length_mm` for a Chinstrap penguin with `body_mass_g` of 3500 is

$$\hat{Y} = 24.9 + 0.004 \cdot 3500 + 9.91 = 48.81$$

- The predicted `bill_length_mm` for a Gentoo penguin with `body_mass_g` of 4000 is

$$\hat{Y} = 24.9 + 0.004 \cdot 4000 + 3.54 = 44.44$$

- The predicted `bill_length_mm` for a Adelie penguin with `body_mass_g` of 3500 is

$$\hat{Y} = 24.9 + 0.004 \cdot 3500 = 38.9$$

# Interpreting the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

# Interpreting the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The coefficient on the quantitative variable still tells us how much the response variable changes per unit increase in the explanatory variable.

# Interpreting the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The coefficient on the quantitative variable still tells us how much the response variable changes per unit increase in the explanatory variable.
  - Every 1 gram increase in body mass corresponds to a 0.004 mm increase in bill-length

# Interpreting the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The coefficient on the quantitative variable still tells us how much the response variable changes per unit increase in the explanatory variable.
  - Every 1 gram increase in body mass corresponds to a 0.004 mm increase in bill-length
- The intercept of the model depends on the penguin species.

# Interpreting the Multiple Regression Equation

Multiple Regression Equation:

$$\hat{Y} = 24.9 + 0.004x + 9.91 \cdot I_{\text{Chinstrap}}(s) + 3.54 \cdot I_{\text{Gentoo}}(s)$$

- The coefficient on the quantitative variable still tells us how much the response variable changes per unit increase in the explanatory variable.
  - Every 1 gram increase in body mass corresponds to a 0.004 mm increase in bill-length
- The intercept of the model depends on the penguin species.
  - The linear model for Gentoo penguins is

$$\hat{Y} = 24.9 + 0.004x + 3.54 = 28.44 + 0.004x$$

# Plotting the Multiple Regression Equation

We can plot the lines of best fit by adding a `geom_parallel_slopes` layer.

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm, color = species))+
  geom_point()+
  geom_parallel_slopes(se=F)
```

# Plotting the Multiple Regression Equation

We can plot the lines of best fit by adding a `geom_parallel_slopes` layer.

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm, color = species))+
  geom_point()+
  geom_parallel_slopes(se=F)
```


Is parallel slopes an appropriate assumption?

# Interaction Model

We can remove the parallel slopes assumption by including **interaction terms** in the linear model:

$$\hat{Y} = \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_3 \cdot I_{\text{Gentoo}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) + \beta_5 x \cdot I_{\text{Gentoo}}(s)$$

# Interaction Model

We can remove the parallel slopes assumption by including **interaction terms** in the linear model:

$$\hat{Y} = \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_3 \cdot I_{\text{Gentoo}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) + \beta_5 x \cdot I_{\text{Gentoo}}(s)$$

- In this model, each species of penguin has its own intercept AND its own slope.

# Interaction Model

We can remove the parallel slopes assumption by including **interaction terms** in the linear model:

$$\hat{Y} = \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_3 \cdot I_{\text{Gentoo}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) + \beta_5 x \cdot I_{\text{Gentoo}}(s)$$

- In this model, each species of penguin has its own intercept AND its own slope.
- For example, the linear model for Chinstrap penguins is

$$\begin{aligned}\hat{Y} =& \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) \\ =& (\beta_0 + \beta_1) + (\beta_1 + \beta_4) x\end{aligned}$$

# Interaction Model

We can remove the parallel slopes assumption by including **interaction terms** in the linear model:

$$\hat{Y} = \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_3 \cdot I_{\text{Gentoo}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) + \beta_5 x \cdot I_{\text{Gentoo}}(s)$$

- In this model, each species of penguin has its own intercept AND its own slope.
- For example, the linear model for Chinstrap penguins is

$$\begin{aligned}\hat{Y} =& \beta_0 + \beta_1 x + \beta_2 \cdot I_{\text{Chinstrap}}(s) + \beta_4 x \cdot I_{\text{Chinstrap}}(s) \\ =& (\beta_0 + \beta_1) + (\beta_1 + \beta_4) x\end{aligned}$$

- These interaction terms correspond to possible synergy between the values of $x$ and $s$.
  - That is, the effect of `body_mass_g` on `bill_length_mm` **depends** on species

# Building the Interaction Model

To create the interaction model, we use a * in the model formula instead of +

```
interaction_penguins <-lm(bill_length_mm ~ body_mass_g*species,
                          data = penguins)
get_regression_table(interaction_penguins, digits = 4)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1     1.63        16.6     0       2.39e+1  30.3
## 2 body_mass_g                0.0032  0.0004       7.23    0       2.30e-3   0.004
## 3 speciesChinstrap           5.06    3.31         1.53    0.127  -1.45e+0  11.6
## 4 speciesGentoo             -0.575   2.79        -0.206   0.837  -6.07e+0   4.92
## 5 body_mass_g:speciesChi~    0.0013  0.0009       1.48    0.141  -4.00e-4   0.003
## 6 body_mass_g:speciesGen~    0.001   0.000600     1.56    0.120  -3.00e-4   0.0022
```

# Building the Interaction Model

To create the interaction model, we use a * in the model formula instead of +

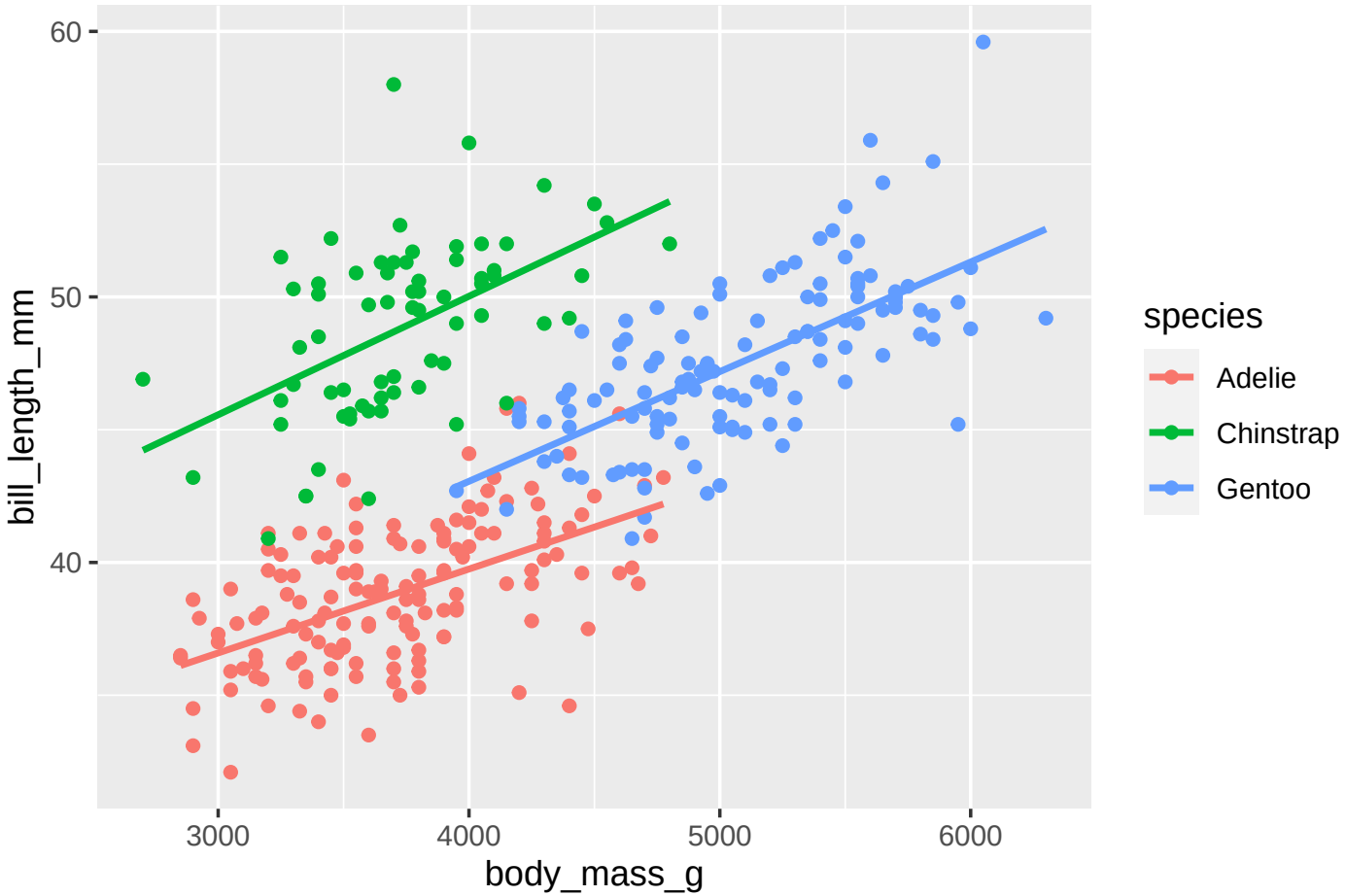
```
interaction_penguins <-lm(bill_length_mm ~ body_mass_g*species,
                          data = penguins)
get_regression_table(interaction_penguins, digits = 4)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1     1.63        16.6     0      2.39e+1  30.3
## 2 body_mass_g                0.0032  0.0004       7.23    0      2.30e-3   0.004
## 3 speciesChinstrap           5.06    3.31         1.53    0.127 -1.45e+0  11.6
## 4 speciesGentoo             -0.575   2.79        -0.206   0.837 -6.07e+0   4.92
## 5 body_mass_g:speciesChi~    0.0013  0.0009       1.48    0.141 -4.00e-4   0.003
## 6 body_mass_g:speciesGen~    0.001   0.000600     1.56    0.120 -3.00e-4   0.0022
```

$$\hat{Y} = 27.1 + 0.0032x + 5.06 \cdot I_{\text{Chinstrap}}(s) - 0.575 \cdot I_{\text{Gentoo}}(s) + 0.0013x \cdot I_{\text{Chinstrap}}(s) + 0.001x \cdot I_{\text{Gentoo}}(s)$$

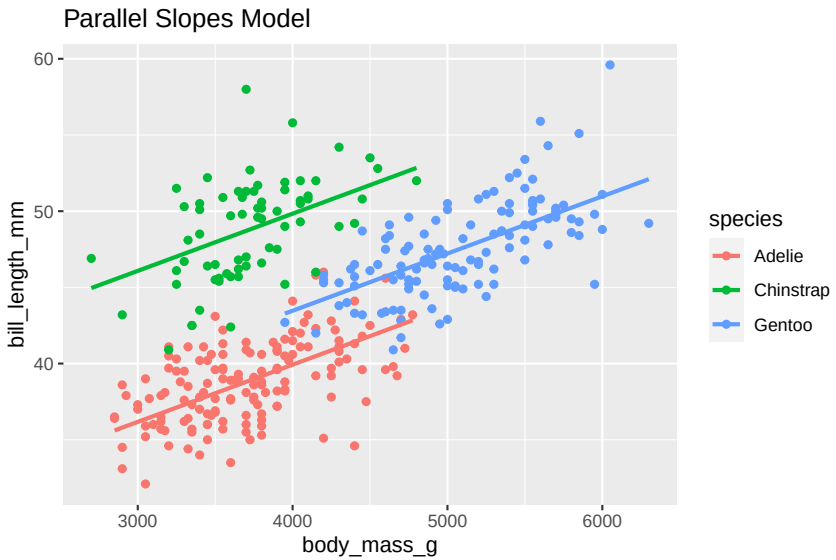
# Interaction Plot

```
ggplot(penguins, aes(x = body_mass_g, y =bill_length_mm, color = species))+
  geom_point()+
  geom_smooth(method = "lm", se=F)
```



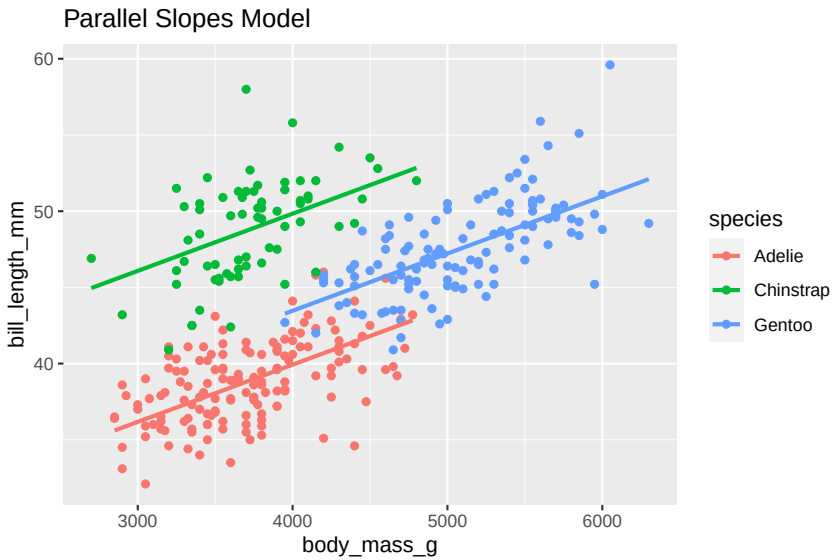
# Parallel Slopes vs. Interaction Model

Which model is preferable?




# Parallel Slopes vs. Interaction Model

Which model is preferable?

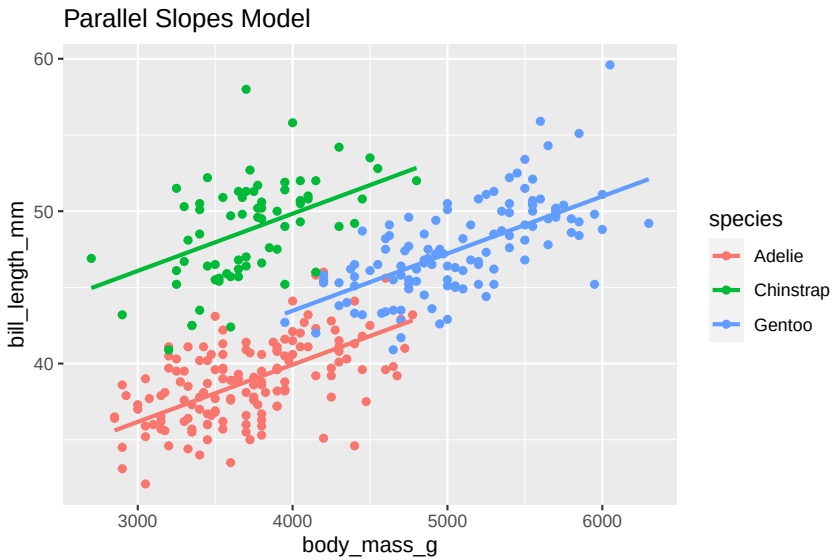




All else equal, we should select the simplest model that reasonably explains the data.

# Parallel Slopes vs. Interaction Model

Which model is preferable?





All else equal, we should select the simplest model that reasonably explains the data.

- In this case, the parallel slopes model is superior.