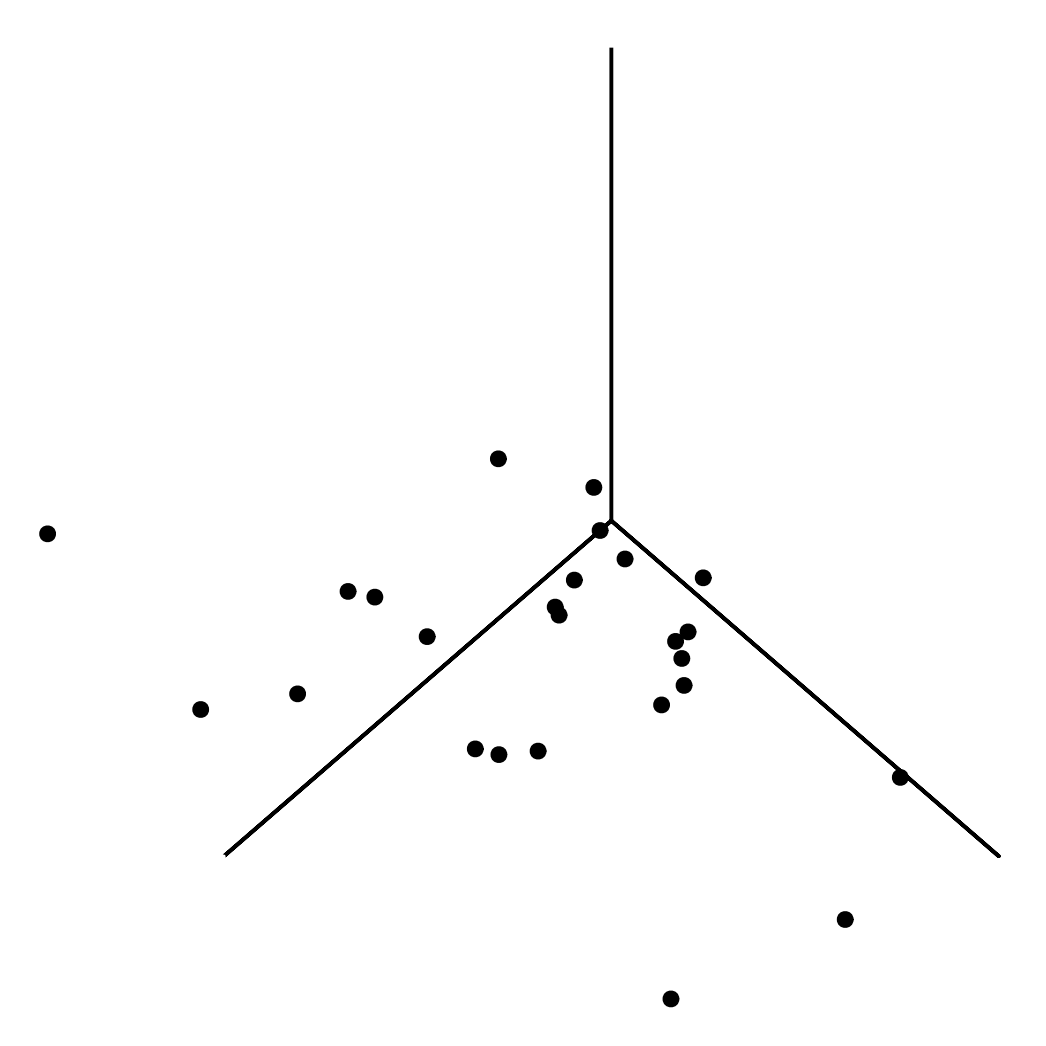
Categorical predictors

Stat 230

April 15 2022

Overview



Today:

- Indicator variables
- No interaction model
- Interaction model
- Model interpretation

MLR Variables

- $Y =$ quantitative response
- $x_1, \ldots, x_p$: p explanatory (predictor) variables
  - X_j can be *either* quantitative or categorical
  - **Today:** adding and interpreting categorical predictors

Quantitative predictors

- **Quantitative** predictor effects on Y are interpreted by quantitative changes: How does the mean response change for
  - a one unit increase in x
  - a 10% reduction in x
- In a "basic" model (e.g. no interactions, polynomial terms, etc)
  - any one unit increase in x_j results in a β_j change in $\mu_{y|x}$

Categorical predictors

Categorical predictor effects on Y are interpreted by changing the level of the categorical variable: How does the mean response differ when

- comparing level "A" to level "B"
- comparing level "A" to level "C"

In a "basic" model (e.g. no interactions, polynomial terms, etc)

- we don't assume that the difference in $\mu_{y|x}$ is the same for all level combinations
- the mean difference between levels "A" and "B" is not necessarily the same as the difference between levels "A" and "C"

`Palmerpenguins` dataset

```
library(palmerpenguins) # package hosting penguins data
library(dplyr) # package for data wrangling
penguins %>% glimpse() # get a glimpse of your data
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
$ sex               <fct> male, female, female, NA, female, male, female, male…
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```

Indicator variables

Categorical variables are represented in regression models as **indicator** (aka dummy) variables

- entries are either "1" or "0"
- a "1" indicates one particular level
- if we have k levels, we need $k - 1$ indicator variables

species	indicator_Chinstrap	indicator_Gentoo
Adelie	0	0
Gentoo	0	1
Adelie	0	0
Chinstrap	1	0
Chinstrap	1	0
...	...	...

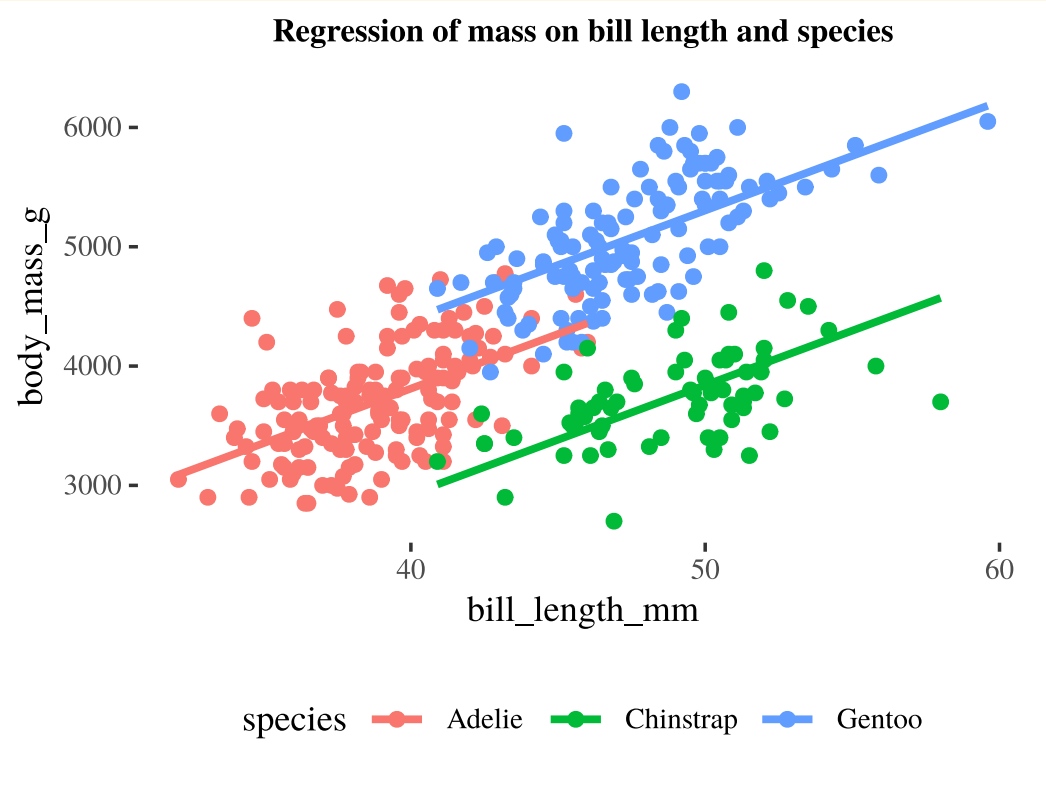
Indicator variables

species	indicator_Chinstrap	indicator_Gentoo
Adelie	0	0
Gentoo	0	1
Adelie	0	0
Chinstrap	1	0
Chinstrap	1	0
...	...	...

- The baseline level is the level that doesn't have an indicator variable
- Adelie is the baseline level in `penguins` example
- R will automatically create indicator variables when using a factor or character variable in an `lm`
- In R, the baseline level is the first level of a variable with order usually determined alphabetically

Mean function without interaction (parallel lines model)

$$\mu_{\text{mass}\,|x} = \beta_0 + \beta_1\ \text{bill}\ + \beta_2\ \text{speciesChinstrap}\ + \beta_3\ \text{speciesGentoo}$$



- Mean function for Adelie

$$\begin{aligned}\mu_{\text{mass}\,|\,\text{bill}\,,\,\text{Adelie}} &= \beta_0 + \beta_1\ \text{bill}\ + \beta_2(0) + \beta_3(0)\\ &= \beta_0 + \beta_1\ \text{bill}\end{aligned}$$

- Mean function for Chinstrap

$$\begin{aligned}\mu_{\text{mass}\,|\,\text{bill}\,,\,\text{Chinstrap}} &= \beta_0 + \beta_1\ \text{bill}\ + \beta_2(1) + \beta_3(0)\\ &= \beta_0 + \beta_2 + \beta_1\ \text{bill}\end{aligned}$$

- Mean function for Gentoo

$$\begin{aligned}\mu_{\text{mass}\,|\,\text{bill}\,,\,\text{Gentoo}} &= \beta_0 + \beta_1\ \text{bill}\ + \beta_2(0) + \beta_3(1)\\ &= \beta_0 + \beta_3 + \beta_1\ \text{bill}\end{aligned}$$

Example: Penguins

- β_1 : effect of bill length on mass, holding species fixed
- β_2 : holding bill length fixed, β_2 is the difference in mean mass between Chinstrap and Adelie

$$\mu_{\text{mass} \mid \text{Chinstrap}} - \mu_{\text{mass} \mid \text{Adelie}} = \beta_2$$

- β_3 : holding bill length fixed, β_3 is the difference in mean mass between Gentoo and Adelie

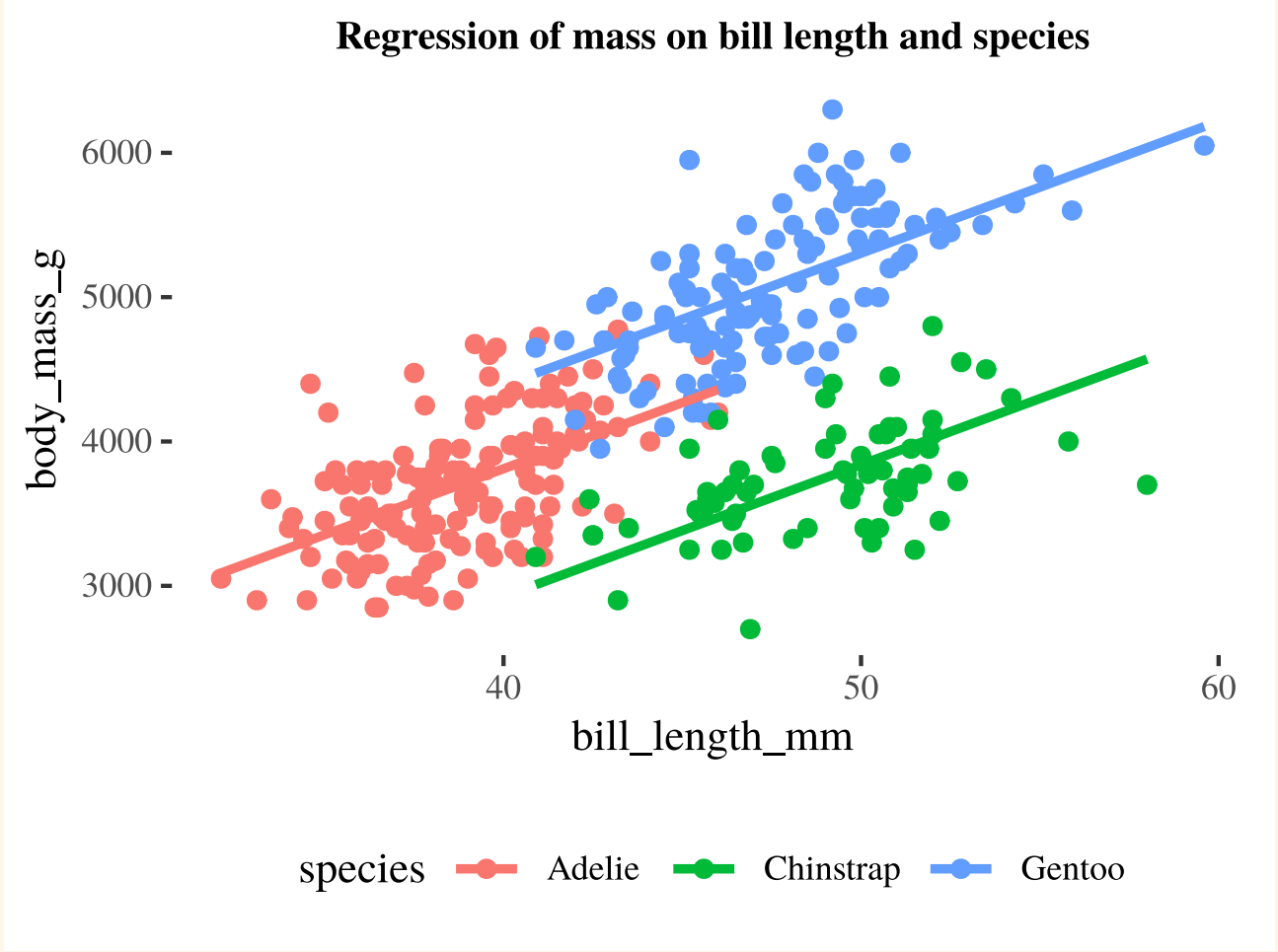
$$\mu_{\text{mass} \mid \text{Gentoo}} - \mu_{\text{mass} \mid \text{Adelie}} = \beta_3$$

- $\beta_2 - \beta_3$: holding bill length fixed, $\beta_2 - \beta_3$ is the difference in mean mass between Chinstrap and Gentoo

$$\mu_{\text{mass} \mid \text{Chinstrap}} - \mu_{\text{mass} \mid \text{Gentoo}} = \beta_2 - \beta_3$$

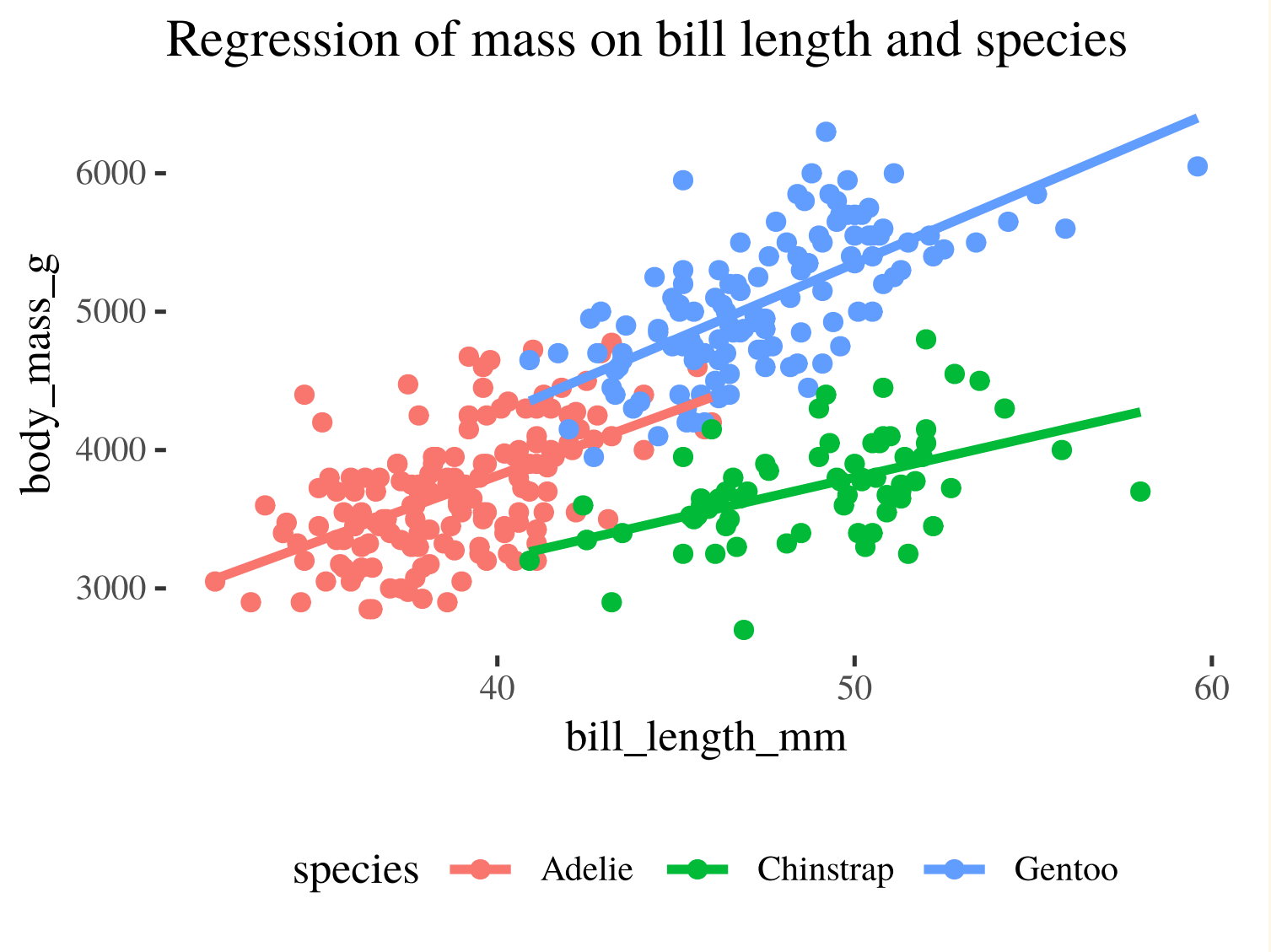
Visualizing the fitted parallel lines (No interaction!)

```
library(moderndive)
peng_parallel_lm <-  lm(body_mass_g ~ bill_length_mm
peng_parallel_points <- get_regression_points(peng_p
ggplot(peng_parallel_points,
 aes(bill_length_mm, body_mass_g, color = species))
 theme(legend.position = "bottom") +
 geom_point() +
 geom_line(aes(y = body_mass_g_hat), size = 1) +
 labs(title = "Regression of mass on bill length and
theme(plot.title = element_text(hjust=0.5, size=9, f
```

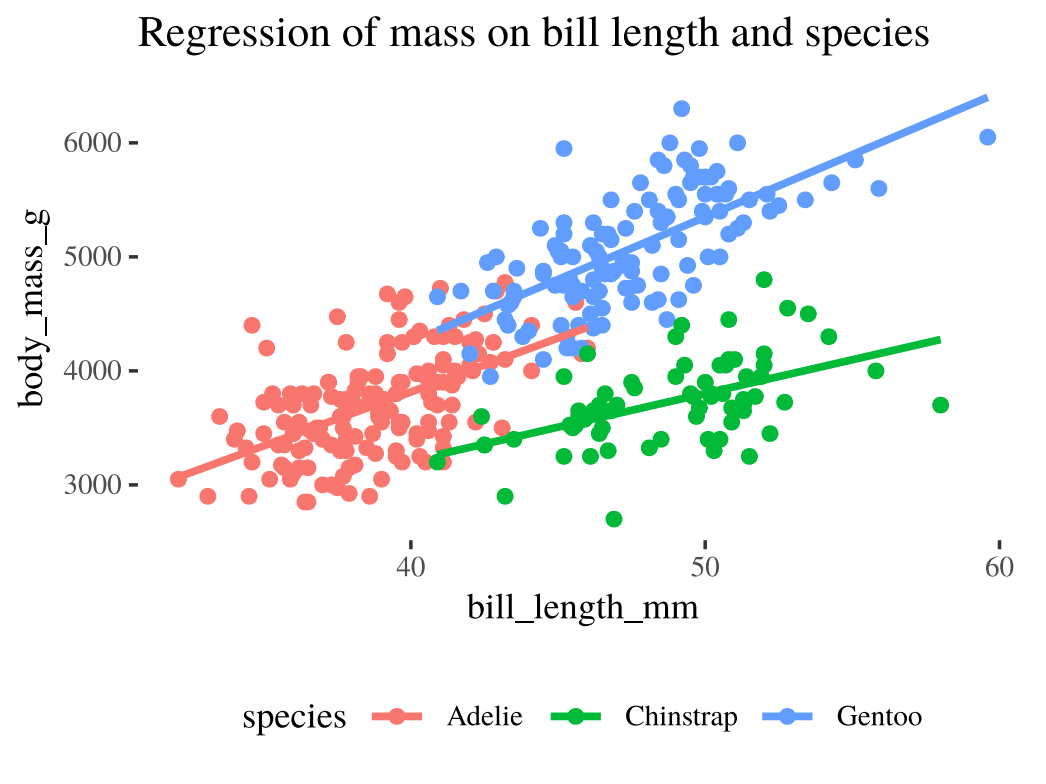


Example: Penguins

```
# adding a color aesthetics fits the m
ggplot(penguins, aes(x = bill_length_m
                     y = body_mass_g,
                     color = species))
theme(legend.position = "bottom")+
geom_point() +
geom_smooth(method = "lm", se = FALSE)
labs(title = "Regression of mass on bi
```



Example: Penguins

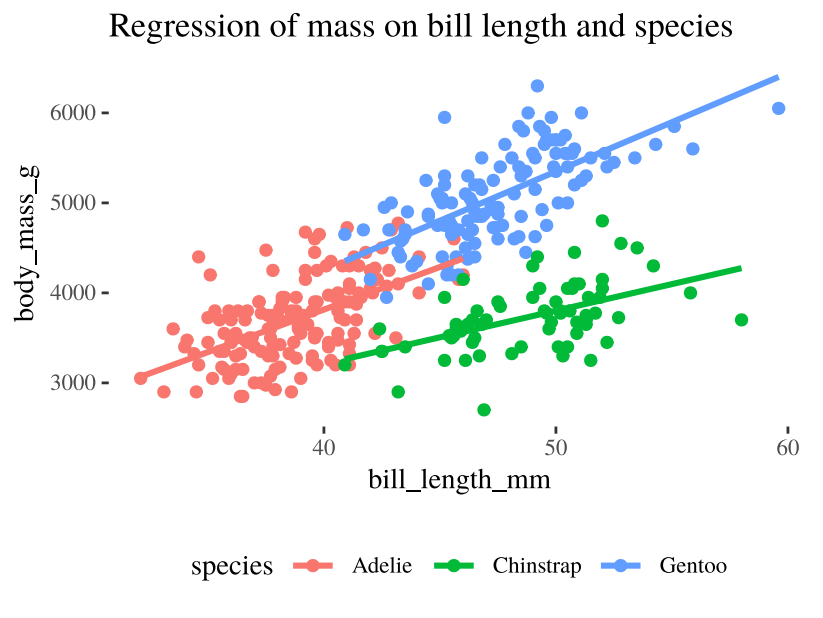


Does the effect of bill length on body mass differed by species?

- If yes, then we need an interaction to create separate lines for each species!

Example: Penguins interaction

$$\mu_{\text{mass}\,|x} = \beta_0 + \beta_1\ \text{bill} + \beta_2\ \text{speciesChinstrap} + \beta_3\ \text{speciesGentoo} + \beta_4\ \text{bill} \times \text{speciesChinstrap} + \beta_5\ \text{bill} \times \text{speciesGentoo}$$



- Mean function for Adelie

$$\mu_{\text{mass}\,|x} = \beta_0 + \beta_1\ \text{bill} + \beta_2(0) + \beta_3(0) + \beta_4 bill \times (0) + \beta_5 bill \times (0)$$
$$= \beta_0 + \beta_1\ \text{bill}$$

- Mean function for Chinstrap

$$\mu_{\text{mass}\,|x} = \beta_0 + \beta_1\ \text{bill} + \beta_2(1) + \beta_3(0) + \beta_4 bill \times (1) + \beta_5 bill \times (0)$$
$$= \beta_0 + \beta_2 + (\beta_1 + \beta_4)\ \text{bill}$$

- Mean function for Gentoo

$$\mu_{mass|x} = \beta_0 + \beta_1\ \text{bill} + \beta_2(0) + \beta_3(1) + \beta_4 bill \times (0) + \beta_5 bill \times (1)$$
$$= \beta_0 + \beta_3 + (\beta_1 + \beta_5)\ \text{bill}$$

Example: Penguins interaction

- Effect of bill length on mass (**slope**) depends on species
  - β_1 : effect of bill length on mass for Adelie (baseline)
  - $\beta_1 + \beta_4$: effect of bill length on mass for Chinstrap
  - $\beta_1 + \beta_5$: effect of bill length on mass for Gentoo

Example: Penguins interaction

- Difference between mean mass between species depends on bill length
- Holding bill length fixed, $\beta_2 + \beta_4 bill$ is the difference in mean mass between Chinstrap and Adelie

$$\begin{aligned}\mu_{\text{mass} \mid \text{Chinstrap}} - \mu_{\text{mass} \mid \text{Adelie}} &= [\beta_0 + \beta_2 + (\beta_1 + \beta_4)\, bill] - [\beta_0 + \beta_1 \text{ bill }] \\ &= \beta_2 + \beta_4 bill\end{aligned}$$

Example: Penguins interaction

- Difference between mean mass between species depends on bill length
- Holding bill length fixed, $\beta_3 + \beta_5 bill$ is the difference in mean mass between Gentoo and Adelie

$$\begin{aligned}\mu_{\text{mass} \mid \text{Gentoo}} - \mu_{\text{mass} \mid \text{Adelie}} &= [\beta_0 + \beta_3 + (\beta_1 + \beta_5)\, bill] - [\beta_0 + \beta_1 \text{ bill }] \\ &= \beta_3 + \beta_5 bill\end{aligned}$$

Example: Penguins interaction

- Difference between mean mass between species depends on bill length
- Holding bill length fixed, $(\beta_2 - \beta_3) + (\beta_4 - \beta_5)\, bill$ is the difference in mean mass between Chinstrap and Gentoo

$$\begin{aligned}\mu_{\text{mass} \mid \text{Chinstrap}} - \mu_{\text{mass} \mid \text{Gentoo}} &= [\beta_0 + \beta_2 + (\beta_1 + \beta_4)\ \text{bill}\,] - [\beta_0 + \beta_3 + (\beta_1 + \beta_5)\ \text{bill}\,] \\ &= (\beta_2 - \beta_3) + (\beta_4 - \beta_5)\, bill\end{aligned}$$

Fit the model with interaction

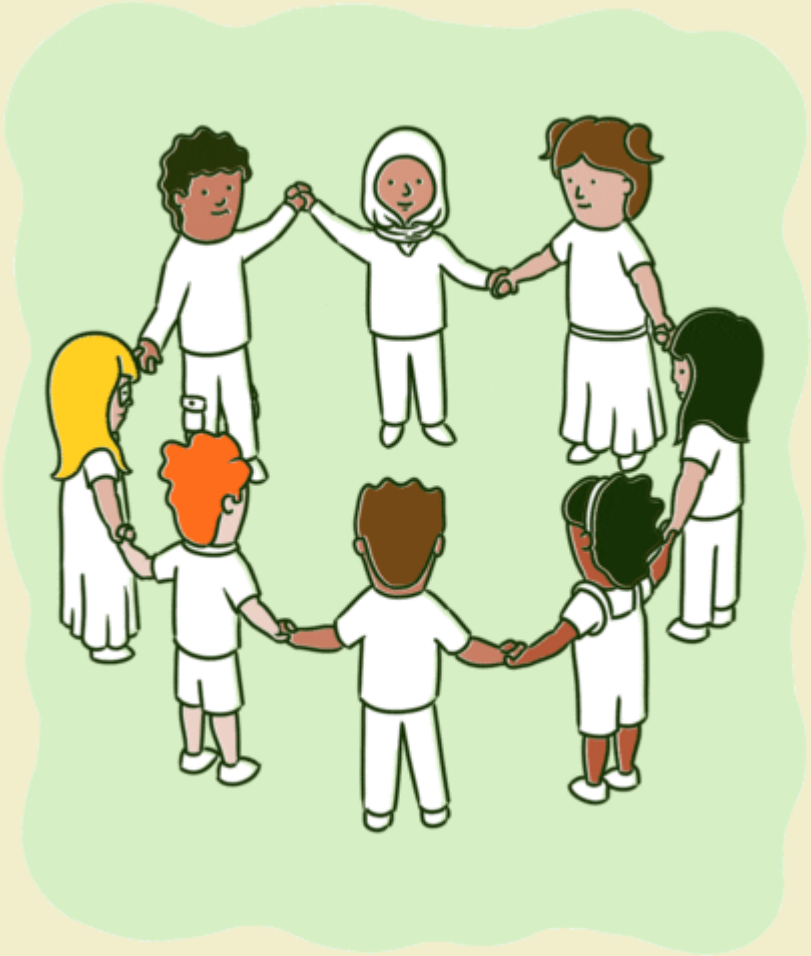
```
library(moderndive) # call the library
# can use either y ~ x1*x2 or y ~ x1 + x2 + x1:x2
peng_interaction_lm <-  lm(body_mass_g ~ bill_length_mm*species, data = penguins)
peng_table_interaction <- get_regression_table(peng_interaction_lm)
knitr::kable(peng_table_interaction, digis= 4, format = "html")
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	34.883	443.176	0.079	0.937	-836.866	906.632
bill_length_mm	94.500	11.398	8.291	0.000	72.080	116.920
species: Chinstrap	811.260	799.806	1.014	0.311	-761.997	2384.517
species: Gentoo	-158.711	683.191	-0.232	0.816	-1502.582	1185.160
bill_length_mm:speciesChinstrap	-35.382	17.747	-1.994	0.047	-70.291	-0.474
bill_length_mm:speciesGentoo	14.959	15.786	0.948	0.344	-16.093	46.012

$$\mu_{\text{mass}\,|x} = 34.88 + 94.50\ \text{bill} + 811.26\ \text{speciesChinstrap} - 158.71\ \text{speciesGentoo} - 35.38\ \text{bill} \times \text{speciesChinstrap} + 14.96\ \text{bill} \times \text{speciesGentoo}$$

Your Turn 2

05:00



Work through Problem 2 of the categorical predictors worksheet with a neighbor