Two Parallel Lines: Crocodiles!!

Multiple Regression

ANOVA models have:

- a quantitative response variable (sepal width of a flower) and
- one categorical explanatory variable (species)
- Separate mean sepal width for each species, individual values normally distributed around the mean

Simple linear regression models have:

- a quantitative response variable (college graduation rate) and
- one quantitative explanatory variable (college acceptance rate)
- Mean graduation rate is a linear function of acceptance rate, individual values normally distributed around the mean

Multiple regression models have:

- a quantitative response variable and
- more than one explanatory variable, may be a mix of categorical and quantitative
- Examples:
  - $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - $\mu(Y|X_1, X_2, X_3, X_4) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$
  - $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2)$
  - $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \beta_4 X_1^2$

We will start by combining one categorical explanatory variable and one quantitative explanatory variable.

Example of Two Parallel Lines

We have measurements of the head length (cm) and total body length (cm) of 32 crocodiles of two different species:

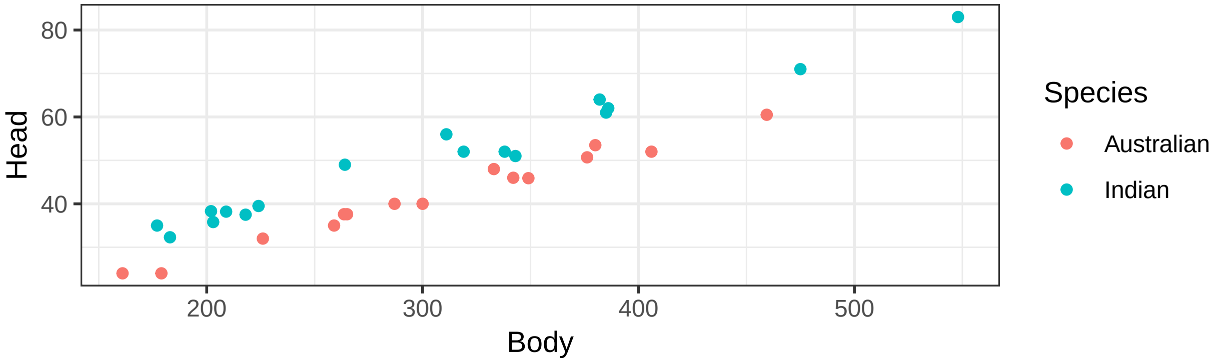
```
head(crocs)
```

```
##    Body Head    Species
## 1 338.0 52.0     Indian
## 2 333.0 48.0 Australian
## 3 202.0 38.3     Indian
## 4 406.0 52.0 Australian
## 5 459.4 60.5 Australian
## 6 264.0 49.0     Indian
```

```
nrow(crocs)
```

```
## [1] 32
```

```
ggplot(data = crocs) +
  geom_point(mapping = aes(x = Body, y = Head, color = Species)) +
  theme_bw()
```



2 lines by filtering to create separate data sets

```r
aus_crocs <- crocs %>% filter(Species == "Australian")
aus_fit <- lm(Head ~ Body, data = aus_crocs)
summary(aus_fit)
```

```
##
## Call:
## lm(formula = Head ~ Body, data = aus_crocs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3529 -0.9968  0.0824  0.7419  2.7973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.463022   1.523732   2.273   0.0407 *
## Body        0.125344   0.004819  26.010 1.35e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.504 on 13 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9797
## F-statistic: 676.5 on 1 and 13 DF,  p-value: 1.35e-12
```

```r
ind_crocs <- crocs %>% filter(Species == "Indian")
ind_fit <- lm(Head ~ Body, data = ind_crocs)
summary(ind_fit)
```

```
##
## Call:
## lm(formula = Head ~ Body, data = ind_crocs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5756 -1.6627 -0.0904  1.2208  4.6261
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.538438   1.861787    5.66 4.53e-05 ***
## Body         0.131304   0.005791   22.68 5.08e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.503 on 15 degrees of freedom
## Multiple R-squared:  0.9717, Adjusted R-squared:  0.9698
## F-statistic: 514.2 on 1 and 15 DF,  p-value: 5.08e-13
```

Questions we'd like to be able to answer (but can't with this output):

1. How strong is the evidence that the intercepts for these lines are different? (today)
2. How strong is the evidence that the slopes for these lines are different? (next class)

2 parallel lines (same slope)

- Our Goal: Equations for two lines

$$\begin{aligned}\text{Estimated Mean Head Length for Australian Crocs } &= \hat{\beta}_0^{Australian} + \hat{\beta}_1 \times (\text{Body Length}) \\ \text{Estimated Mean Head Length for Indian Crocs } &= \hat{\beta}_0^{Indian} + \hat{\beta}_1 \times (\text{Body Length})\end{aligned}$$

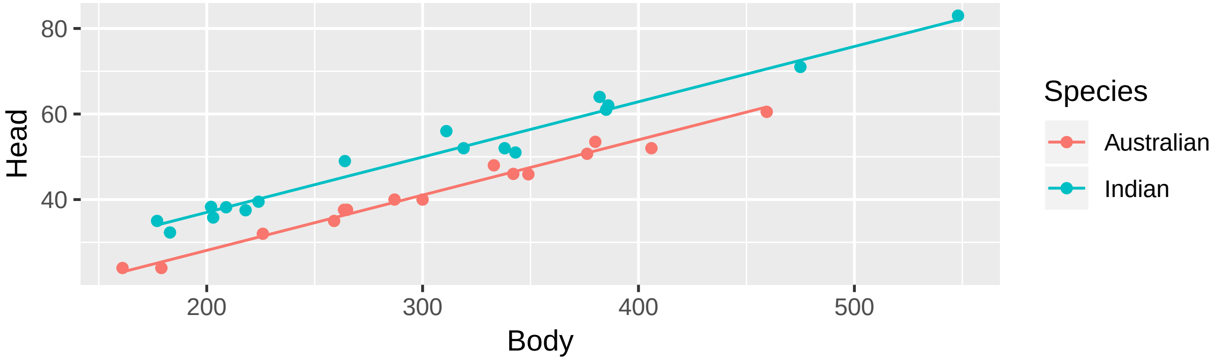
- Note: Different intercepts, same slope.

```r
parallel_lines_fit <- lm(Head ~ Body + Species, data = crocs)
summary(parallel_lines_fit)
```

```
##
## Call:
## lm(formula = Head ~ Body + Species, data = crocs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4959 -1.4218 -0.0842  1.0117  4.6405
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.265418   1.309167    1.73   0.0942 .
## Body          0.129261   0.003904   33.11  < 2e-16 ***
## SpeciesIndian 8.893772   0.737538   12.06 8.05e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 29 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.9755
## F-statistic:   617 on 2 and 29 DF,  p-value: < 2.2e-16
```

```r
crocs <- crocs %>%
  mutate(
    fitted = predict(parallel_lines_fit)
  )

ggplot(data = crocs) +
  geom_point(mapping = aes(x = Body, y = Head, color = Species)) +
  geom_line(mapping = aes(x = Body, y = fitted, color = Species))
```



- R gives us a single combined equation:

$$\text{Estimated Mean Head Length} = \hat{\mu}(\text{Head}|\text{Body}, \text{Species}) = \hat{\beta}_0 + \hat{\beta}_1(\text{Body}) + \hat{\beta}_2\text{SpeciesIndian}$$

$$\hat{\mu}(\text{Head}|\text{Body}, \text{Species}) = 2.27 + 0.13(\text{Body}) + 8.89\text{SpeciesIndian}$$

What is the `SpeciesIndian` variable?

- Behind the scenes, R creates a new **indicator variable** called `SpeciesIndian`:

$$\texttt{SpeciesIndian} = \begin{cases} 1 & \text{if the species for crocodile } i \text{ is Indian.} \\ 0 & \text{otherwise (in this case, the species is Australian)} \end{cases}$$

- R doesn't modify the data frame (it creates a secret copy in the background), but it would look like this:

```
head(crocs)
```

```
##    Body Head    Species   fitted SpeciesIndian
## 1 338.0 52.0     Indian 54.84955             1
## 2 333.0 48.0 Australian 45.30947             0
## 3 202.0 38.3     Indian 37.27000             1
## 4 406.0 52.0 Australian 54.74556             0
## 5 459.4 60.5 Australian 61.64811             0
## 6 264.0 49.0     Indian 45.28421             1
```

Above, we obtained this estimated equation:

$$\hat{\mu}(\text{Head}|\text{Body}, \text{Species}) = 2.27 + 0.13(\text{Body}) + 8.89\text{SpeciesIndian}$$

What is the estimated equation describing the relationship between body length and head length, for Australian crocodiles?

What is the estimated equation describing the relationship between body length and head length, for Indian crocodiles?

What is the interpretation of $\widehat{\beta}_0 = 2.27$?

What is the interpretation of $\widehat{\beta}_1 = 0.13$?

What is the interpretation of $\widehat{\beta}_2 = 8.89$?

Using the output from the summary function, conduct a test of the claim that a single regression line can be used to describe the relationship between body length and head length in the population of all Australian and Indian crocodiles.

Conduct a test of the claim that neither species nor body length are associated with head length in the population of all Australian and Indian crocodiles. (Note: formally, this is a test only of linear association with body length.)

Find and interpret a 95% confidence interval for β_2, the coefficient of `SpeciesIndian`.

```
confint(parallel_lines_fit)
```

```
##                    2.5 %     97.5 %
## (Intercept)   -0.4121302  4.9429659
## Body           0.1212763  0.1372466
## SpeciesIndian  7.3853376 10.4022072
```

Find and interpret a 95% confidence interval for the mean head length of the sub-population of Australian crocodiles that have a total body length of 400cm.

```
predict_data <- data.frame(
  Species = "Australian",
  Body = 400
)

predict(parallel_lines_fit, newdata = predict_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 53.96999 52.63765 55.30233
```