two_lines

Sunday, March 29, 2020 10:24 AM

two_lines

Two 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 Lines

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

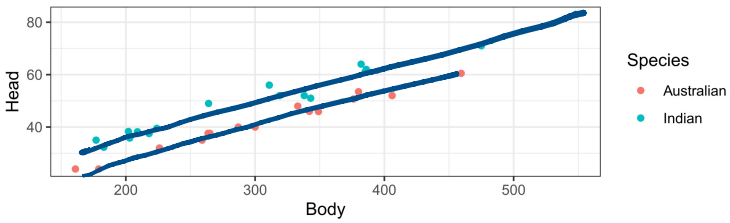
```
head(crocs)
```

```
##   Body Head    Species
## 1  349 45.9 Australian
## 2  183 32.3     Indian
## 3  179 24.0 Australian
## 4  218 37.5     Indian
## 5  311 56.0     Indian
## 6  338 52.0     Indian
```

```
nrow(crocs)
```

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

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


Head
80
60
40
200
300
400
500
Body
Species
Australian
Indian

2 lines by filtering to create separate data sets

```
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
```

$\hat{\beta}_0 = 3.463$

$\hat{\beta}_1 = 0.125$

```
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
```

$\hat{\beta}_0 = 10.538$

$\hat{\beta}_1 = 0.131$

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?
2. How strong is the evidence that the slopes for these lines are different?

2 parallel lines (same slope)

- Our Goal: Equations for two lines

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

- Note: Different intercepts, same slope.

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

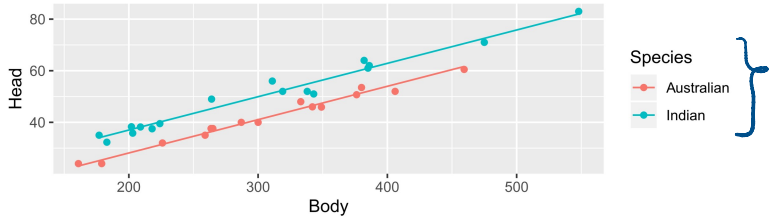
```
##
## Call:
## lm(formula = Head ~ Body + Species, data = crocs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4959 -1.4218 -0.0842  1.0117  4.6405
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.265418   1.309167    1.73   0.0942 .
## 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
```

$\hat{\beta}_2$

$\hat{\beta}_{Australian} = \hat{\beta}_0$, $\hat{\beta}_1$

$\hat{\beta}_{Indian} = \hat{\beta}_0 + \hat{\beta}_2$

```
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:

$$\text{Estimated Mean Head Length} = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \text{Body} + \hat{\beta}_2 \text{SpeciesIndian}$$

$$\hat{\mu} = 2.27 + 0.13\text{Body} + 8.89\text{SpeciesIndian}$$

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  349 45.9 Australian 47.37765             0
## 2  183 32.3      Indian 34.81403             1
## 3  179 24.0 Australian 25.40321             0
## 4  218 37.5      Indian 39.33818             1
## 5  311 56.0      Indian 51.35949             1
## 6  338 52.0      Indian 54.84955             1
```

Above, we obtained this estimated equation:

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

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

$$\hat{\mu}(Y|X_1 = x_1, X_2 = 0) = 2.27 + 0.13x_1 + 8.89 \times 0$$
$$= 2.27 + 0.13(\text{Body})$$

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

$$\hat{\mu}(Y|X_1 = x_1, X_2 = 1) = 2.27 + 0.13x_1 + 8.89 \times 1$$
$$= 11.18 + 0.13(\text{Body})$$

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

For a body length of 0 cm for an Australian crocodile, the expected head length is 2.27 cm on average, in a similar population.

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

For a 1 cm increase in body length, we expect a 0.13 cm increase in head length on average, in a similar population of crocodiles.

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

For an Indian crocodile, we expect the head length to be 8.89 cm longer, on average, than for an Australian crocodile, with a body length of 50 cm.

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.

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

There is very strong evidence ($p\text{-value} \approx 8 \times 10^{-13}$) that there is a different intercept for Indian crocodiles than for Australian crocodiles. The intercept is greater for Indian crocs.

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.)

$H_0: \beta_1 = 0$ "$\beta_2 = 0$" "$\beta_1 = 0$"

$H_A: \beta_1 \neq 0$

There is very strong evidence that body length is linearly associated with head length. The relationship is positive.

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
```

We are 95% confident that the mean difference in head length for an Indian croc versus an Australian croc w/ body length 0cm is between 7.385cm and 10.402cm. . . . Explain meaning of 95% CI.

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")
```

X_1

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

We are 95% confident that the mean head length for an Australian crocodile with a total body length of 400cm is between 52.638cm and 55.302cm.

1 model, 2 lines (different slopes)

- Our Goal: Equations for two lines

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

- Note: Different intercepts and slopes.
- To allow for different slopes, you have two options for the formula (the model that it fits is the same either way, the second is just a shorthand for the first):
  - `Body + Species + Body:Species`
  - `Body * Species`
- The term `Body:Species` is called the **interaction** between `Body` and `Species`. It is just the product of those two variables.

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

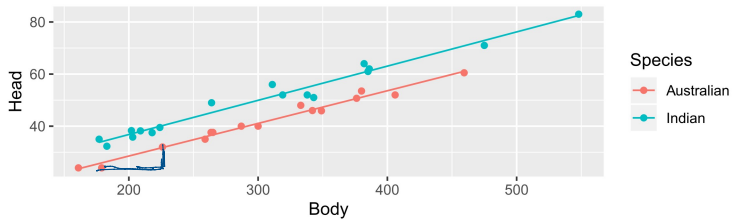
```
##
## Call:
## lm(formula = Head ~ Body + Species + Body:Species, data = crocs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5756 -1.3294 -0.0040  0.9646  4.6261
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.463022   2.126572   1.628   0.1146
## Body               0.125344   0.006726  18.637   <2e-16 ***
## SpeciesIndian      7.075415   2.638253   2.682   0.0121 *
## Body:SpeciesIndian 0.005959   0.008296   0.718   0.4785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.099 on 28 degrees of freedom
## Multiple R-squared:  0.9775, Adjusted R-squared:  0.975
## F-statistic: 404.6 on 3 and 28 DF,  p-value: < 2.2e-16
```

β_0 β_1 β_2 β_3

$H_0: \beta_3 = 0$

$H_A: \beta_3 \neq 0$

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



What is the estimated equation for the mean from this model?

$$\hat{\mu}(Y \mid X_1 = x_1, X_2 = x_2, X_3 = x_3) = 3.463 + 0.125 x_1 + 7.075 x_2 + 0.00596 x_3$$

(Body; Species Indian; Body : Species Indian)

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

$$\hat{\mu}(Y \mid X_1 = x_1, X_2 = 0, X_3 = 0) = 3.463 + 0.125 x_1$$
$$= 3.463 + 0.125(\text{Body})$$

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

$$\hat{\mu}(Y \mid X_1 = x_1, X_2 = 1, X_3 = 1 \cdot x_1) = (3.463 + 0.125 x_1) + (7.075(1) + 0.00596(1 \cdot x_1))$$
$$= 10.538 + 0.131 x_1 = 10.538 + 0.131(\text{Body})$$

What is the interpretation of $\hat{\beta}_0 = 3.463$?

For an Australian croc with full body length 0 cm, the expected head length will be 3.463cm on average.

What is the interpretation of $\hat{\beta}_1 = 7.075$? (contribution of Species to intercept)

For an Indian croc with a full body length of 0cm, the expected head length will be 7.075 cm larger, on average, than for an Australian croc of the same body length.

What is the interpretation of $\hat{\beta}_2 = 0.125$? (slope)

For an Australian croc, for a 1 cm increase in body length, we expect a 0.125 cm increased in head length, on average.

What is the interpretation of $\hat{\beta}_3 = 0.006$?

For an Indian croc, for a 1 cm increase in body length, we expect an additional 0.006 cm increase in head length, on average, relative to an Australian croc.

Using the output from the summary function, conduct a test of the claim that the *slope* of the line describing the relationship between body length and head length in the population of all Australian crocodiles is the same as the *slope* of the line describing the relationship between body length and head length in the population of all Indian crocodiles.

$H_0: \beta_3 = 0$

$H_A: \beta_3 \neq 0$

p-value = 0.4785

There is no statistical evidence to say that the slope of the line describing the relationship between body length an head length in the population of Indian crocodiles is different from that for Australian crocodiles.

'that for Australian crocodiles.