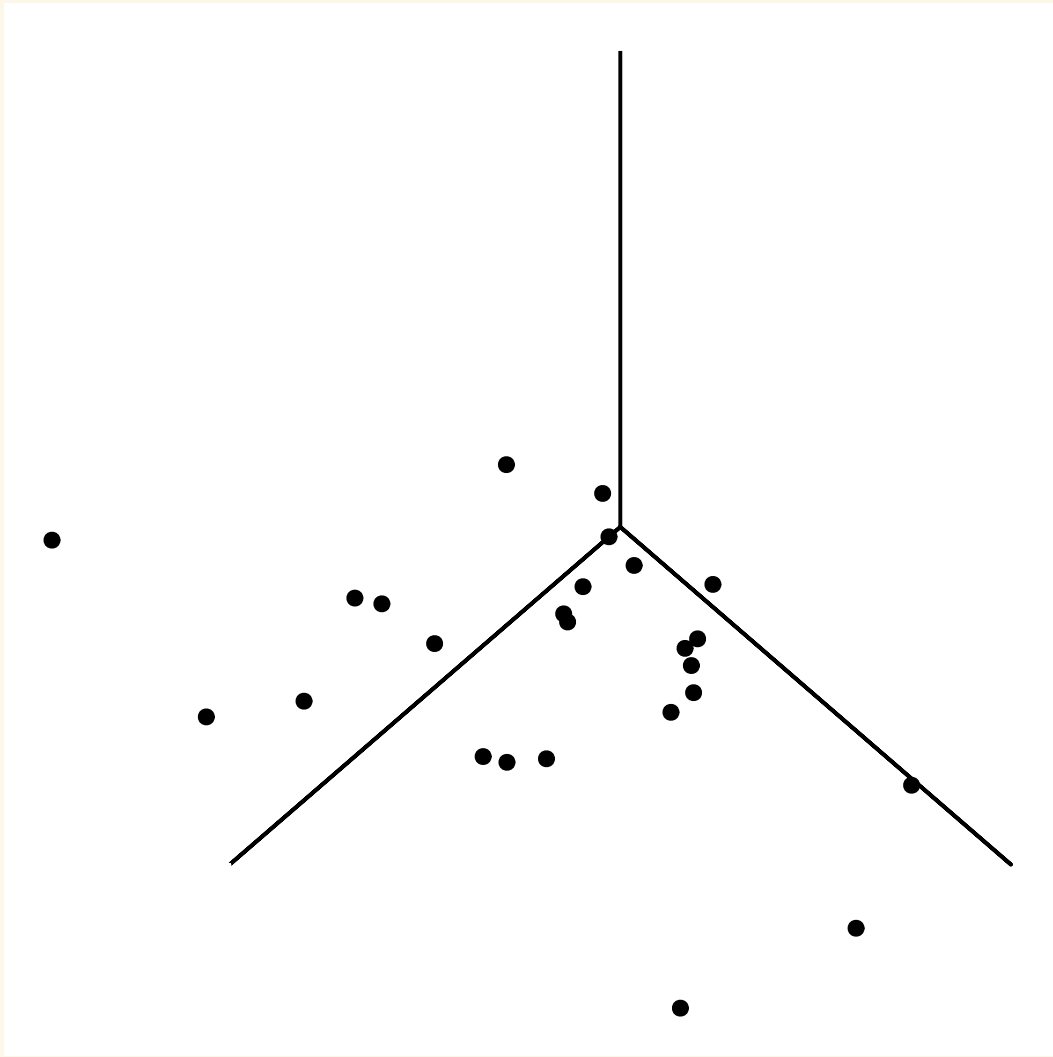


ANOVA for MLR

Stat 230

April 24 2022

Overview

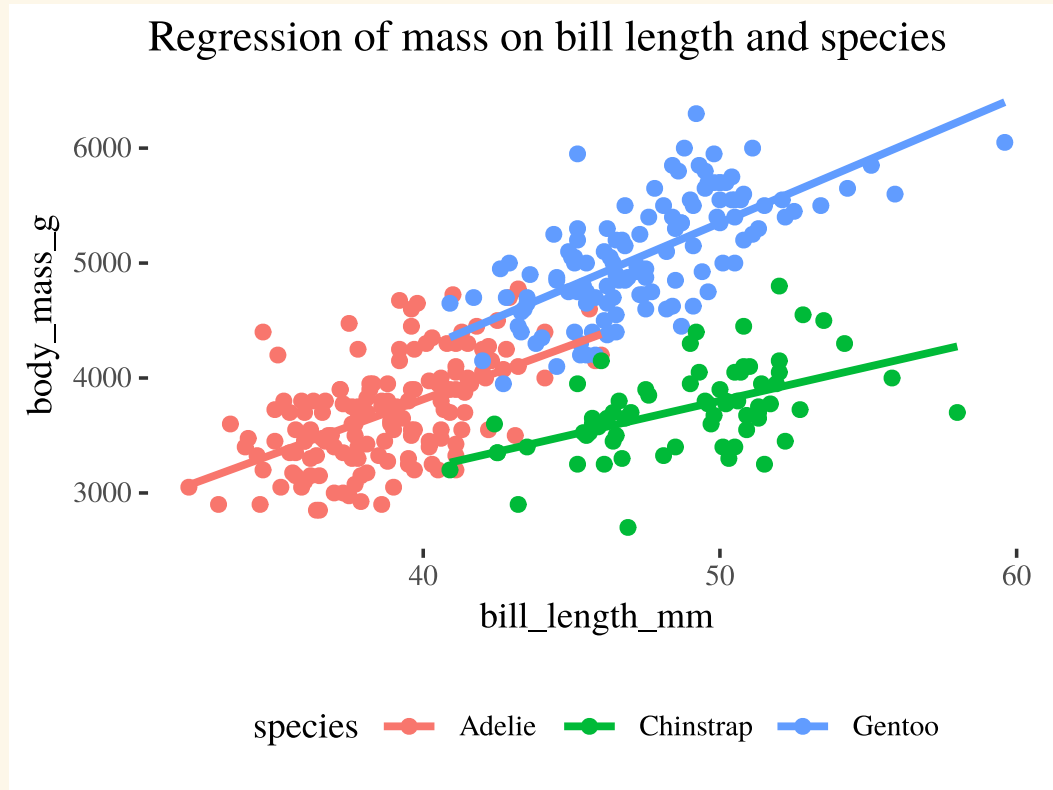


Today:

- ANOVA for MLR
- MLR Inference
- R^2 adjusted
- ANOVA F-test

Penguins example

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



- We will test whether the effect of bill length on mass depends on species:

$$H_0 : \beta_4 = \beta_5 = 0$$

Comparing models

Is the effect of bill length the same for Gentoo and Adelie (baseline)?

$$H_0 : \beta_5 = 0$$

$$H_A : \beta_5 \neq 0$$

Equivalent test:

$$H_0 : \mu_{\text{mass} | x} = \beta_0 + \beta_1 \text{ bill} + \beta_2 \text{ speciesChinstrap} + \beta_3 \text{ speciesGentoo} \\ + \beta_4 \text{ bill} \times \text{speciesChinstrap}$$

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

Individual t-tests: Is this term significant assuming all others are in the model?

Comparing models

Does the effect of bill length on mass depends on species?

$$H_0 : \beta_4 = \beta_5 = 0$$

$$H_A : \text{at least one } \beta_4, \beta_5 \neq 0$$

Equivalent test:

$$H_0 : \mu_{\text{mass} | x} = \beta_0 + \beta_1 \text{ bill} + \beta_2 \text{ speciesChinstrap} + \beta_3 \text{ speciesGentoo}$$

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

Our new test ("F-test"): Is at least one of these terms significant assuming all others are in the model?

Comparing models

$$H_0 : \beta_4 = \beta_5 = 0$$

$$H_A : \text{at least one } \beta_4, \beta_5 \neq 0$$

We can no longer use

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

We will use Analysis of Variance (ANOVA)

- **idea:** how much more explanatory power does the bigger model H_A have compared to the smaller model H_0 ?

ANOVA for MLR (basically same as SLR!)

$$SST = SS_{reg} + SSR$$

- SST: Total variation Total sum of squares

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2$$

- SSR: Unexplained variation Residual sum of squares

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (n - p - 1)\hat{\sigma}^2$$

- SSreg: Explained variation Regression sum of squares

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

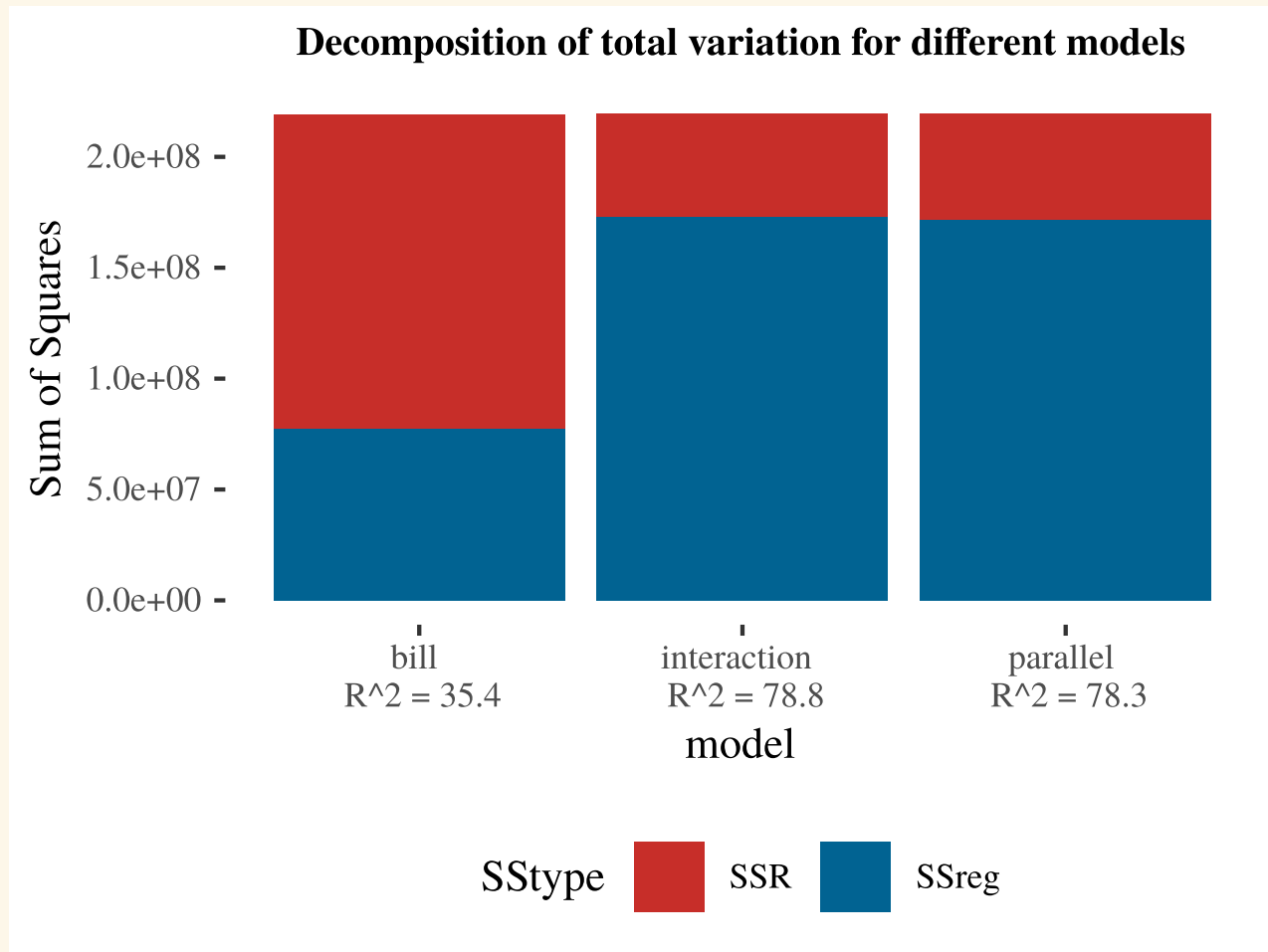
R-squared

R-squared (coefficient of determination) measures the proportion of variability observed in the response Y which can be explained by the regression of \mathbf{Y} on x_1, \dots, x_p .

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{SSR}{SST}$$

- About 35.4% of the variation in mass is explained by the regression of mass on bill length.
- About 78.3% of the variation in mass is explained by the regression of mass on bill length and species.
- About 78.8% of the variation in mass is explained by the regression of mass on bill length, species and their interaction.

Penguins: ANOVA for 3 models



$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{SSR}{SST}$$

- SS_{reg} (explained) is largest for the biggest model
- SSR (unexplained) is smallest for the biggest model

ANOVA for MLR

$$SST = SS_{reg} + SSR$$

- **Mathematically:** bigger models have bigger SS_{reg} :

$$SS_{reg}(x_1, x_2, x_1 : x_2) > SS_{reg}(x_1, x_2) > SS_{reg}(x_1)$$

- **Mathematically:** bigger models have smaller SSR :

$$SSR(x_1, x_2, x_1 : x_2) < SSR(x_1, x_2) < SSR(x_1)$$

- Bigger models have more explanatory power (higher SS_{reg}) and larger R^2 than a smaller model
- Assume the smaller model is a nested version of the larger
- But is the bigger model "worth it" if it is more complex and only has slightly more SS_{reg} explained?

ANOVA for MLR

Mean square values divide SS by their degrees of freedom

- Degrees of freedom (basically) measure how much stuff we've added up in the SS
- **MST:** Average total variation Mean Square of Total

$$MST = \frac{SST}{n - 1} = s_y^2 = \text{sample variance of Y}$$

- **MSR:** Average unexplained variation residual mean square

$$MSR = \frac{SSR}{n - (p + 1)} = \hat{\sigma}^2 = \text{estimated model variance}$$

- **MSreg:** Explained variation Regression mean square

$$MSreg = \frac{SSreg}{p}$$

ANOVA for MLR

$$SST = SSreg + SSR$$

- **Mathematically:** mean square residual and regression values won't equal the total

$$MST \neq MSreg + MSR$$

- **Mathematically:** mean square regression values won't be ordered based on their size

$$MSreg(x_1, x_2, x_1 : x_2)? MSreg(x_1, x_2)? MSreg(x_1)$$

- **Mathematically:** mean square residual values won't be ordered based on their size

$$MSR(x_1, x_2, x_1 : x_2)? MSR(x_1, x_2)? MSR(x_1)$$

Adjusted R-squared

Adjusted R^2 is similar to R^2 , but it uses mean square values to account for the size of a model.

$$R_a^2 = R_{\text{adjust}}^2 = 1 - \frac{MSR}{MST}$$

- R_a^2 could actually increase if unneeded predictors are removed from a model.
- R^2 would never increase when predictors are removed.

Penguins: Adjusted R^2 for 3 models

```
penguins <- penguins %>% tidyr::drop_na(bill_length_mm, body_mass_g, species) # drop missing values
peng_bill_lm <- lm(body_mass_g ~ bill_length_mm, data = penguins) # simple linear regression model
peng_nointeraction_lm <- lm(body_mass_g ~ bill_length_mm + species, data = penguins) # parallel lines model
peng_interaction_lm <- lm(body_mass_g ~ bill_length_mm*species, data = penguins) # interaction model
```

```
get_regression_summaries(peng_bill_lm) %>%
  select(r_squared, adj_r_squared) # just want these two
# A tibble: 1 × 2
  r_squared adj_r_squared
    <dbl>      <dbl>
1    0.354      0.352
```

```
get_regression_summaries(peng_nointeraction_lm) %>%
  select(r_squared, adj_r_squared)
# A tibble: 1 × 2
  r_squared adj_r_squared
    <dbl>      <dbl>
1    0.783      0.781
```

```
get_regression_summaries(peng_interaction_lm) %>%
  select(r_squared, adj_r_squared)
# A tibble: 1 × 2
  r_squared adj_r_squared
    <dbl>      <dbl>
1    0.788      0.785
```

glimpse(penguins)

```
Rows: 342
Columns: 8
$ species      <fct> Adelie, Adelie,
$ island       <fct> Torgersen, Torge
$ bill_length_mm <dbl> 39.1, 39.5, 40.3
$ bill_depth_mm <dbl> 18.7, 17.4, 18.6
$ flipper_length_mm <int> 181, 186, 195, 1
$ body_mass_g   <int> 3750, 3800, 3250
$ sex          <fct> male, female, fe
$ year         <int> 2007, 2007, 2007
```

Penguins: Adjusted R^2 for 3 models

- Removing the interaction terms results in a small drop in adjusted R^2 (0.785 to 0.781).
- The explanatory power of both the larger interaction model and smaller no-interaction models looks similar.
- **Next:** let's verify this observation with an ANOVA F-test

Comparing models: ANOVA F-test

- **Hypotheses:**

$$H_0 : \mu_{Y|x} = \text{smaller, nested model}$$

$$H_A : \mu_{Y|x} = \text{larger model}$$

where the nested model is arrived at by setting one or more β_j in the larger model equal to 0

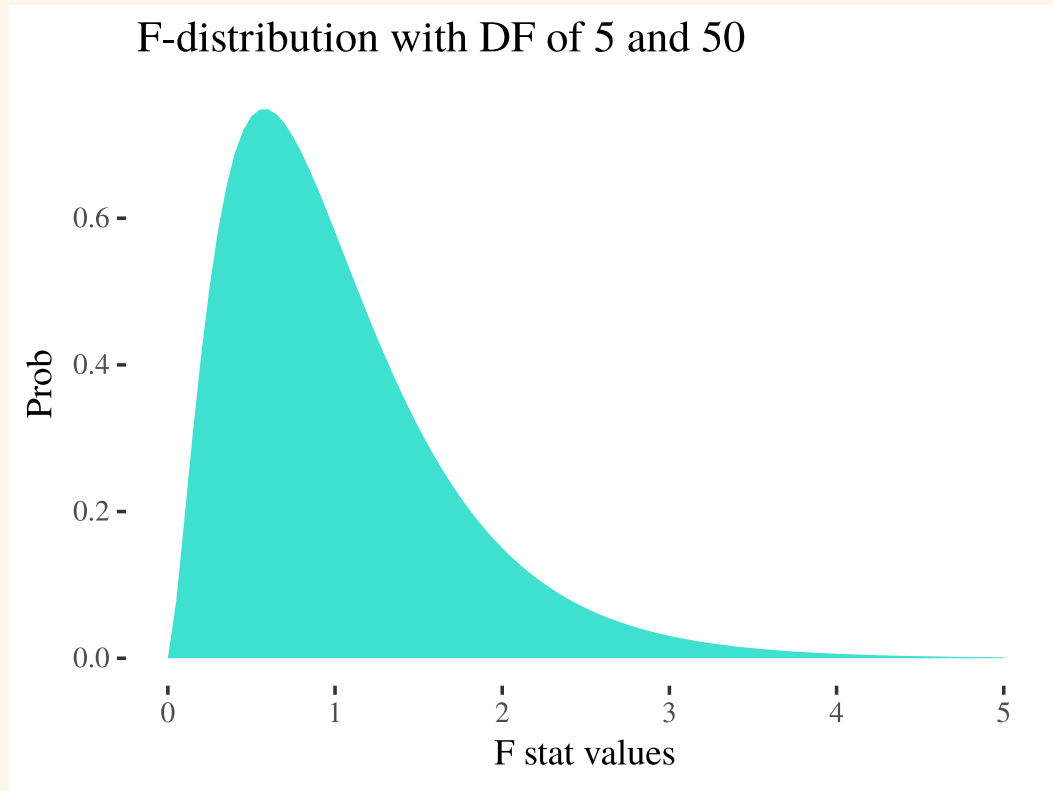
- **Test stat:**

$$F = \frac{\text{extraSS} / (\# \text{ terms tested})}{MSR_{\text{larger}}}$$

where extra SS is the difference in explained SS between the larger H_A and smaller H_0 models

$$\begin{aligned} \text{extraSS} &= SS_{\text{reg}}(\text{larger model}) - SS_{\text{reg}}(\text{smaller model}) \\ &= SSR(\text{smaller model}) - SSR(\text{larger model}) \end{aligned}$$

Comparing models: ANOVA F-test



- **P-value:** Probability of getting an T test stat larger than one observed

$$\text{p-value} = P(F > f_{\text{observed}})$$

using the F-distribution with degrees of freedom equal to the # of terms tested (numerator) and $n - (p + 1)$ from the larger H_A model (denominator)

ANOVA F-test in R

R-code

```
anova(smaller_lm, larger_lm)
```

Let's compare the interaction and parallel lines models in penguins dataset:

$$H_0 : \mu_{\text{mass} | x} = \beta_0 + \beta_1 \text{ bill} + \beta_2 \text{ speciesChinstrap} + \beta_3 \text{ speciesGentoo}$$

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

- Same as testing β_4 and β_5 in the larger model

$$H_0 : \beta_4 = \beta_5 = 0$$

$$H_A : \text{at least one } \beta_4, \beta_5 \neq 0$$

ANOVA F-test in R

```
anova(peng_nointeraction_lm , peng_interaction_lm)
```

Analysis of Variance Table

Model 1: `body_mass_g ~ bill_length_mm + species`

Model 2: `body_mass_g ~ bill_length_mm * species`

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	338	47613707				
2	336	46447006	2	1166702	4.22	0.01549 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ANOVA F-test in R

```
anova(peng_nointeraction_lm , peng_interaction_lm)
```

Analysis of Variance Table

Model 1: `body_mass_g ~ bill_length_mm + species`

Model 2: `body_mass_g ~ bill_length_mm * species`

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	338	47613707				
2	336	46447006	2	1166702	4.22	0.01549 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Res.Df column

- model DF for smaller = $342 - 4 = 338$
- model DF for larger = $342 - 6 = 336$

RSS Df column

- RSS(smaller)= 47,613,707
- RSS(larger)= 46,447,006

ANOVA F-test in R

```
anova(peng_nointeraction_lm , peng_interaction_lm)
```

Analysis of Variance Table

```
Model 1: body_mass_g ~ bill_length_mm + species
```

```
Model 2: body_mass_g ~ bill_length_mm * species
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	338	47613707				
2	336	46447006	2	1166702	4.22	0.01549 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Df column

- number of terms tested
 $= 6 - 4 = 338 - 336 = 2$

Sum of Sq column

- Extra SS for this test = $\text{RSS}(\text{smaller}) - \text{RSS}(\text{larger}) = 1,166,702$

ANOVA F-test in R

```
anova(peng_nointeraction_lm , peng_interaction_lm)
```

Analysis of Variance Table

Model 1: body_mass_g ~ bill_length_mm + species

Model 2: body_mass_g ~ bill_length_mm * species

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	338	47613707				
2	336	46447006	2	1166702	4.22	0.01549 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

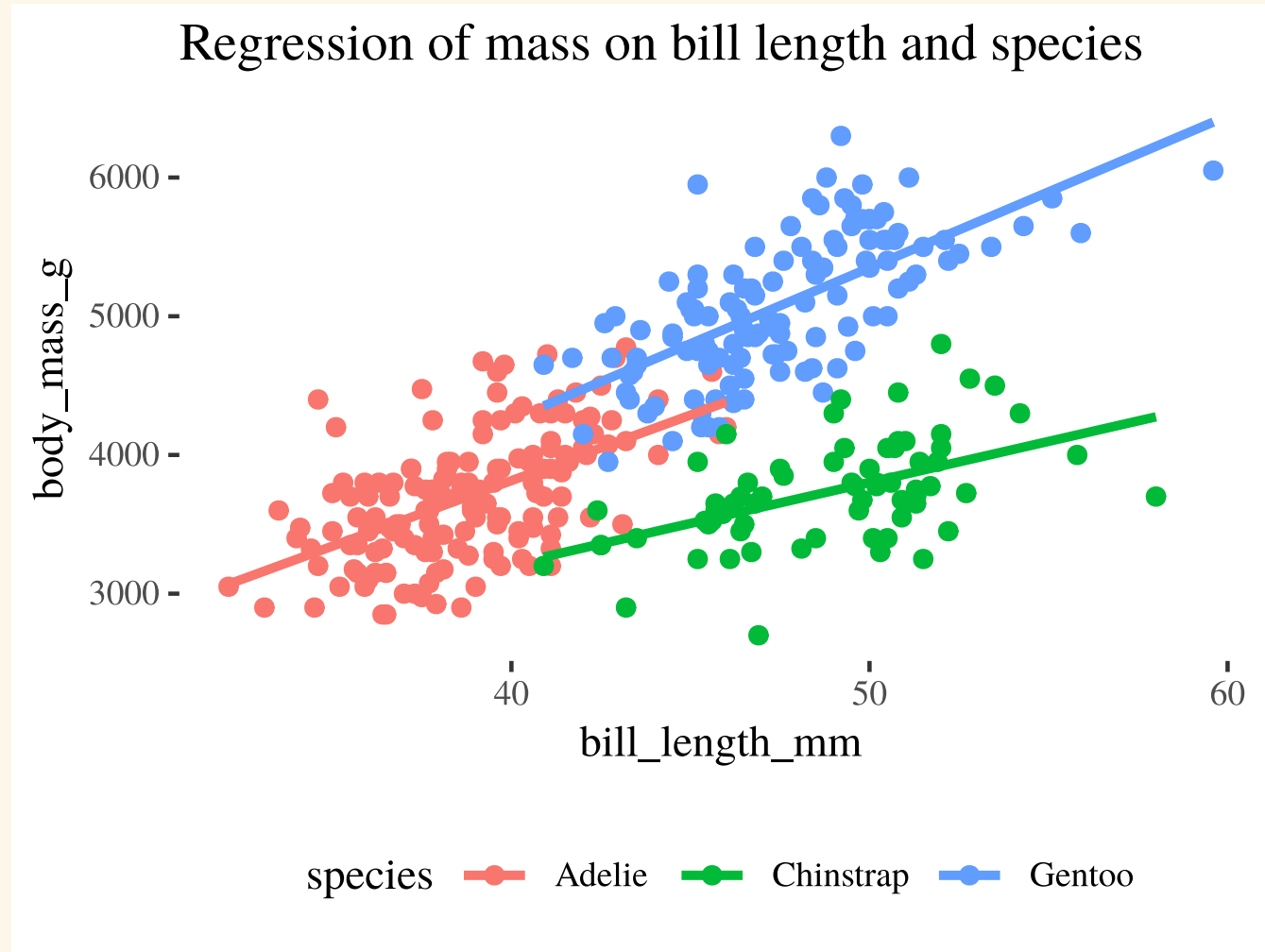
F column

- **F** test stat = $F = \frac{1,166,702/2}{46,447,006/336} = 4.22$

Pr(> F) column

- p-value for this test
= $P(F_{2,336} > 4.22) = 1 - \text{pf}(4.22, 2, 336)$
= 0.01549

Penguins wrapup



The slopes are different enough to warrant an interaction model

Penguins wrapup

```
summary(peng_interaction_lm)$coefficients # simple way of getting summaries
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.88299	443.17604	0.07871136	9.373091e-01
bill_length_mm	94.49982	11.39794	8.29095491	2.733899e-15
speciesChinstrap	811.26034	799.80552	1.01432201	3.111590e-01
speciesGentoo	-158.71092	683.19141	-0.23230813	8.164401e-01
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373192	4.698940e-02
bill_length_mm:speciesGentoo	14.95935	15.78642	0.94760841	3.440099e-01

Penguins wrapup

```
summary(peng_interaction_lm)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.88299	443.17604	0.07871136	9.373091e-01
bill_length_mm	94.49982	11.39794	8.29095491	2.733899e-15
speciesChinstrap	811.26034	799.80552	1.01432201	3.111590e-01
speciesGentoo	-158.71092	683.19141	-0.23230813	8.164401e-01
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373192	4.698940e-02
bill_length_mm:speciesGentoo	14.95935	15.78642	0.94760841	3.440099e-01

- The difference in slopes (effect of bill length on mass) between Adelie (baseline) and Gentoo isn't statistically significance ($t = 0.948$, $df = 336$, $p = 0.344$)

Penguins wrapup

```
summary(peng_interaction_lm)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.88299	443.17604	0.07871136	9.373091e-01
bill_length_mm	94.49982	11.39794	8.29095491	2.733899e-15
speciesChinstrap	811.26034	799.80552	1.01432201	3.111590e-01
speciesGentoo	-158.71092	683.19141	-0.23230813	8.164401e-01
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373192	4.698940e-02
bill_length_mm:speciesGentoo	14.95935	15.78642	0.94760841	3.440099e-01

- The difference in slopes (effect of bill length on mass) between Adelie (baseline) and Chinstrap is weakly statistically significance ($t = -1.994$, $df = 336$, $p = 0.047$).

Penguins wrapup

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.88299	443.17604	0.07871136	9.373091e-01
bill_length_mm	94.49982	11.39794	8.29095491	2.733899e-15
speciesChinstrap	811.26034	799.80552	1.01432201	3.111590e-01
speciesGentoo	-158.71092	683.19141	-0.23230813	8.164401e-01
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373192	4.698940e-02
bill_length_mm:speciesGentoo	14.95935	15.78642	0.94760841	3.440099e-01

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

- To compare Chinstrap and Gentoo, either use the test

$$H_0 : \beta_1 + \beta_4 = \beta_1 + \beta_5 \Rightarrow \beta_4 - \beta_5 = 0$$

Penguins wrapup

- OR! Relevel the species variable to make Chinstrap the baseline - it's slope seem most different (flatter) compared to the other two species

$$\begin{aligned}\mu_{\text{mass} | x} = & \beta_0^* + \beta_1^* \text{bill} + \beta_2^* \text{speciesAdelie} + \beta_3^* \text{speciesGentoo} \\ & + \beta_4^* \text{bill} \times \text{speciesAdelie} + \beta_5^* \text{bill} \times \text{speciesGentoo}\end{aligned}$$

- Same variables but changing up the indicator variables changes our parameter interpretation
- that's why we used β^* here (not the same values as β in the previous model)
- but won't change an inference conclusions or effect estimation!!

Penguins wrapup

- OR! Relevel the species variable to make Chinstrap the baseline
- Use `fct_relevel()` from the `forcats` package:

```
library(forcats)
# just need to specify the first level
penguins$new_species <- fct_relevel(penguins$species, "Chinstrap")
```

Check order using a quick table command:

```
table(penguins$species)
```

Adelie	Chinstrap	Gentoo
151	68	123

```
table(penguins$new_species)
```

Chinstrap	Adelie	Gentoo
68	151	123

Penguins wrapup

```
peng_interaction_lm_CS <- lm(body_mass_g ~ bill_length_mm * new_species, data = penguins)
summary(peng_interaction_lm_CS)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	846.14333	665.79566	1.270875	2.046521e-01
bill_length_mm	59.11773	13.60261	4.346059	1.838939e-05
new_speciesAdelie	-811.26034	799.80552	-1.014322	3.111590e-01
new_speciesGentoo	-969.97126	844.76586	-1.148213	2.516974e-01
bill_length_mm:new_speciesAdelie	35.38208	17.74666	1.993732	4.698940e-02
bill_length_mm:new_speciesGentoo	50.34143	17.44503	2.885718	4.157658e-03

Penguins wrapup

```
peng_interaction_lm_CS <- lm(body_mass_g ~ bill_length_mm * new_species, data = penguins)
summary(peng_interaction_lm_CS)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	846.14333	665.79566	1.270875	2.046521e-01
bill_length_mm	59.11773	13.60261	4.346059	1.838939e-05
new_speciesAdelie	-811.26034	799.80552	-1.014322	3.111590e-01
new_speciesGentoo	-969.97126	844.76586	-1.148213	2.516974e-01
bill_length_mm:new_speciesAdelie	35.38208	17.74666	1.993732	4.698940e-02
bill_length_mm:new_speciesGentoo	50.34143	17.44503	2.885718	4.157658e-03

- The difference in slopes (effect of bill length on mass) between Chinstrap (baseline) and Adelie is weakly statistically significance ($t = 1.994$, $df = 336$, $p = 0.047$).

Penguins wrapup

```
summary(peng_interaction_lm_CS)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	846.14333	665.79566	1.270875	2.046521e-01
bill_length_mm	59.11773	13.60261	4.346059	1.838939e-05
new_speciesAdelie	-811.26034	799.80552	-1.014322	3.111590e-01
new_speciesGentoo	-969.97126	844.76586	-1.148213	2.516974e-01
bill_length_mm:new_speciesAdelie	35.38208	17.74666	1.993732	4.698940e-02
bill_length_mm:new_speciesGentoo	50.34143	17.44503	2.885718	4.157658e-03

- The difference in slopes (effect of bill length on mass) between Chinstrap (baseline) and Gentoo is statistically significant ($t = 2.886$, $df = 336$, $p = 0.00416$).