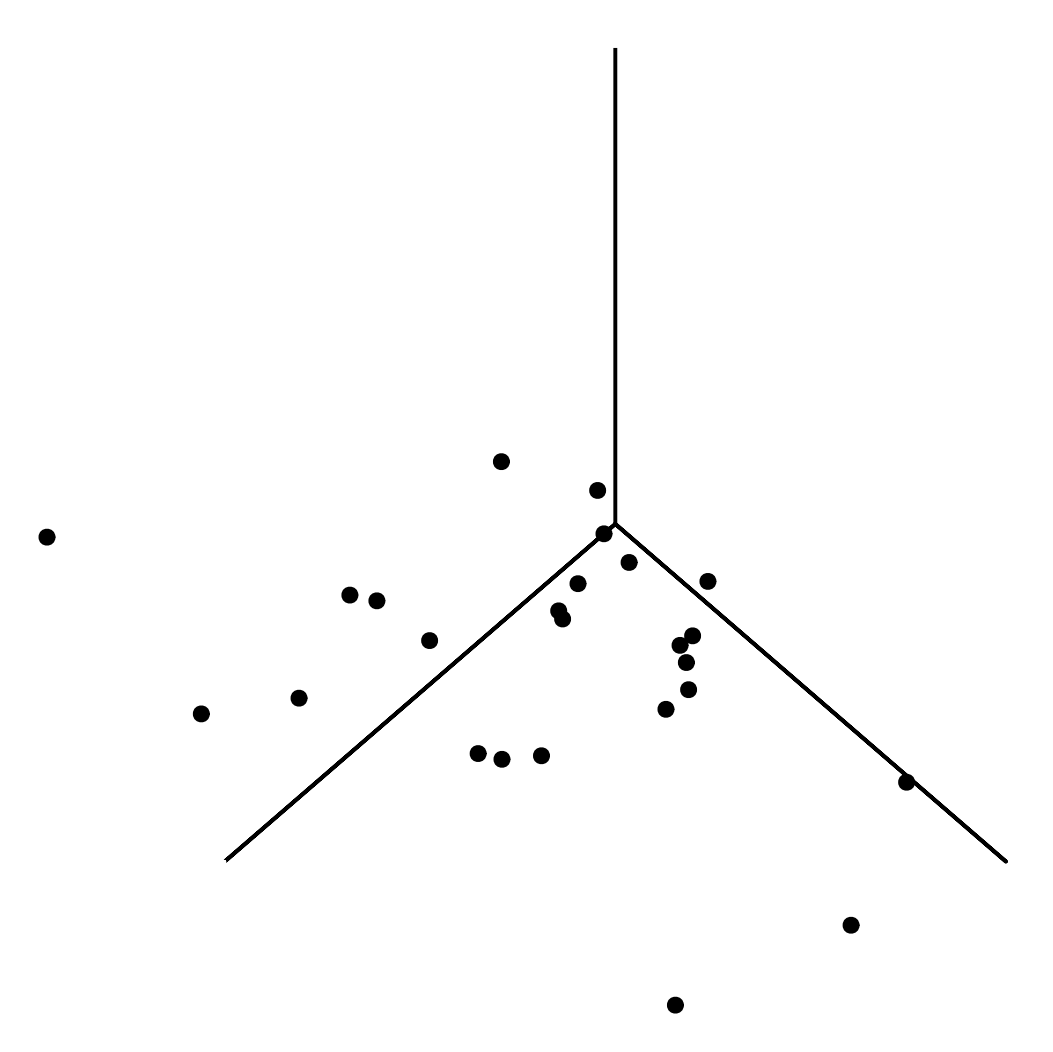
MLR: More ANOVA and Diagnostics using residuals

Stat 230

April 25 2022

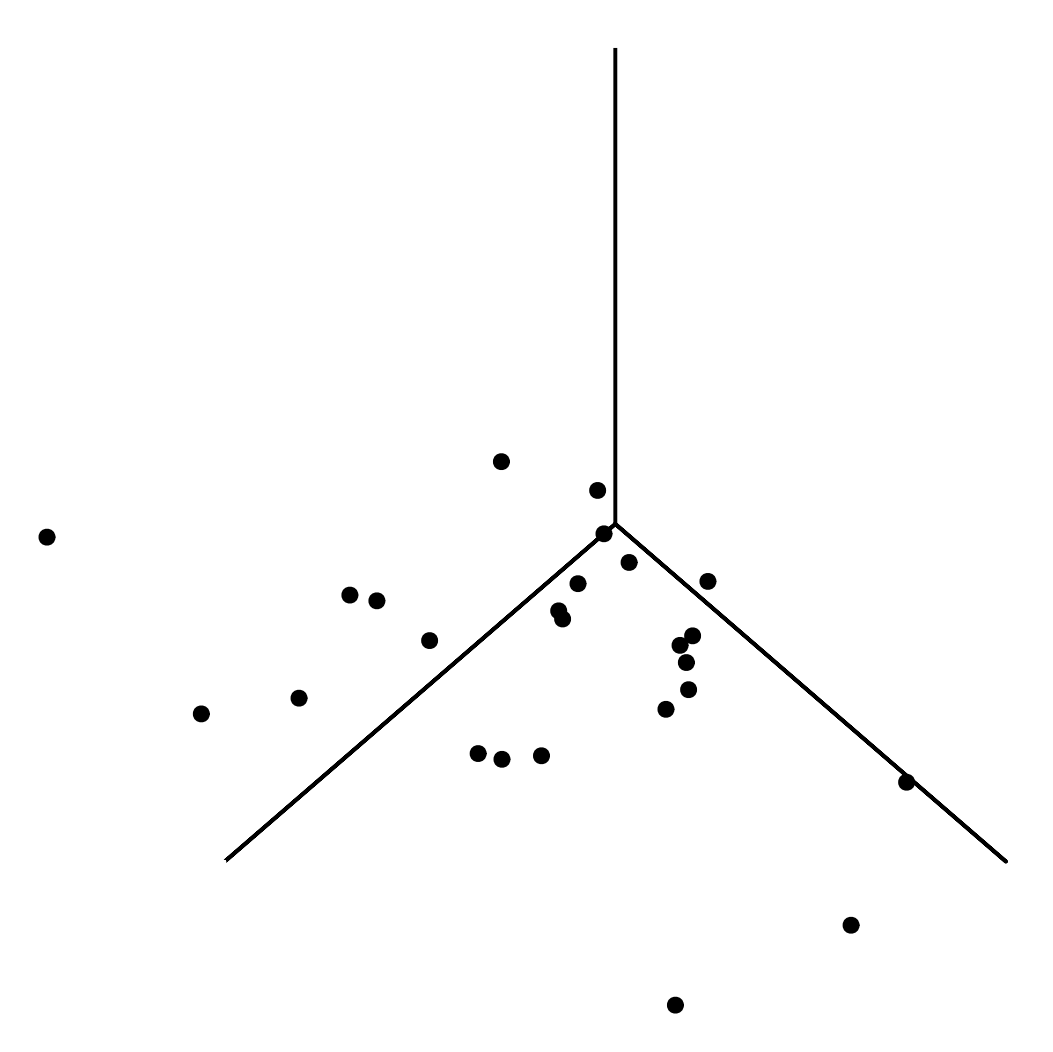
Overview



Today:

- Some additional topics on ANOVA
- Assessing Model Assumptions

Overview



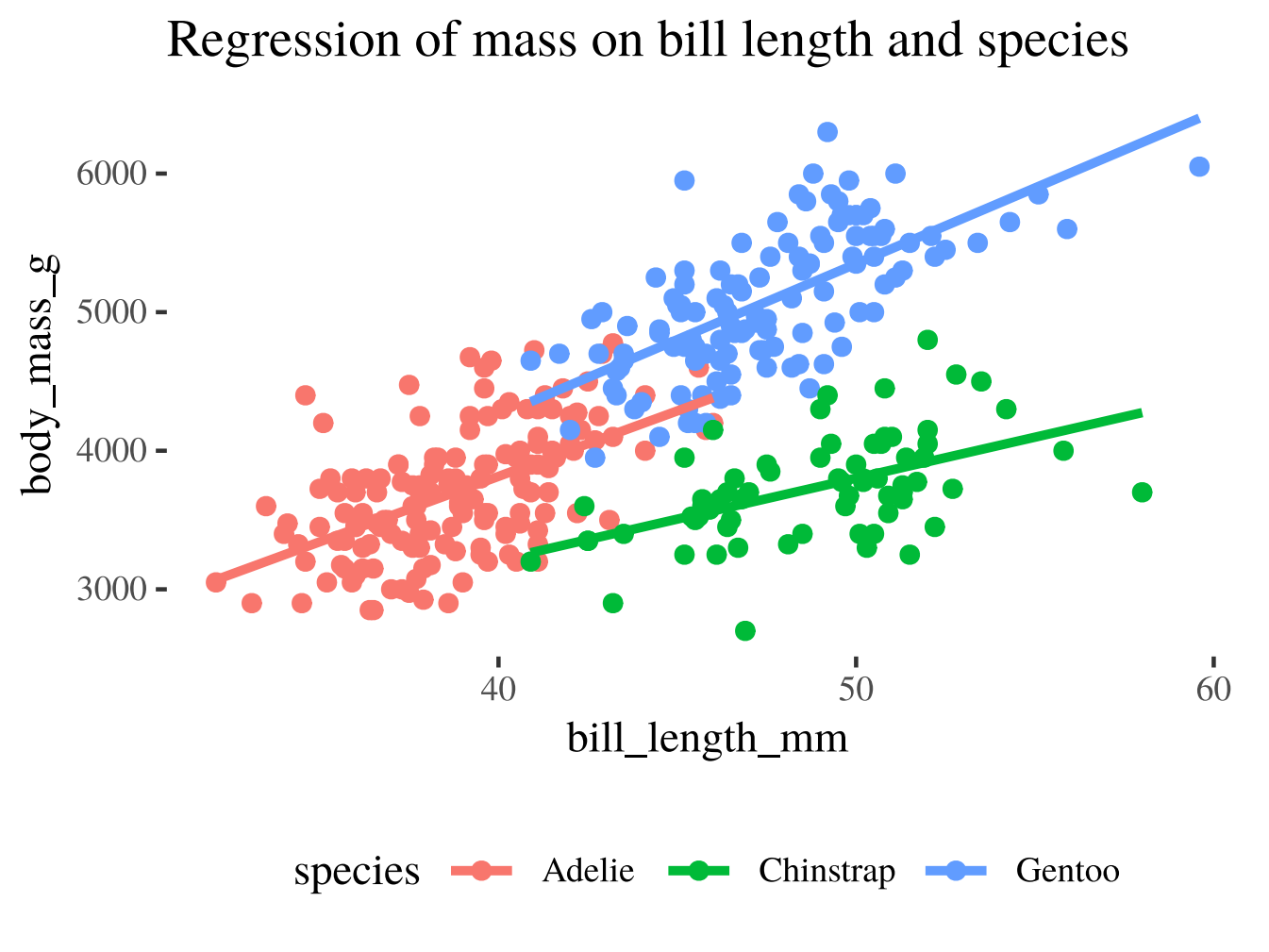
Today:

- Some additional topics on ANOVA
- Assessing Model Assumptions

Important Comment:

- The class lecture notes use p to represent the number of terms in your regression model (# betas minus the intercept).
- Your textbook uses p to denote the number of betas in your model. The textbook p will always be 1 greater than lecture notes p.

Penguins Example



One-term F-test

$$H_0 : \mu_{Y|x} = \beta_0 + 0 + \beta_2 x_2 + \cdots + \beta_p x_p \Rightarrow \beta_1 = 0$$
$$H_A : \mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \Rightarrow \beta_1 \neq 0$$

Question being asked: Is the effect of x_1 on μ_y statistically significant holding all other predictors $x_2, \ldots, x_p$ fixed?

- We can test "one term" with either a t-test or F-test.
- t-test is "easier" since we get these results from `summary` or `tidy` without having to fit a reduced model!

Overall F-test

$$H_0 : \mu_{Y|x} = \beta_0$$
$$H_A : \mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Question being asked: Are any of our predictors effects statistically significant??

```
summary(peng_interaction_lm)

... output omitted to save space! ....

Residual standard error: 371.8 on 336 degrees of freedom
Multiple R-squared:  0.7882,    Adjusted R-squared:  0.7851
F-statistic: 250.1 on 5 and 336 DF,  p-value: < 2.2e-16
```

At least one term in the interaction model has a statistically significant effect on mass (F = 250.1, df = 5, 336, p < 0.0001)

More on ANOVA

```
# R-code
anova(my_model)
```

anova on one lm model gives a table of extra SS for sequentially increasing terms in a model

```
anova(peng_interaction_lm)
Analysis of Variance Table

Response: body_mass_g
                      Df   Sum Sq  Mean Sq F value  Pr(>F)
bill_length_mm         1 77669072 77669072  561.86 < 2e-16 ***
species                2 94024918 47012459  340.09 < 2e-16 ***
bill_length_mm:species   2  1166702   583351    4.22 0.01549 *
Residuals              336 46447006   138235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

More on ANOVA

```
Analysis of Variance Table

Response: body_mass_g
                      Df   Sum Sq  Mean Sq F value  Pr(>F)
bill_length_mm         1 77669072 77669072  561.86 < 2e-16 ***
species                2 94024918 47012459  340.09 < 2e-16 ***
bill_length_mm:species 2  1166702   583351    4.22 0.01549 *
Residuals            336 46447006   138235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum Sq column:

$$SSreg(\text{ Bill length }) = 77,669,072$$

Df column: including this variable adds 1 term to the model

F value, $\Pr(> F)$ **columns:** ignore!! (not a valid F test)

More on ANOVA

```
Analysis of Variance Table

Response: body_mass_g
                      Df   Sum Sq  Mean Sq F value  Pr(>F)
bill_length_mm         1 77669072 77669072  561.86 < 2e-16 ***
species                2 94024918 47012459  340.09 < 2e-16 ***
bill_length_mm:species   2  1166702   583351    4.22 0.01549 *
Residuals              336 46447006   138235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum Sq column: extra SS for adding species to a model already containing bill_length_mm

$$\text{SSreg}(\text{Bill length, species}) - \text{SSreg}(\text{Bill length}) = 94,024,918$$

Df column: including this variable adds 2 terms to the model

F value, $\Pr(> F)$ **columns:** ignore!! (not a valid F test)

More on ANOVA

```
Analysis of Variance Table

Response: body_mass_g
                     Df   Sum Sq  Mean Sq F value  Pr(>F)
bill_length_mm        1 77669072 77669072  561.86 < 2e-16 ***
species               2 94024918 47012459  340.09 < 2e-16 ***
bill_length_mm:species  2  1166702   583351    4.22 0.01549 *
Residuals           336 46447006   138235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum Sq column: extra SS for adding bill_length_mm: species to a model already containing bill_length_mm and species

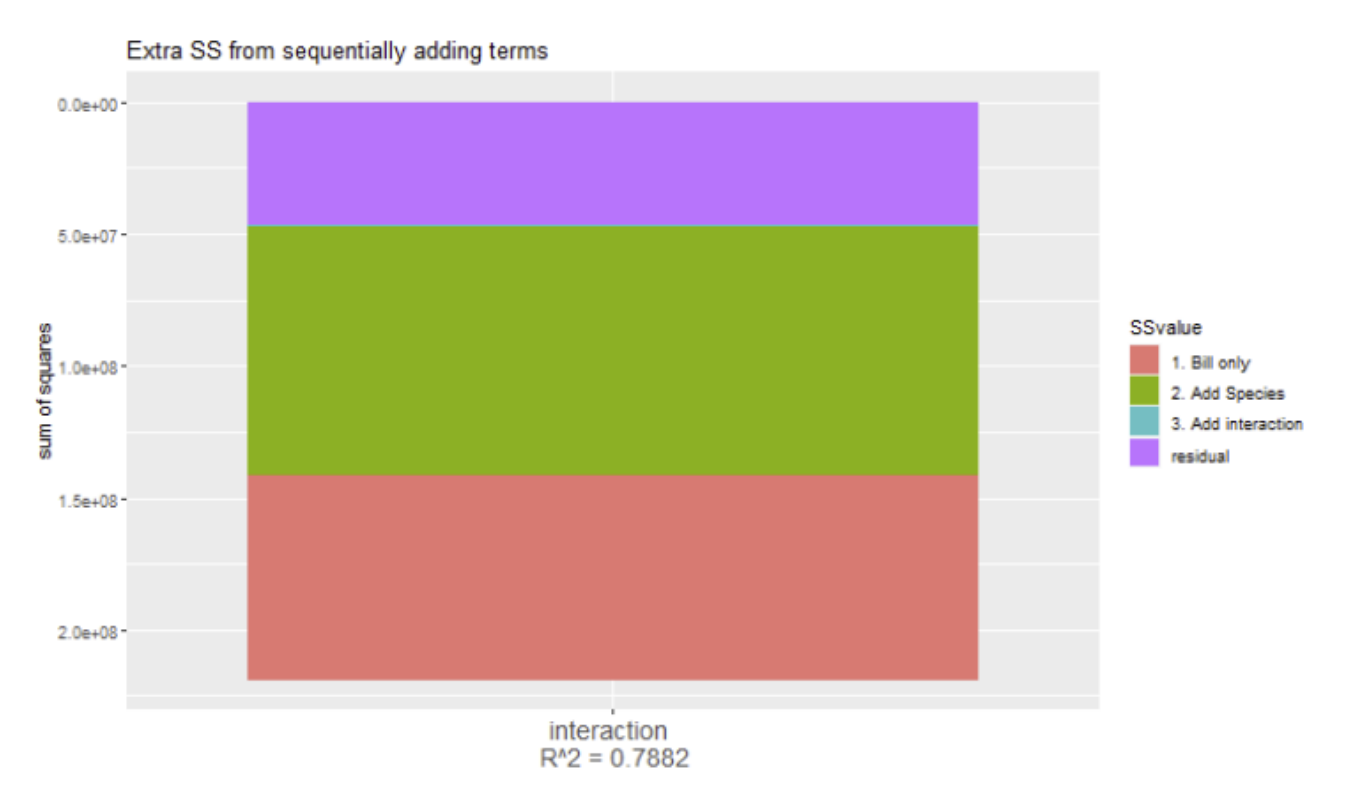
$$\text{SSreg}(\text{Bill length, species, interaction}) - \text{SSreg}(\text{Bill length, species}) = 1,166,702$$

Df column: including this variable adds 2 terms to the model

F value, $\Pr(> F)$ **columns:** The last row in this table is a valid F-test comparing the model with no interaction (null) to the interaction model (alt).

More on ANOVA

Here we can see the sequential extra SS "add up" to the overall SSreg for the interaction model.



Let's talk about disgnostics of MLR with focus on residuals

Multiple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots \beta_p x_{p,i} + \epsilon_i \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma)$$

- i.i.d = independent and identically distributed
- These model errors ϵ_i 's are independent of x !
- We only need to check that the errors are i.i.d. $N(0, \sigma)$.
- We don't actually see the model errors ϵ_i but we can inspect their equivalent, r_i the residuals, from the fitted regression model.

Diagnostic Tools for model checking

Linearity

- Residual plot(s): Plot residuals $r_i = y_i - \hat{y}_i$ against the fitted values $\hat{y}_i$ and all explanatory variables

Equal/Constant variance a.k.a. homoscedastic

- Residual plot(s)
- (Non-constant variance test)

Normality

- Normal QQ plot of residuals

Independence

- Plot residuals vs. time (temporal association) or explore spatial association within residuals.

What if assumption is violated? (same as SLR)

What if model assumptions are violated? Possible Solutions:

- **Linearity, Variance, Normality:** Transform one or both variables
- **Linearity:** change mean function, use non-linear regression
- **Variance:** weighted regression, "robust SEs"
- **Independence:** use time variable in model, or use time series or spatial regression model, or random effects (mixed model) for correlated data

How “robust” is regression against violation of the assumptions? (same as SLR)

- **Robust:** can violate and still get valid inference results
- **Normality:** the t-tests and CI for model parameters and the mean response are saved by the Central Limit Theorem when n is large, even if your subpopulation of responses are not normally distributed.
- **Not very robust:** can give misleading results if violated

How “robust” is regression against violation of the assumptions? (same as SLR)

- **Linearity:** if the mean function is wrong then your estimated effects, mean response, or predicted response will be biased!
- **Constant variance and independence:** if you are not correctly modeling your response variability, then your SEs will not be an accurate reflection of your actual uncertainty (meaning CIs/tests might be misleading)
- **Normality only when computing prediction intervals:** these intervals need the normal subpopulation assumption to hold

Residuals plot in R using `ggResidpanel`

`resid_xpanel(my_lm)`: all residuals vs. explanatory plots

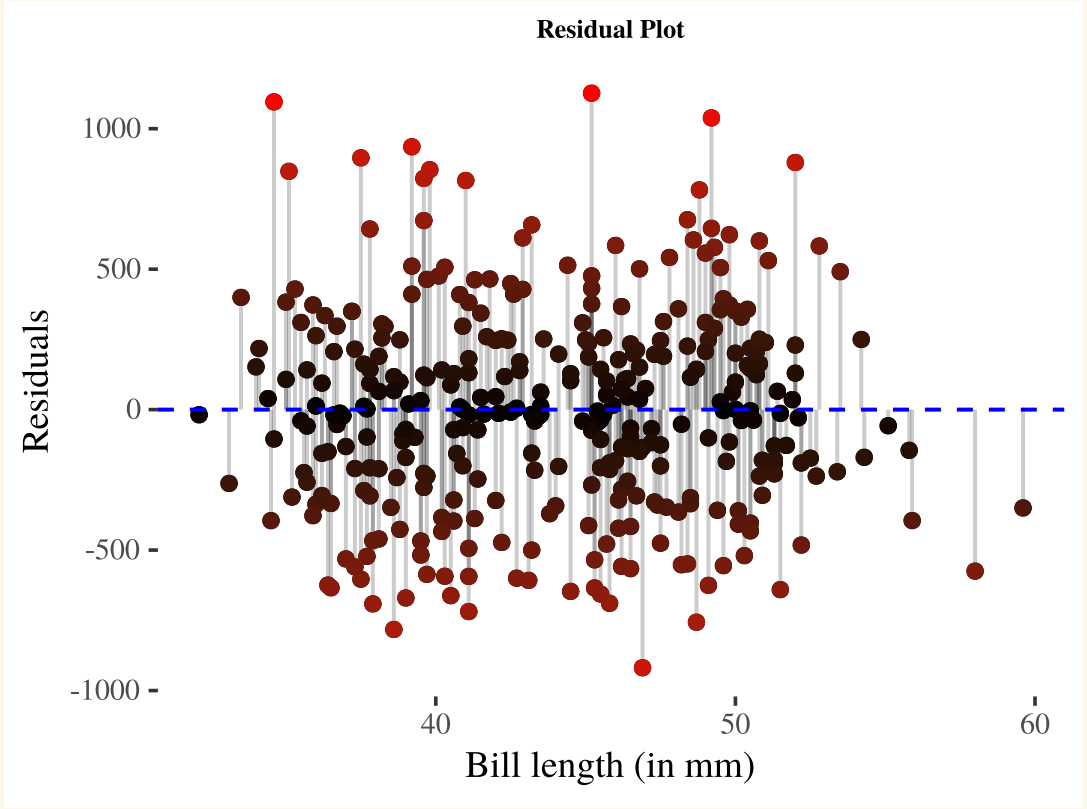
`resid_panel(my_lm)`:

- residual vs fitted (linearity, variance)
- QQ plot and histogram of residuals (normality)
- residual vs row number (not all that useful!)

`resid_panel(my_lm, plots = "resid")`: residual vs fitted

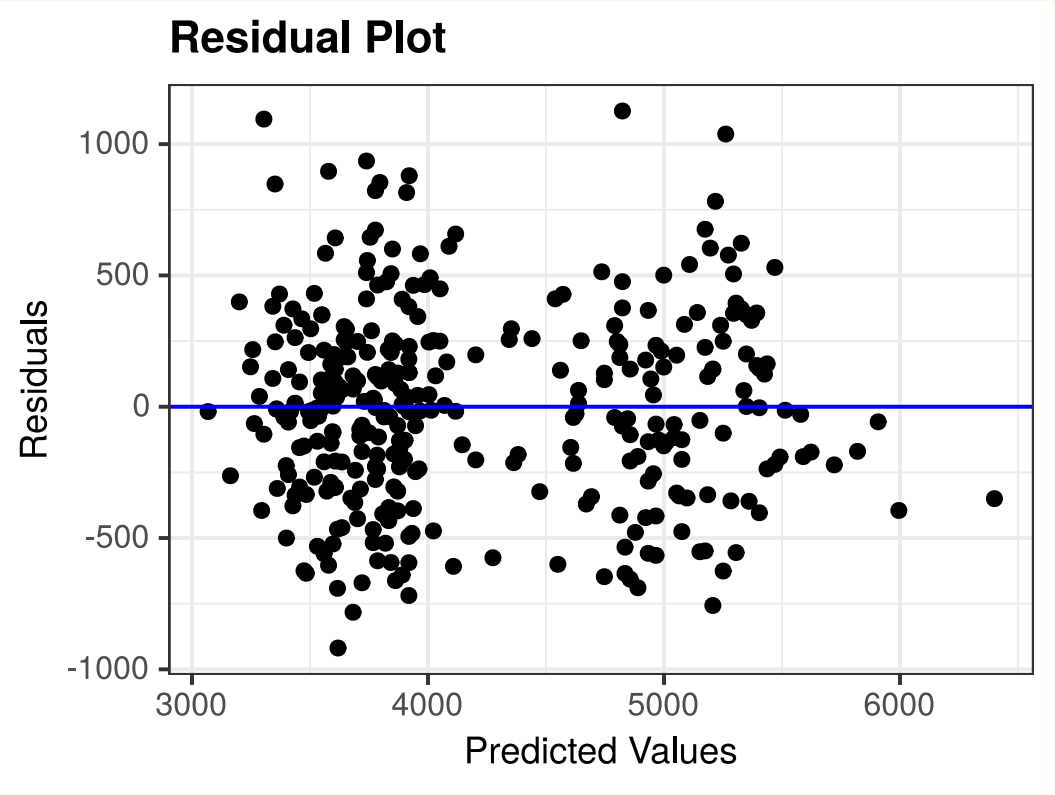
Penguins example

```
peng_interaction_lm <-  lm(body_mass_g ~ bill_length_mm*species, data = p
peng_regression_points <- get_regression_points(peng_interaction_lm)
ggplot(peng_regression_points, aes(x = bill_length_mm, y = residual)) +
 geom_point() +
  theme(legend.position = "none") +
  geom_segment(aes(xend = bill_length_mm, yend = 0), alpha = .2) +
  scale_color_continuous(low = "black", high = "red") +  # Colors to use
  geom_point(aes(color = abs(residual))) +
  geom_hline(yintercept = 0, col = "blue", size = 0.5, linetype = "dashed
   labs(x = "Bill length (in mm)",
        y = "Residuals",
        title = "Residual Plot") +
  theme(plot.title = element_text(hjust=0.5, size=7, face='bold'))
```



Penguins example

```
library(ggResidpanel)
peng_interaction_lm <-  lm(body_mass_g ~ bill_length_mm*species, data = penguins)
resid_panel(peng_interaction_lm, plots = "resid")
```

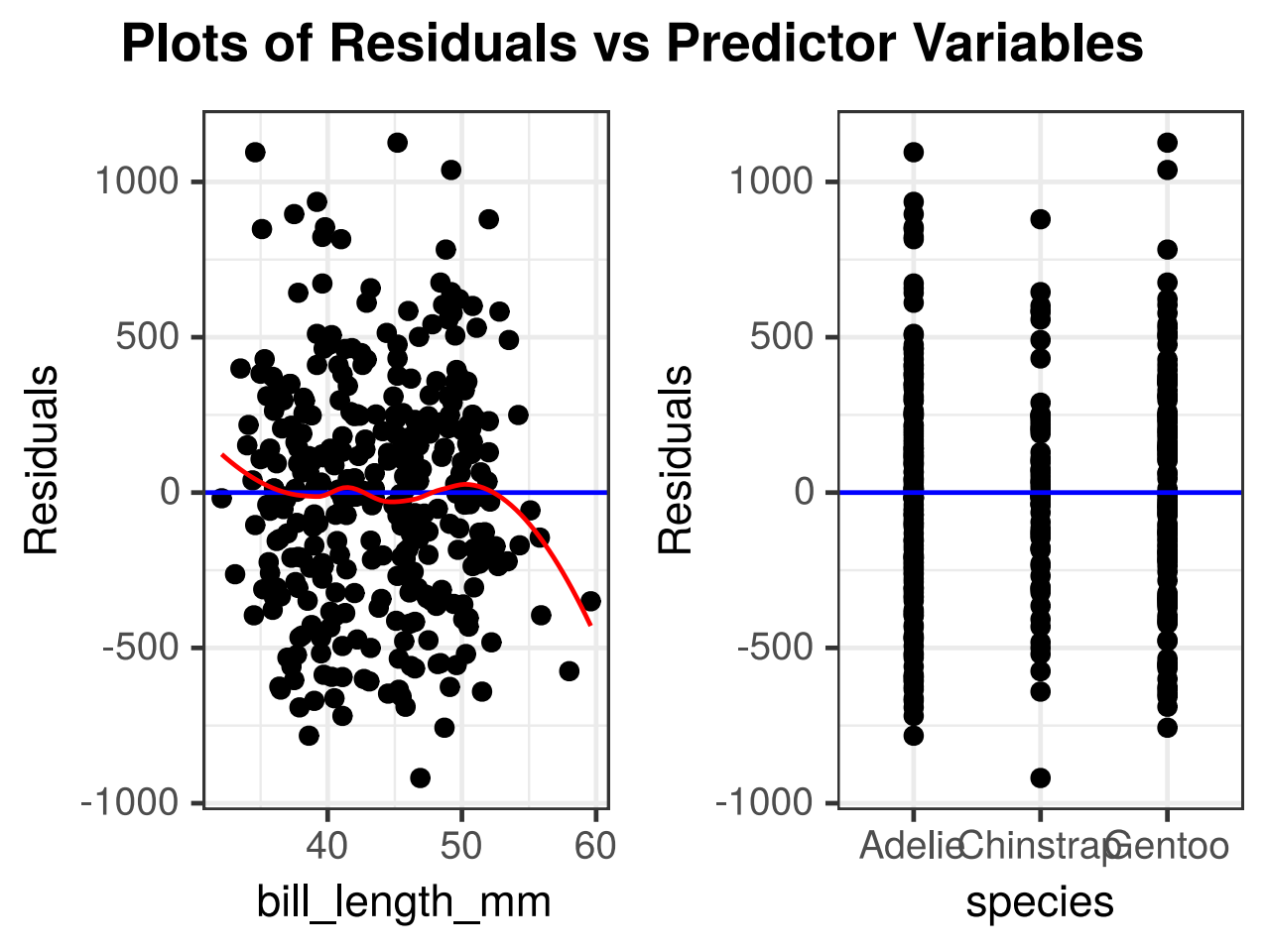


residuals vs fitted:

- **linearity:** seems ok (no systematic over/under estimation, expect maybe for a few large penguins)
- **constant variance:** seems ok, though the larger fitted group may have slightly less variation than the lower group

Penguins example

```
resid_xpanel(peng_interaction_lm, smoother = TRUE)
```



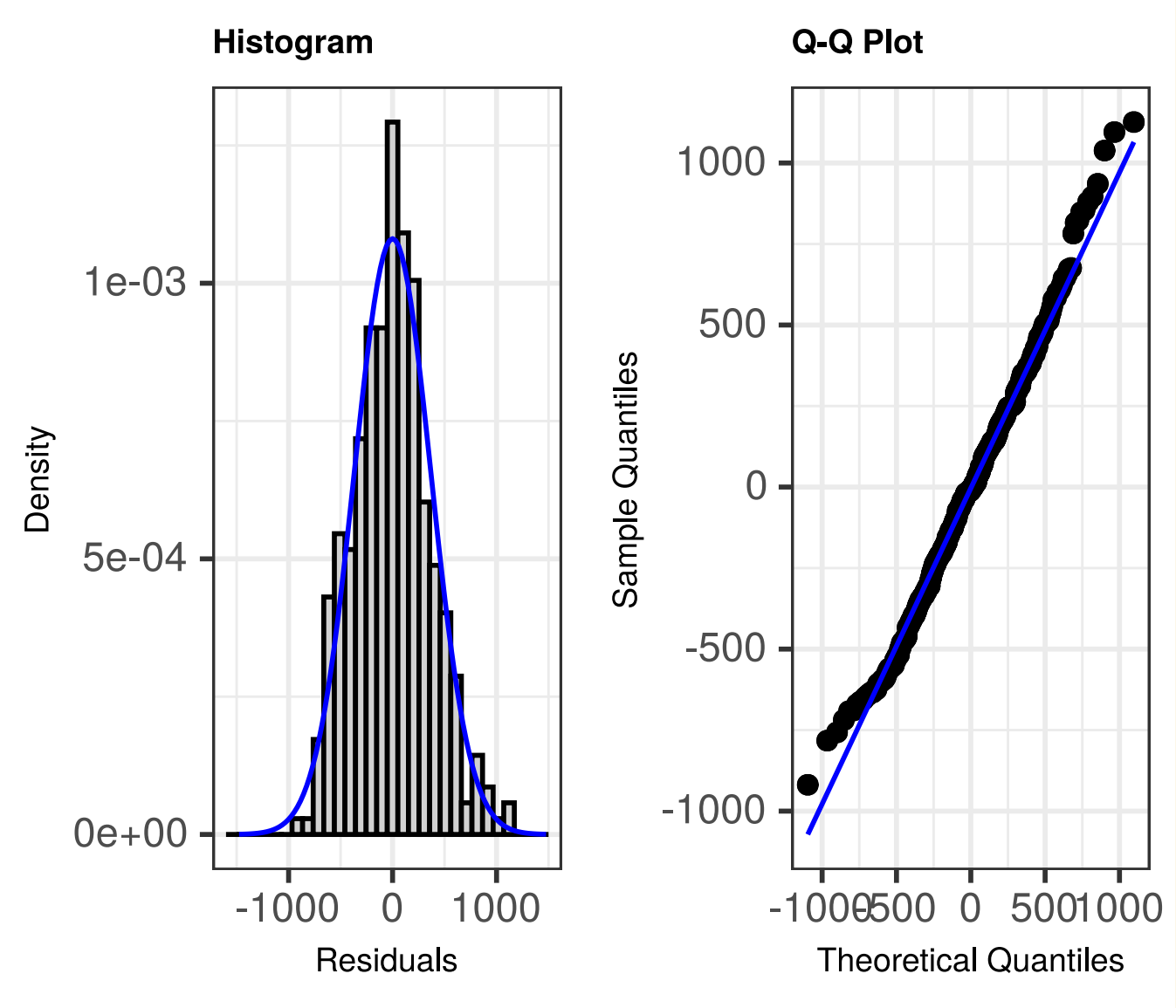
residuals vs bill:

- looks good

residuals vs. species

- roughly centered around 0 , variation is a bit less in Chinstrap

Penguins example



Penguins example

Top 5 largest residual cases

```
peng_regression_points %>%
  slice_max(residual, n= 5)
# A tibble: 5 × 6
     ID body_mass_g bill_length_mm species body_mass_g_hat residual
  <int>       <int>          <dbl> <fct>              <dbl>    <dbl>
1   231        5950           45.2 Gentoo             4824.    1126.
2    14        4400           34.6 Adelie             3305.    1095.
3   169        6300           49.2 Gentoo             5262.    1038.
4     7        4675           39.2 Adelie             3739.     936.
5   133        4475           37.5 Adelie             3579.     896.
```

Penguins example

Top 5 smallest (most negative) residual cases

```
peng_regression_points %>%
  slice_min(residual, n= 5)
# A tibble: 5 × 6
     ID body_mass_g bill_length_mm species   body_mass_g_hat residual
  <int>       <int>          <dbl> <fct>               <dbl>    <dbl>
1   313        2700           46.9 Chinstrap           3619.    -919.
2   116        2900           38.6 Adelie              3683.    -783.
3   154        4450           48.7 Gentoo              5207.    -757.
4    12        3200           41.1 Adelie              3919.    -719.
5   104        2925           37.9 Adelie              3616.    -691.
```

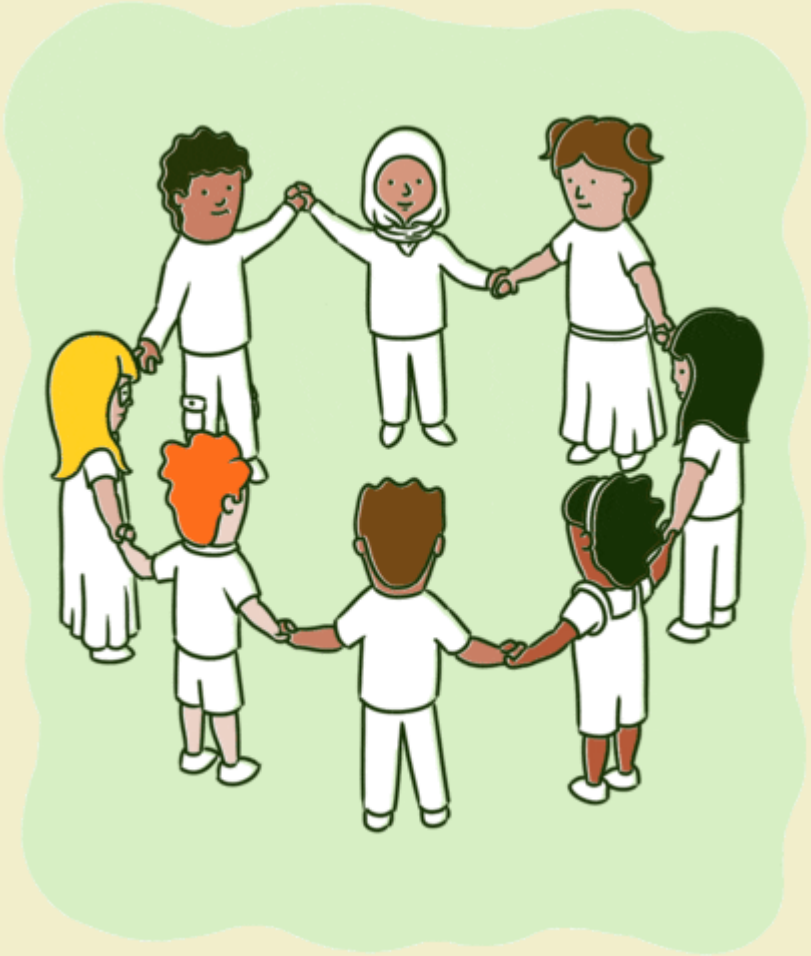
Penguins example

Top 5 largest predicted mass cases

```
peng_regression_points %>%
  slice_max(body_mass_g_hat, n= 5)
# A tibble: 5 × 6
     ID body_mass_g bill_length_mm species body_mass_g_hat residual
  <int>       <int>          <dbl> <fct>              <dbl>    <dbl>
1   185        6050           59.6 Gentoo             6400.   -350.
2   253        5600           55.9 Gentoo             5995.   -395.
3   267        5850           55.1 Gentoo             5907.    -57.4
4   215        5650           54.3 Gentoo             5820.   -170.
5   259        5500           53.4 Gentoo             5721.   -221.
```

Your Turn 1

05:00



- Get the in class activity file from moodle
- Skim through it with your group members