MA22004

Seminar 8

Dr Eric Hall

26/11/2020

Announcements

Reminders

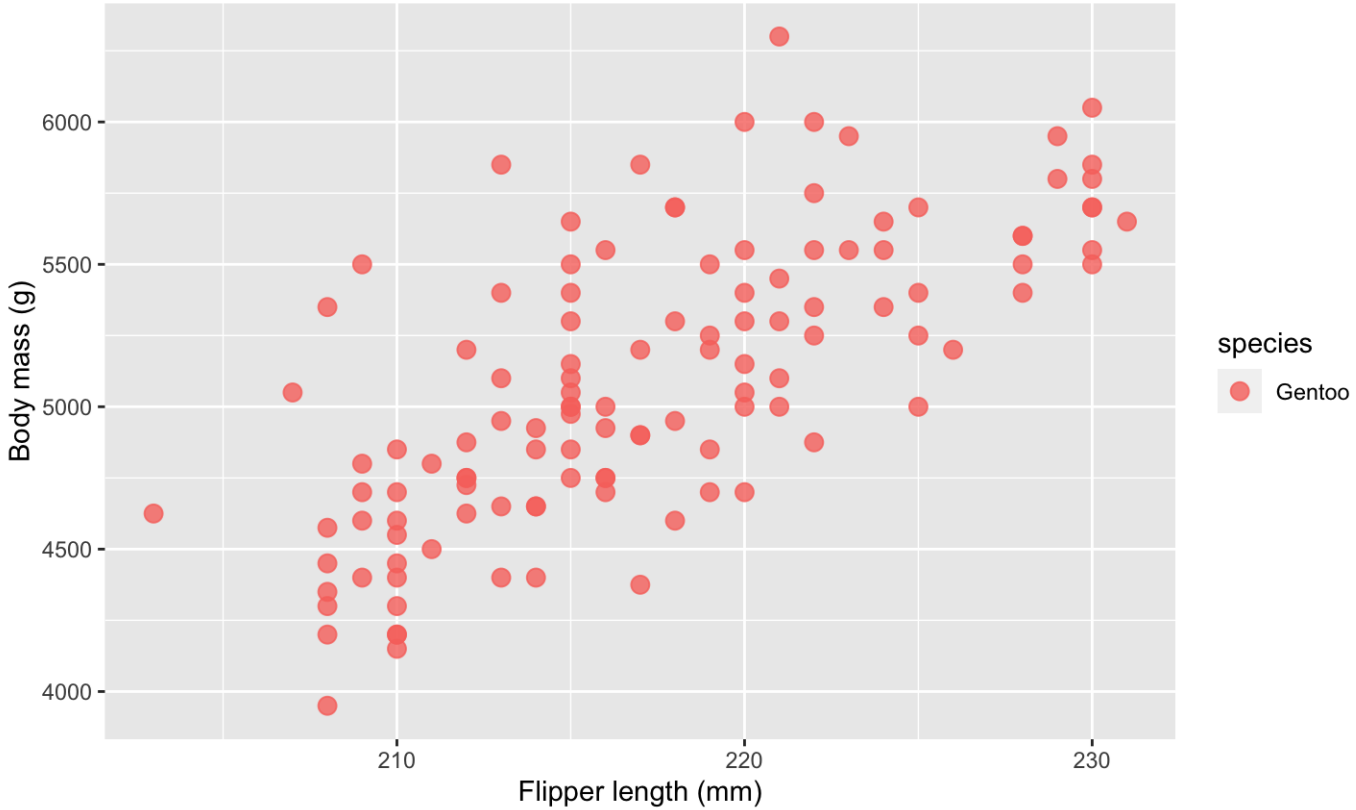
- It is week 8! You should have read the remainder of §6 of the notes on **Perusall**.
- Class test 2 was yesterday.

Linear Regression (what is it?)

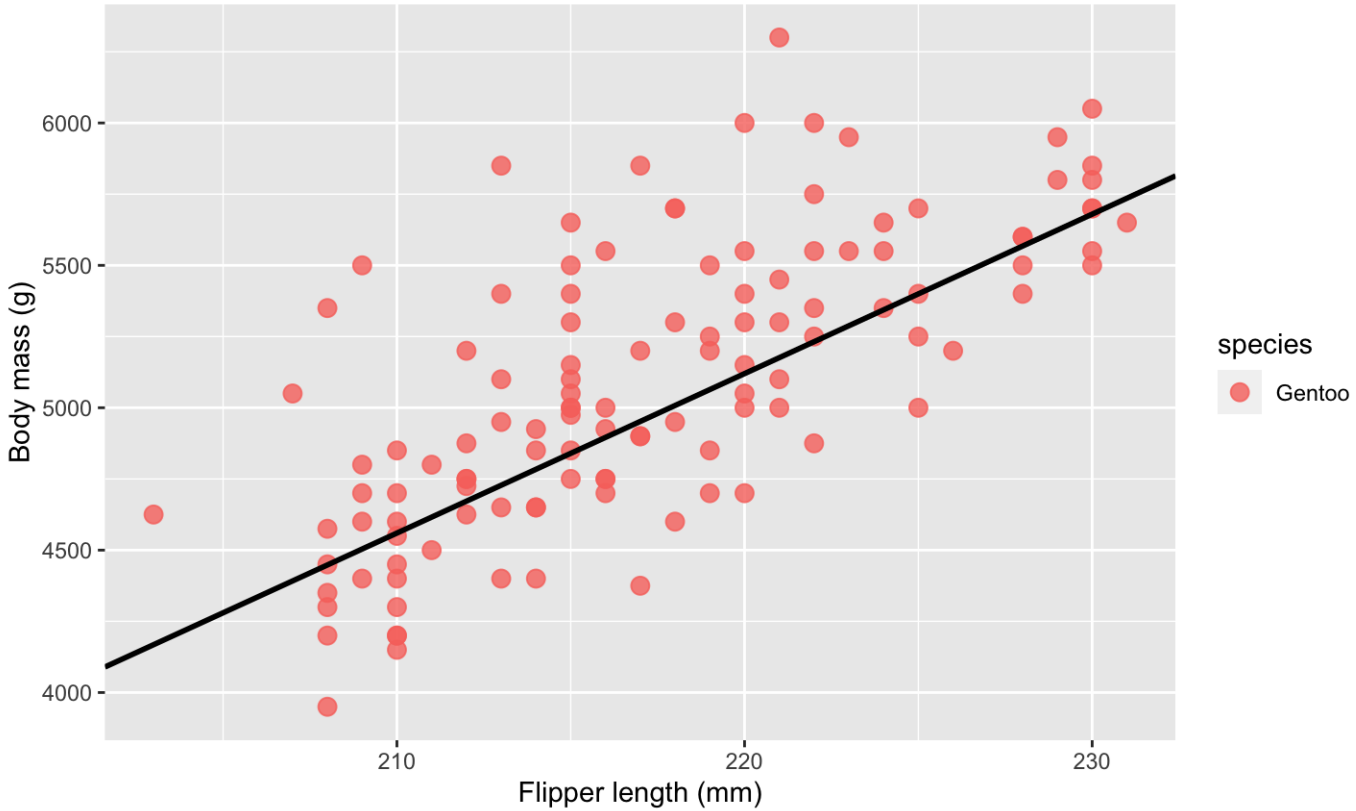
- Assumes relationship between two variables X and Y can be modeled by a straight line
- Perfect linear relationship: we would know the exact value of Y just by knowing the exact value of X

$$y = \beta_0 + \beta_1 x$$

Palmer penguins (Gentoo)



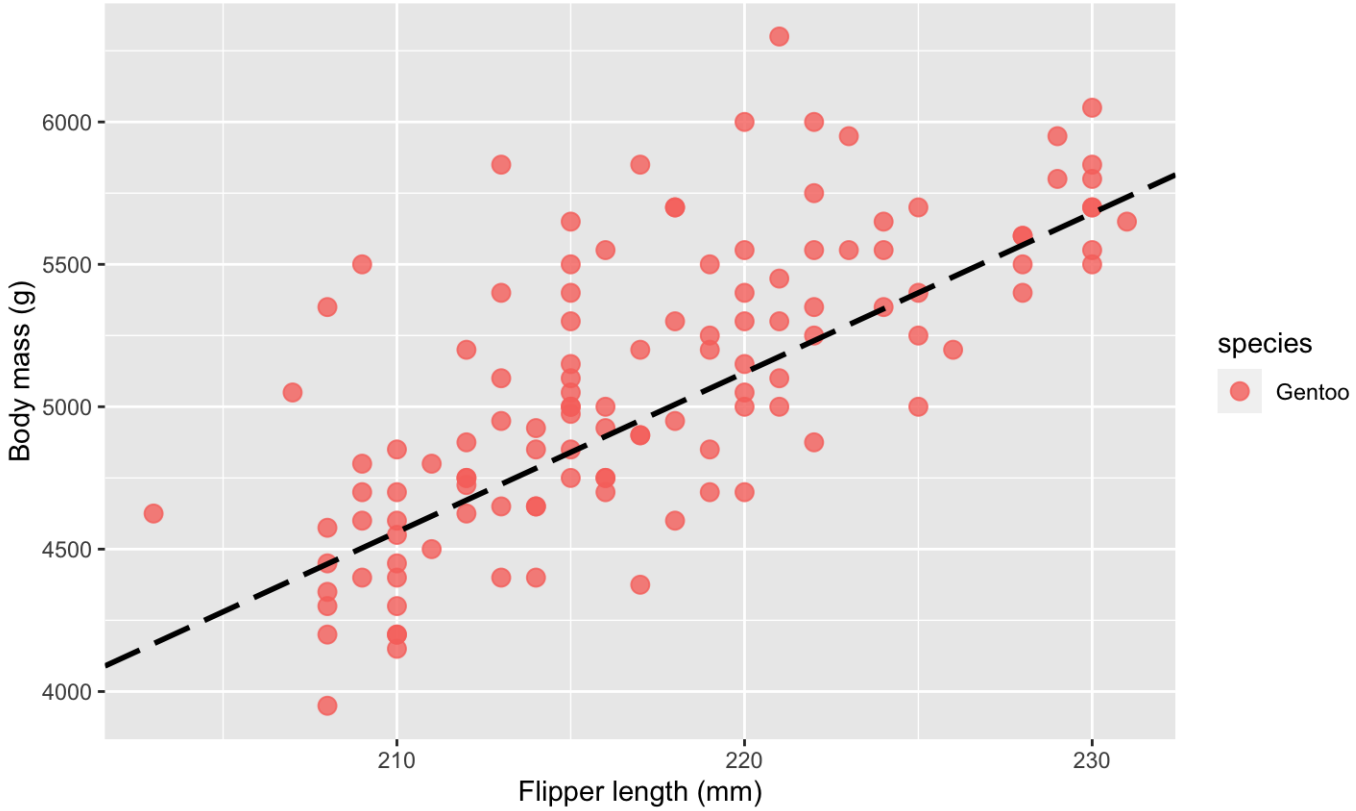
Fitting a line by eye



Residuals (every observation has a friend)

Data = Fit + Residual

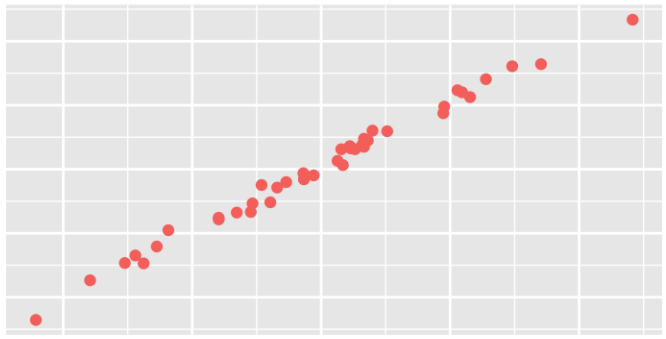
One goal in regression is to pick the linear model to minimize residuals.



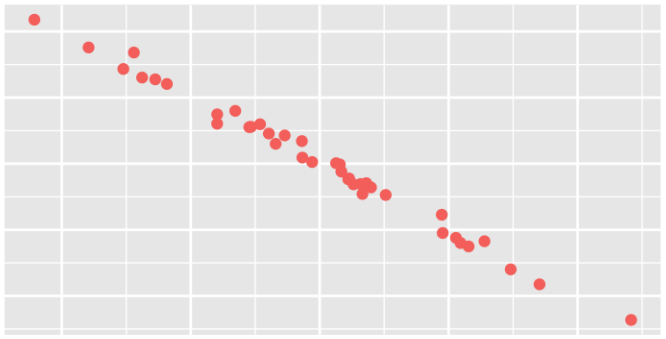
Correlation

Correlation describes the strength of a linear relationship.

(a) $r \approx +1$, linear relationship

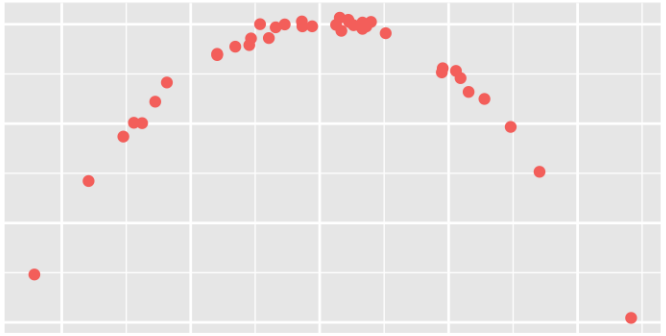


(b) $r \approx -1$, linear relationship



(c) $r \approx 0$, no relationship apparent

(d) $r \approx 0$, nonlinear relationship



Finding "best fit"

We want a line that minimizes the residuals.

- Choose the line that minimizes the sum of the squared residuals

$$\epsilon_1^2 + \epsilon_2^2 + \ldots \epsilon_m^2$$

- Commonly referred to as **least squares line**

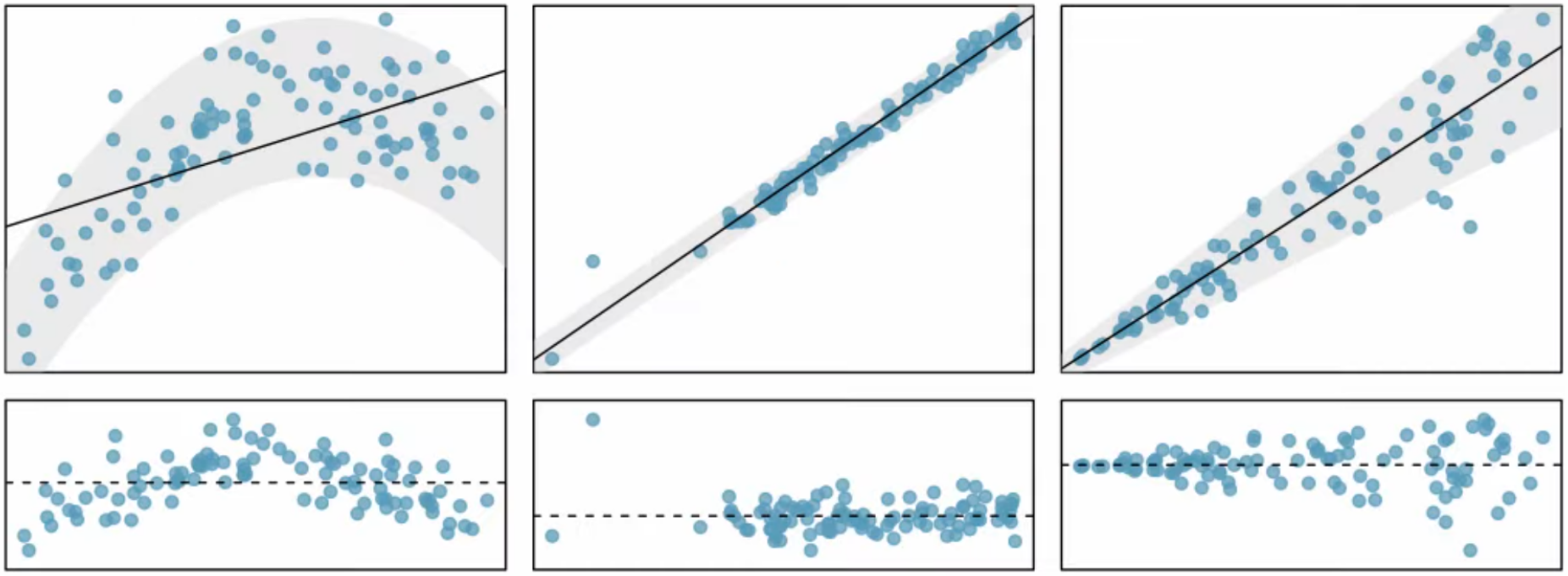
Conditions for linear regression

For fitting a least squares line we focus on three requirements:

- **Linearity**: data should show linear trend
- **Nearly normal residuals**: (usually fails due to outliers)
- **Constant variability**: variability of points around least squares line

Conditions for linear regression (in pictures)

For fitting a least squares line we focus on three requirements:



Your turn!

What conditions is this linear model obviously violating?

Finding the least squares line

Least squares estimates for $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are given by

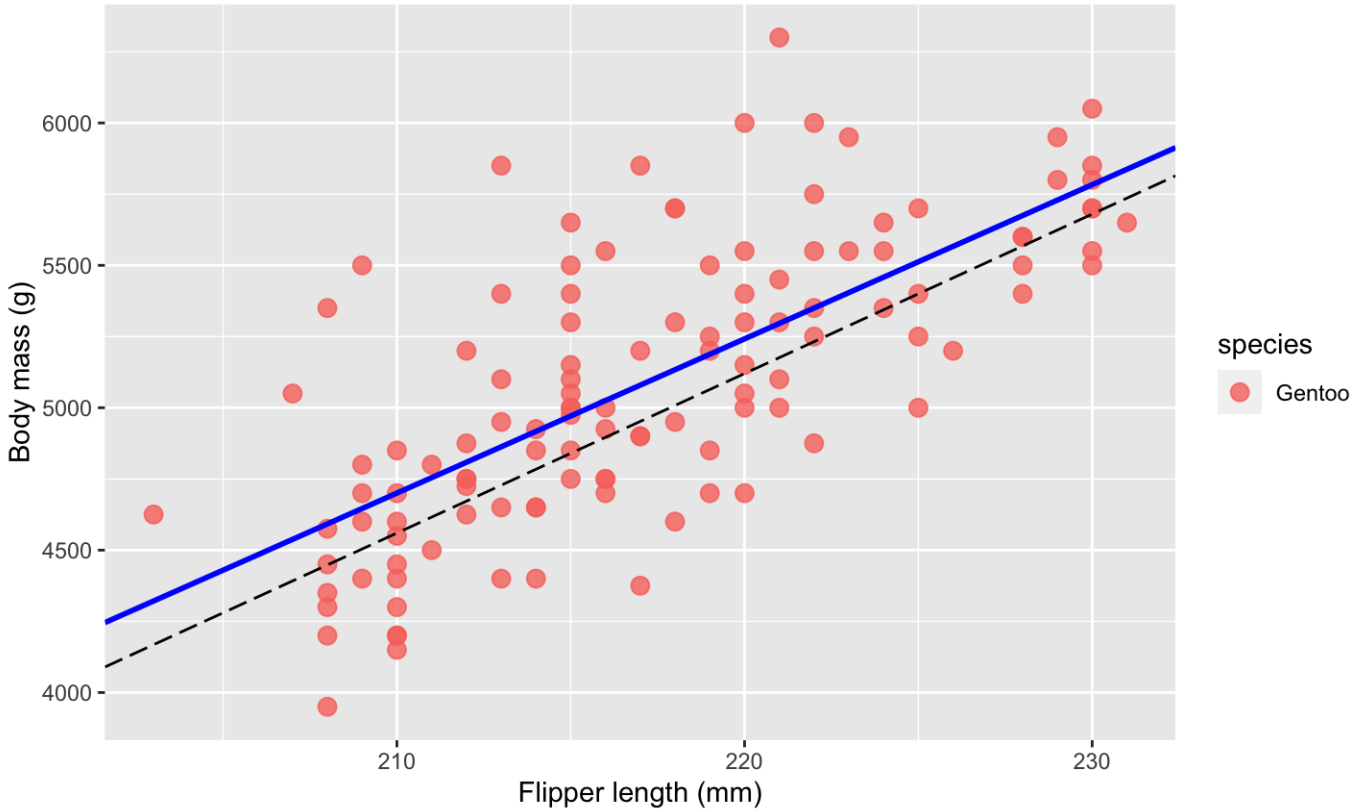
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{m}(X_i - \overline{X})^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

```
xbar <- mean(gentoo$flipper_length_mm)
ybar <- mean(gentoo$body_mass_g)
sxy <- cov(gentoo$flipper_length_mm, gentoo$body_mass_g)
sxx <- var(gentoo$flipper_length_mm)

b1 <- sxy/sxx
b0 <- ybar - b1*xbar
```

Least squares fit



What do these numbers mean?

- **SLOPE**: For each additional 10 mm flipper length, we would expect the penguin to weigh 541.7 g **more**.
- **INTERCEPT**: -6674.20 g describes the average weight if a penguin had flipper length 0…

Extrapolation can be treacherous (here there are no observations near zero).

Summary

Today we discussed least squares regression.