Class Notes

Nickhil Sethi

Hello there! If you're reading this, it means you're in Vitaly Druker's STAT 706 class. Awesome! And if you're anything like me, you're not amazing at taking notes in class, so here's some notes I typed up after re-watching the lectures online.

Most of the material is in the slides, so here I try to detail just the elements that aren't self-explanatory. Note that the professor said to pay the most attention to how to interpret the results, rather than mathematical underpinnings. Happy reading.

Class 1: Introduction

Let's take a look at the palmer penguins dataset:

```
# create a dataframe
data(package = 'palmerpenguins')
```

Table 1: Data sets in palmerpenguins

Item	Title
penguins	Size measurements for adult foraging penguins near Palmer Station, Antarctica
penguins_raw (penguins)	Penguin size, clutch, and blood isotope data for foraging adults near Palmer Station, Antarctica

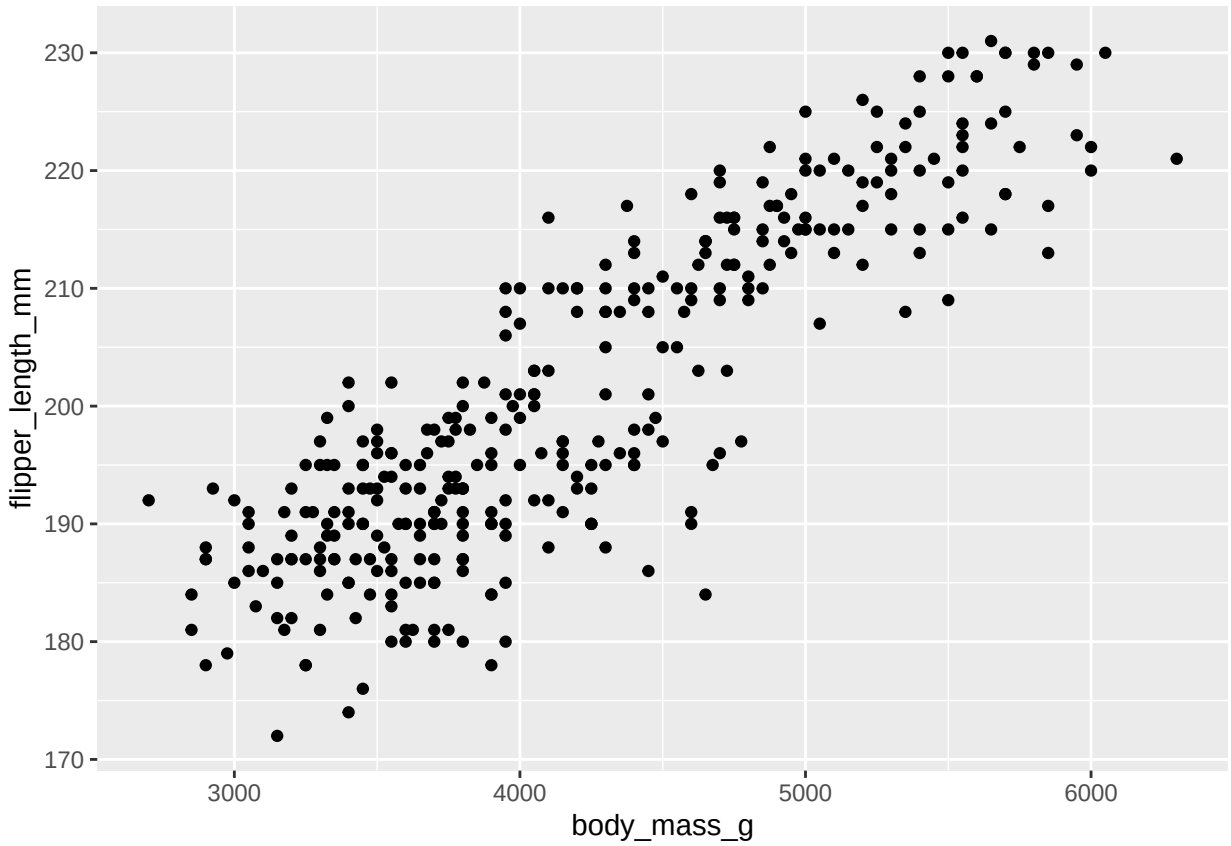
```
head(penguins)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

- Let's talk about the statistical relationship between flipper length and body mass. If we had to guess, what would we expect it to look like?
- Loosely speaking, it's probably an increasing relationship. Bigger penguins probably have larger flippers overall. Let's take a look at the data:

```
ggplot(data = penguins, mapping = aes(x = body_mass_g, y = flipper_length_mm)) + geom_point()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



- Let's define a simple model for the relationship between these two variables. Let y_i stand for flipper length and X_i stand for body mass. Here's a formula expressing a linear relationship between the two:

$$y_i = \beta_0 + \beta_1 \cdot X_i$$

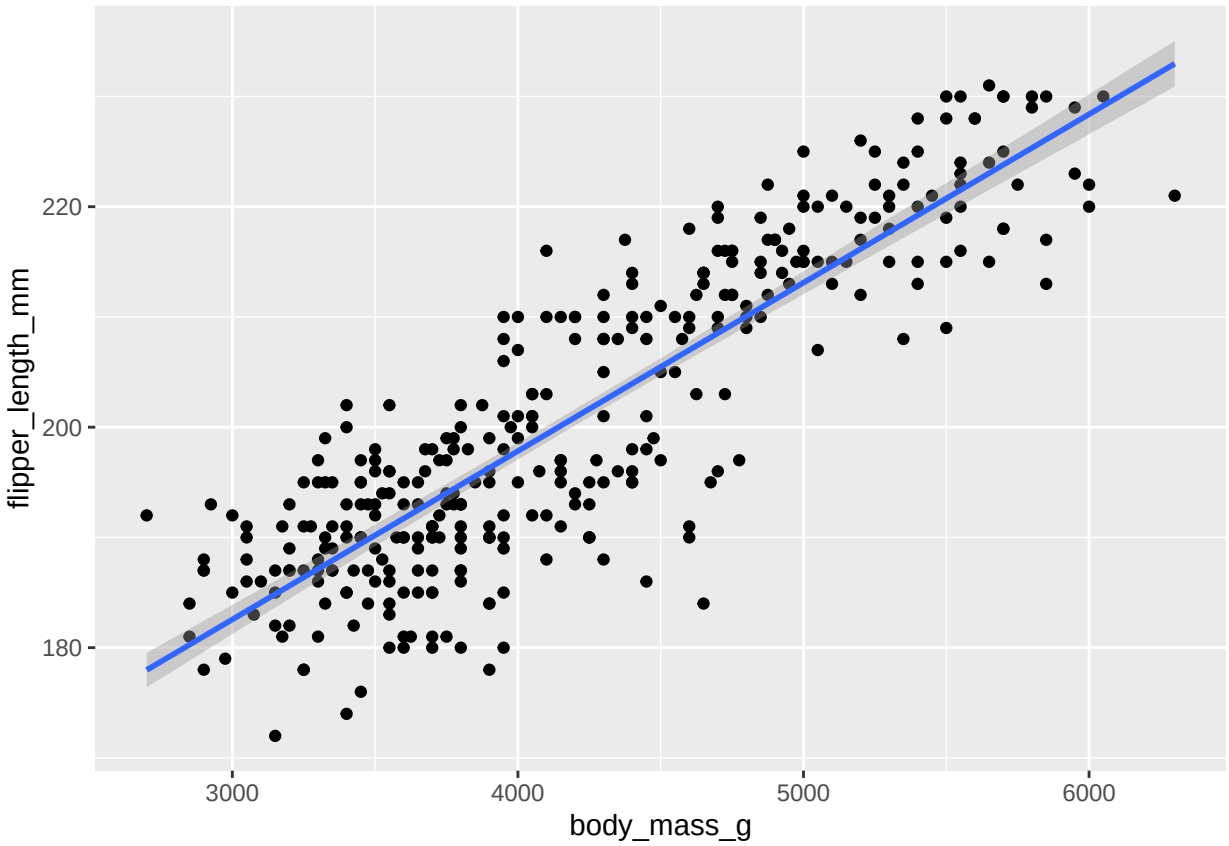
- β_0 and β_1 are free parameters that are chosen based on the dataset. Let's graph the best fit line (we'll see how to actually choose the parameters later):

```
ggplot(
  data = penguins,
  mapping = aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



- As we can see, the model doesn't perfectly capture the data. What's missing is a notion of error, which accounts for the fact that the relationship between the two variables isn't exactly a line.

$$y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$$

- How do we interpret the parameters in the model?
- β_0 is the predicted flipper length associated with a 0g penguin. This is just to say that the intercept term isn't by itself meaningful, as there's no such thing as a 0 mass penguin.
- β_1 is the slope of the line – the change in flipper length associated with a 1g increase in body mass of a penguin.

Class 2: OLS

- Picking up where we left off before, how do we interpret the parameters of the model?

$$y_i = \beta_0 + \beta_1 \cdot X_i$$

- β_0 predicted flipper length of a 0 mass penguin. i.e. β_0 by itself is not meaningful.
- β_1 the change in y associated with a 1-unit change in X
- What's missing from the model? ϵ_i captures noise / error.
- Let's think about how to model the data generating process. What should the noise ϵ_i look like?
- Symmetric above and below our regression line. i.e. the model isn't

Ordinary Linear Regression

- estimate parameters
- estimate variance
- least squares vs MLE
- residuals and their properties
- Confidence intervals vs Prediction Intervals
- Diagnostics (R squared)
- Multiple Variables
- Qualitative

-AIC

Bernoiulli/Binomial Data

- link functions
- predictions
- odds, probability
- risk ratio vs odds ratio
- goodness of fit
- Diagnostics (Pearson Residuals)
- different scoring functions
- confusion matrix and properties (sensitivity, specificity, PPV, accuracy, NPV)
- comparing nested models

-overdispersion

- f statistics
- quasibinomial