Week 9 R demo

Nathalie Moon

15/03/2021

Load penguins data

```
# Let's take a first look at the data
glimpse(penguins);
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
## $ sex               <fct> male, female, female, NA, female, male, female, m...
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

For this demo, we'll be trying to predict the length of a penguin's bill using a linear regression model.

```
# First step let's look at the association between bill depth and bill length

# Let's get rid of the warning

# Now we can fit a linear regression model to start exploring this association more deeply
```

Do you think this fitted linear model is effectively representing the association between bill depth and bill length? What other variable could we add to this model?

```
#glimpse(penguins_clean)

# New variable to add to the model:

library(broom)

# What about an interaction term?
```

Comparing the prediction accuracy of multiple models

```
# Set up


# Create training dataset
# Testing dataset includes all observations NOT in the training data



# Fit models to training data



# Make predictions for testing data using training model


# Make predictions for training data using training model


# Calculate RMSE for testing data

# Calculate RMSE for training data
```

What does it mean if the RMSE based on test data is *smaller* than the RMSE based on the training data?