Model Selection

Richard J. Telford

04 June 2021

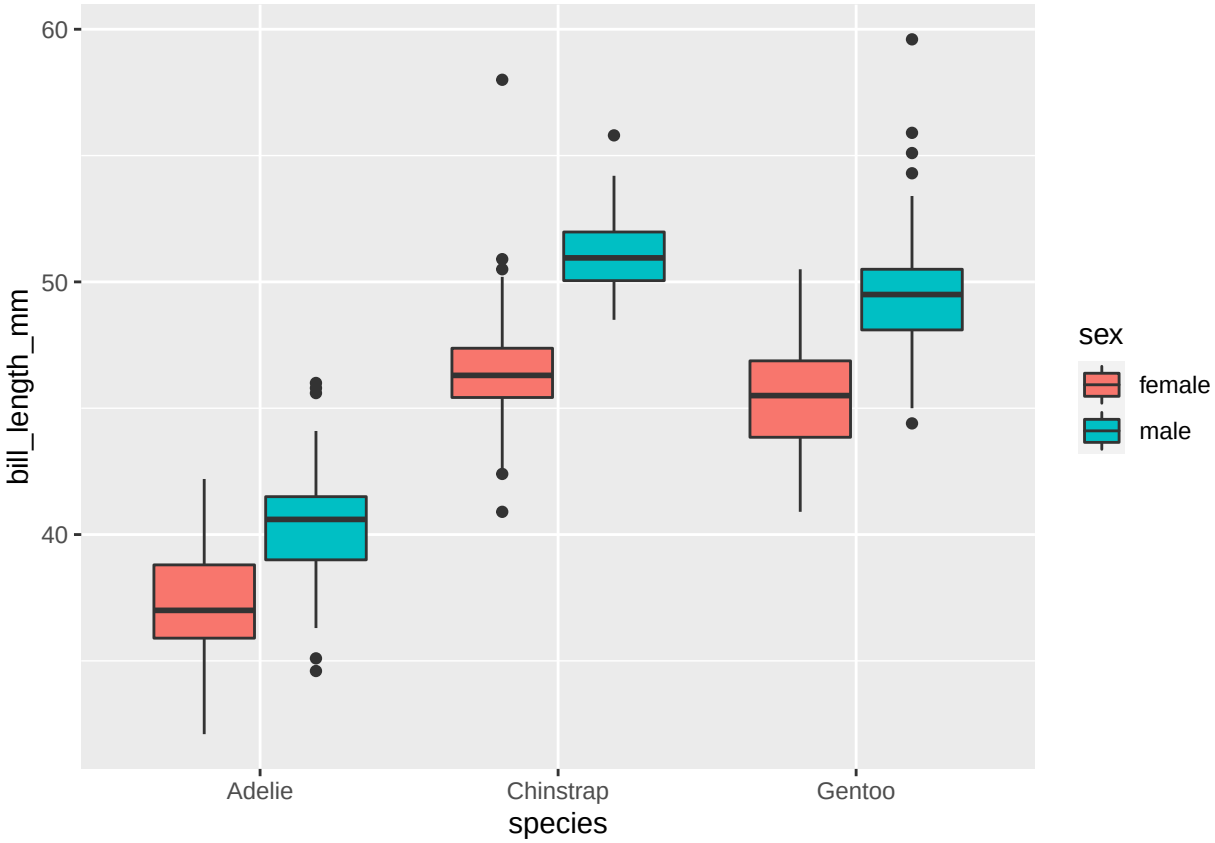
hypothesis testing

1. Import the palmerpenguin data

```
penguins_df <- penguins %>%
  drop_na()
```

2. Test the hypothesis that bill length differs between species.

```
penguins_df %>%
  ggplot(aes(x = species, y = bill_length_mm, fill = sex)) +
  geom_boxplot()
```



```
lm_df <- lm(bill_length_mm~species, data = penguins_df)
anova(lm_df)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2 7015.4  3507.7   397.3 < 2.2e-16 ***
## Residuals 330 2913.5     8.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Test the hypothesis that bill length differs by sex in addition to species.

```
lm_df2 <- lm(bill_length_mm~species + sex, data = penguins_df)

anova(lm_df,lm_df2)
```

```
## Analysis of Variance Table
##
## Model 1: bill_length_mm ~ species
## Model 2: bill_length_mm ~ species + sex
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    330 2913.5
## 2    329 1777.8  1    1135.7 210.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. How should the p-values be interpreted.

Exploratory model building

Normally you would be doing this on a separate data set

5. Use forward selection to find the best model to explain bill length.
6. Build a set of candidate models to explain bill depth using one or two predictors.
7. Extract the AIC from each models (hint use function `AIC`). Which is the better model?
8. Calculate the deltaAIC for each model.
9. Calculate the AIC weights for each model. Interpret these weights.

Collinearity

12. Make a model predicting bill_length from all other variables. Find the VIF of each predictor. Are there any problem variables? `olsrr::ols_vif_tol`
13. Use `GGally::ggpairs()` to plot the data to try to identify the cause of any high vif.
14. Use `MASS::mvrnorm()` to simulate 100 observation of two predictor variables (x and z) with a given correlation. Simulate a response variable y = b0 + b1x + b2z. Test how the uncertainty in the coefficients changes with the correlation (and hence vif) of the predictor variables.