Lab 06: Adelie Penguins

Logistic regression intro

Stats is 'Fun' - Dav King, Luke Thomas, Thomas Barker, Harry Liu

2022-11-08

Setup

Load packages

```
library(tidyverse)
library(tidymodels)
library(palmerpenguins)
library(knitr)
library(patchwork)

penguins <- penguins %>%
  mutate(adelie = factor(if_else(species == "Adelie", 1, 0)))
penguins %>%
  count(adelie, species)
```

```
# A tibble: 3 x 3
  adelie species       n
  <fct>  <fct>     <int>
1 0      Chinstrap    68
2 0      Gentoo      124
3 1      Adelie      152
```

[Select this page for the "Workflow & formatting" in Gradescope.]

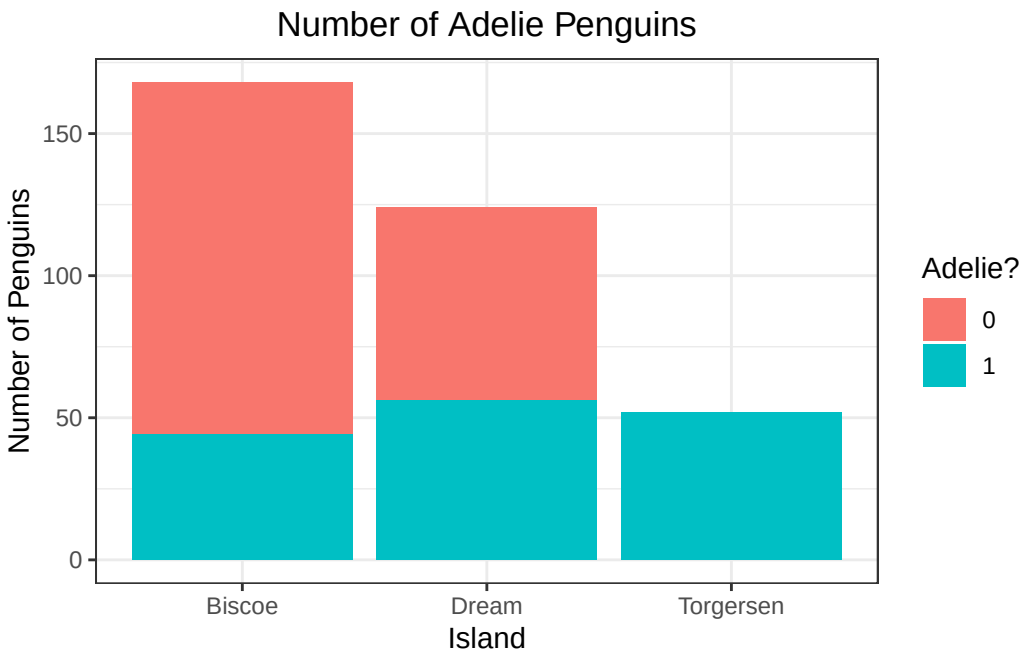
Exercises

Exercise 1

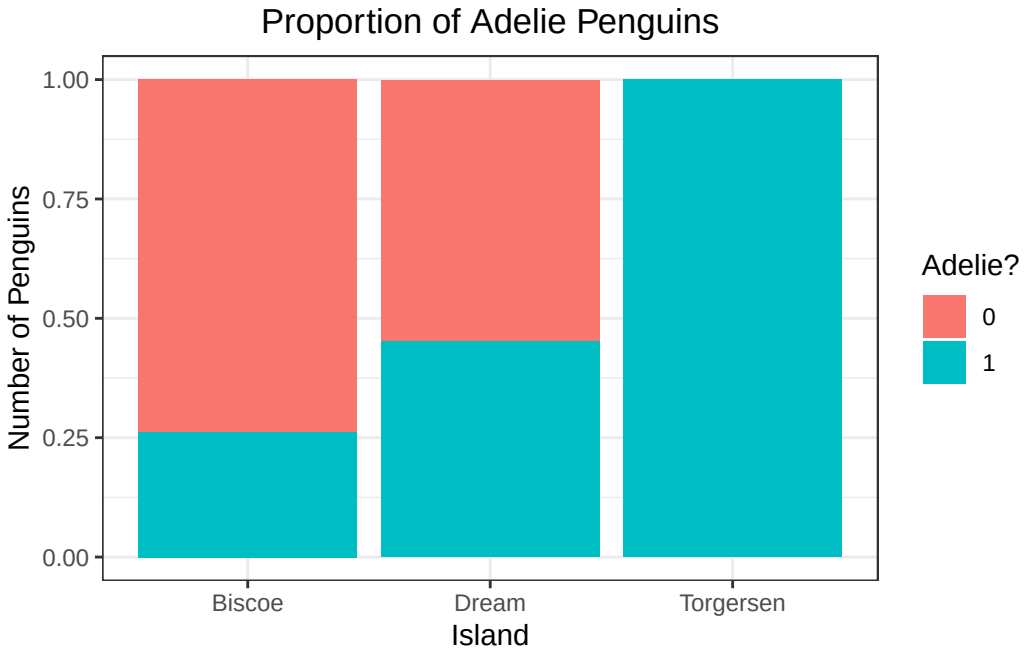
```
tot <- ggplot(penguins, aes(x = island, fill = adelie)) +
  geom_bar() +
  theme_bw() +
  labs(x = "Island", y = "Number of Penguins", fill = "Adelie?",
       title = "Number of Adelie Penguins") +
  theme(plot.title = element_text(hjust = 0.5))

perc <- ggplot(penguins, aes(x = island, fill = adelie)) +
  geom_bar(position = "fill") +
  theme_bw() +
  labs(x = "Island", y = "Number of Penguins", fill = "Adelie?",
       title = "Proportion of Adelie Penguins") +
  theme(plot.title = element_text(hjust = 0.5))

tot
```




```
perc
```



One thing we observe is that there are roughly comparable raw numbers of Adelie penguins on every island, even though the numbers of total penguins across species differ greatly by island. Thus, Biscoe island penguins are only ~25% adelie, while Dream island penguins are close to 50% Adelie and 100% of Torgersen island penguins are Adelie.

Exercise 2

```
penguins %>%
  count(island, adelie) %>%
  pivot_wider(names_from = adelie, values_from = n, values_fill = 0)
```

```
# A tibble: 3 x 3
  island      `0`   `1`
  <fct>     <int> <int>
1 Biscoe      124    44
2 Dream        68    56
3 Torgersen     0    52
```

When a value is missing (or `NA`), `values_fill` here specifies that it should replace all of those missing values with a 0. That means that, for Torgersen island, the number of non-adelie penguins (which was 0, or `NA`), should be filled with the value 0.

…

Exercise 3

```r
penguins |>
  filter(island == 'Biscoe') |>
  count(adelie) |>
  mutate(prob = round(n / sum(n), 3)) |>
  mutate(odds = round(prob / (1 - prob), 3))
```

```
# A tibble: 2 x 4
  adelie     n  prob  odds
  <fct>  <int> <dbl> <dbl>
1 0        124 0.738 2.82
2 1         44 0.262 0.355
```

The **probability** a randomly selected penguin is from the Adelie species if it was recorded on Biscoe island is 0.262, and the **odds** a randomly selected penguin is from the Adelie species if it was recorded on Biscoe island is 0.355.

Exercise 4

```
adelie_fit <- logistic_reg() |>
  set_engine("glm") |>
  fit(adelie ~ island, data = penguins, family = "binomial")

tidy(adelie_fit) |> kable(digits = 3)
```

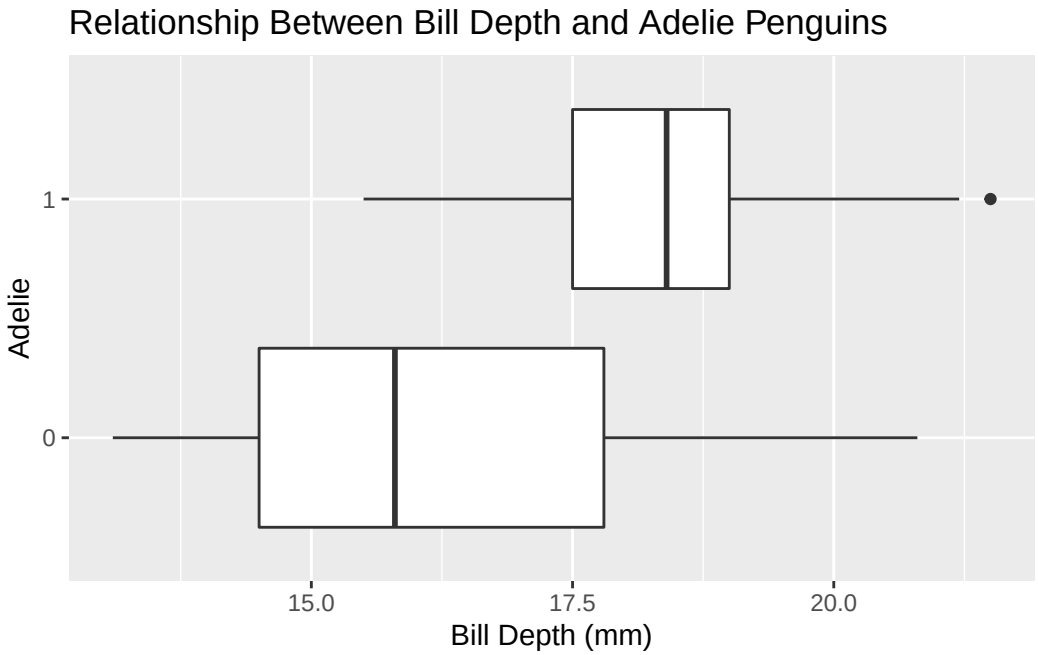
term	estimate	std.error	statistic	p.value
(Intercept)	-1.036	0.175	-5.904	0.000
islandDream	0.842	0.252	3.345	0.001
islandTorgersen	19.602	904.527	0.022	0.983

The predicted odds of a penguin being from the Adelie species if it was recorded on Biscoe island is $e^{-1.036} = 0.356$, and the predicted odds of a penguin being from the Adelie species if it was recorded on Dream island is $e^{-1.036 + 0.842} = 0.824$.

Exercise 5

```
ggplot(data = penguins, aes(x = bill_depth_mm, y = adelie)) +
  geom_boxplot() +
  labs(title = "Relationship Between Bill Depth and Adelie Penguins",
       x = "Bill Depth (mm)",
       y = "Adelie")
```

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



Observation:

Adelie penguins appear to have deeper bills, as the minimum, median, and maximum bill lengths of Adelie penguins are greater than the minimum, median, and maximum bill lengths of non-Adelie penguins.

Exercise 6

```
depth_fit <- logistic_reg() |>
  set_engine("glm") |>
  fit(adelie ~ island + bill_depth_mm, data = penguins, family = "binomial")

tidy(depth_fit) |> kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-14.676	1.881	-7.804	0.000
islandDream	-0.892	0.359	-2.481	0.013
islandTorgersen	18.132	822.821	0.022	0.982
bill_depth_mm	0.836	0.113	7.416	0.000

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -14.676 + -0.892 \times islandDream + 18.132 \times islandTorgersen + 0.836 \times bill\_depth\_mm$$

Exercise 7

```
log_odds <- -14.676 - 0.892*1 + 0.836*17
log_odds_new <- -14.676 - 0.892*1 + 0.836*20
log_odds_change <- log_odds_new - log_odds
log_odds_change
```

```
[1] 2.508
```

```
exp(log_odds_change)
```

```
[1] 12.28034
```

The log-odds of being from the Adelie species are expected to be 2.508 more for those penguins with bill depth of 20 mm compared to those with bill depth of 17 mm, given that the penguins were recorded on the Dream island and all else equal.

The odds of being from the Adelie species for those penguins with bill depth of 20 mm are expected to be 12.28 (exp(2.508)) times the odds for those penguins with bill depth of 17 mm, given that the penguins were recorded on the Dream island and all else equal.

Exercise 8

```
log_odds2 <- -14.676 + 0.836*18
log_odds2_new <- -14.676 - 0.892*1 + 0.836*21
log_odds2_change <- log_odds2_new - log_odds2
log_odds2_change
```

```
[1] 1.616
```

```
exp(log_odds2_change)
```

```
[1] 5.032918
```

The log-odds of being from the Adelie species are expected to be 1.616 more for those penguins with bill depth of 21 mm recorded on Dream island compared to those with bill depth of 18 mm recorded on Biscoe island, given all else equal.

The odds of being from the Adelie species for those penguins with bill depth of 21 mm recorded on Dream island are expected to be 5.03 (exp(1.616)) times the odds for those penguins with bill depth of 18 mm recorded on Biscoe island, given all else equal.