HW 02: Multiple linear regression

Dav King

2022-10-18

Set up

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

> **Note**
>
> Select this page for Workflow & formatting".

Exercises

Exercise 1

```
penguins <- drop_na(penguins)

set.seed(123)
penguinSplit <- initial_split(penguins)
penguinTrain <- training(penguinSplit)
penguinTest <- testing(penguinSplit)
```

Exercise 2

```
penguinRec <- recipe(body_mass_g ~ ., data = penguinTrain) %>%
  update_role(island, new_role = "ID") %>%
  update_role(year, new_role = "ID") %>%
  step_center(bill_length_mm, bill_depth_mm, flipper_length_mm) %>%
  step_dummy(all_nominal_predictors())

penguinSpec <- linear_reg() %>%
  set_engine("lm")

penguinWFlow <- workflow() %>%
  add_model(penguinSpec) %>%
  add_recipe(penguinRec)

penguinFit <- penguinWFlow %>%
  fit(data = penguinTrain)

tidy(penguinFit) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	3693.104	73.811	50.034	0.000
bill_length_mm	24.267	8.147	2.979	0.003
bill_depth_mm	76.549	22.516	3.400	0.001
flipper_length_mm	13.788	3.278	4.206	0.000
species_Chinstrap	-312.150	89.932	-3.471	0.001
species_Gentoo	1044.917	142.580	7.329	0.000
sex_male	396.164	54.581	7.258	0.000

$\hat{body\_mass\_g} = 3693.104 - 312.150 \times Chinstrap + 1044.917 \times Gentoo + 24.267 \times bill\_length\_mm(mean) + 76.549 \times bill\_depth\_mm(mean) + 13.788 \times flipper\_length\_mm(mean) + 396.164 \times male$

Exercise 3

Holding all other predictors constant, for every 1 millimeter increase in a penguin's bill length, we would expect its mass to increase by 24.267 grams, on average.

Holding all other predictors constant, for every 1 millimeter increase in a penguin's bill depth, we would expect its mass to increase by 76.549 grams, on average.

Holding all other predictors constant, for every 1 millimeter increase in a penguin's flipper length, we would expect its mass to increase by 13.788 grams, on average.

Holding all other predictors constant, if a penguin is a Chinstrap penguin, we would expect its mass to decrease by 312.150 grams, on average, compared to an otherwise identical Adelie penguin.

Holding all other predictors constant, if a penguin is a Gentoo penguin, we would expect its mass to increase by 1044.917 grams, on average, compared to an otherwise identical Adelie penguin.

Holding all other predictors constant, if a penguin is male, we would expect its mass to increase by 396.164 grams, on average, compared to an otherwise identical female penguin.

For a female Adelie penguin with a bill length of 43.82892 mm, a bill depth of 17.08273 mm, and a flipper length of 201.2329 mm, we would expect to see a body mass of 3693.104 grams.

Exercise 4

```
sample <- tibble(species = "Adelie", bill_length_mm = 39.1,
                 bill_depth_mm = 18.7, flipper_length_mm = 181, sex = "male",
                 island = NA, year = NA)
3750 - predict(penguinFit, sample)
```

```
     .pred
1 -69.34174
```

```
glance(penguinFit)$r.squared
```

```
[1] 0.8833681
```

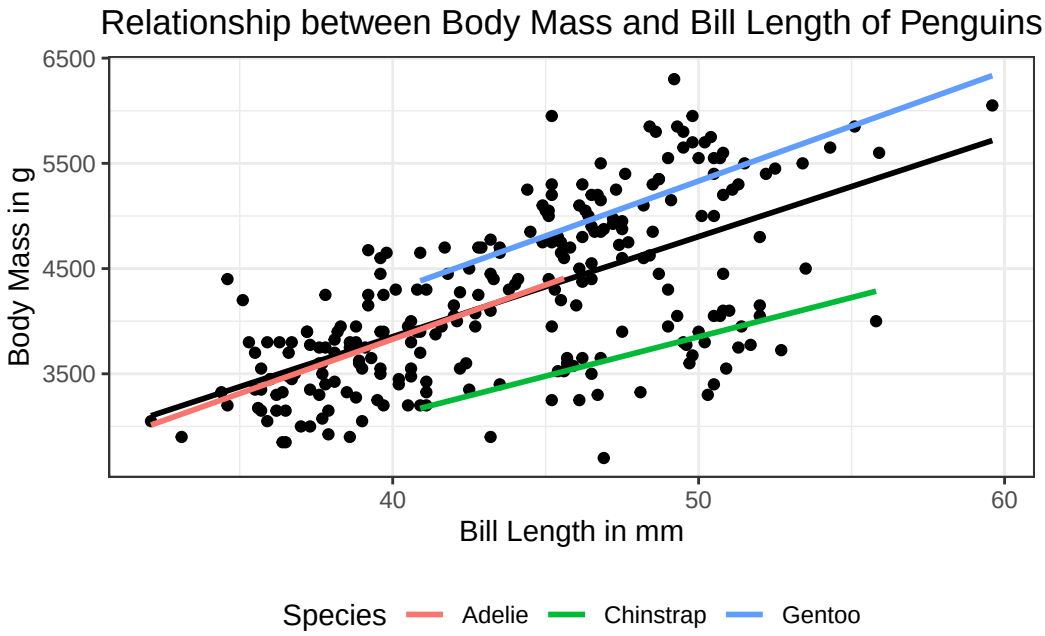
This model gives us a negative residual for this penguin, meaning it overpredicts the penguin's weight.

The R^2 value of 0.883 tells us that, based on the training data, 88.3% of the variance in a penguin's weight can be explained by its species, bill length, bill depth, flipper length, and sex.

Exercise 5

```
ggplot(penguinTrain, aes(x = bill_length_mm, y = body_mass_g)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, color = "black") +
  geom_smooth(method = "lm", se = F, aes(color = species)) +
  theme_bw() +
  labs(x = "Bill Length in mm", y = "Body Mass in g", color = "Species",
       title = "Relationship between Body Mass and Bill Length of Penguins") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "bottom")
```

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



This visualization does not provide much evidence for an interaction effect between bill length and species in the prediction of body mass. While upon initial inspection, there are three clear clusters of data points for the different species of penguins, when we fit regression lines through them, they seem to have different intercepts but more or less identical slopes. Thus, we would not suspect an interaction between bill length and species.

Exercise 6

```
interactRec <- recipe(body_mass_g ~ bill_length_mm + species,
                      data = penguinTrain) %>%
  step_center(bill_length_mm) %>%
  step_dummy(species) %>%
  step_interact(terms = ~ bill_length_mm:starts_with("species"))

interactWFlow <- workflow() %>%
  add_model(penguinSpec) %>%
  add_recipe(interactRec)

interactFit <- interactWFlow %>%
  fit(data = penguinTrain)

tidy(interactFit) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	4223.895	83.146	50.801	0.000
bill_length_mm	103.180	14.547	7.093	0.000
species_Chinstrap	-833.092	126.108	-6.606	0.000
species_Gentoo	464.424	102.277	4.541	0.000
bill_length_mm_x_species_Chinstrap	-28.508	22.297	-1.279	0.202
bill_length_mm_x_species_Gentoo	1.148	18.958	0.061	0.952

Exercise 7

```
multiModelPred <- predict(penguinFit, penguinTest) %>%
  bind_cols(penguinTest)
interactModelPred <- predict(interactFit, penguinTest) %>%
  bind_cols(penguinTest)

rsq(multiModelPred, truth = body_mass_g, estimate = .pred)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.846
```

```
rsq(interactModelPred, truth = body_mass_g, estimate = .pred)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.771
```

```
rmse(multiModelPred, truth = body_mass_g, estimate = .pred)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        309.
```

```
rmse(interactModelPred, truth = body_mass_g, estimate = .pred)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        375.
```

By multiple measures, the model fit in exercise 2 outperforms the model fit in exercise 6 in terms of predicting body mass within the testing data. It has a higher R^2 value (0.846 vs 0.771), suggesting that it can explain around 7.5% more of the variance in body mass based on the predictors it uses. Additionally, it has a lower RMSE value (309 vs 375), suggesting that its average error in predicting body mass is less than the average error of the model created in exercise 6.

Exercise 8

Holding all other predictors constant, for every 1 standard deviation increase in participant age, we would expect to see an increase of 0.072 in their perception of the threat of Covid, on average.

Holding all other predictors constant, if a participant was living in the European Union, we would expect their perception of the threat of Covid to decrease by 0.614, on average, compared to identical participants living in the US or Canada.

Exercise 9

Holding all other predictors constant, the effect of a participant's emotionality on their perception of the threat of Covid differs by 0.101 when the participant lives in the EU compared to participants who live in the US/Canada. However, do keep in mind that their model found this effect to be not significant.

If a person lives in the US/Canada, we expect their perception of the threat of Covid to increase by 0.188 for each additional standard deviation in emotionality, holding all other predictors constant.

If a person lives in the EU, we expect their perception of the threat of Covid to increase by 0.289 for each additional standard deviation in emotionality, holding all other predictors constant.

Exercise 10

For every additional 1% of the population living in urban areas, the median GDP of a country is expected to multiply by a factor of 1.042894 [exp(0.042)].

For a country with 0% of its population living in urban areas, the median GDP is expected to be 450.34 US$ [exp(6.11)].