PSTAT 131 HW 2

Ezra Torio

2022-10-16

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.1     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggthemes)
tidymodels_prefer()
```

```
getwd()
```

```
## [1] "/Users/ezratorio"
```

```
data <- read_csv("~/Desktop/abalone.csv")
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

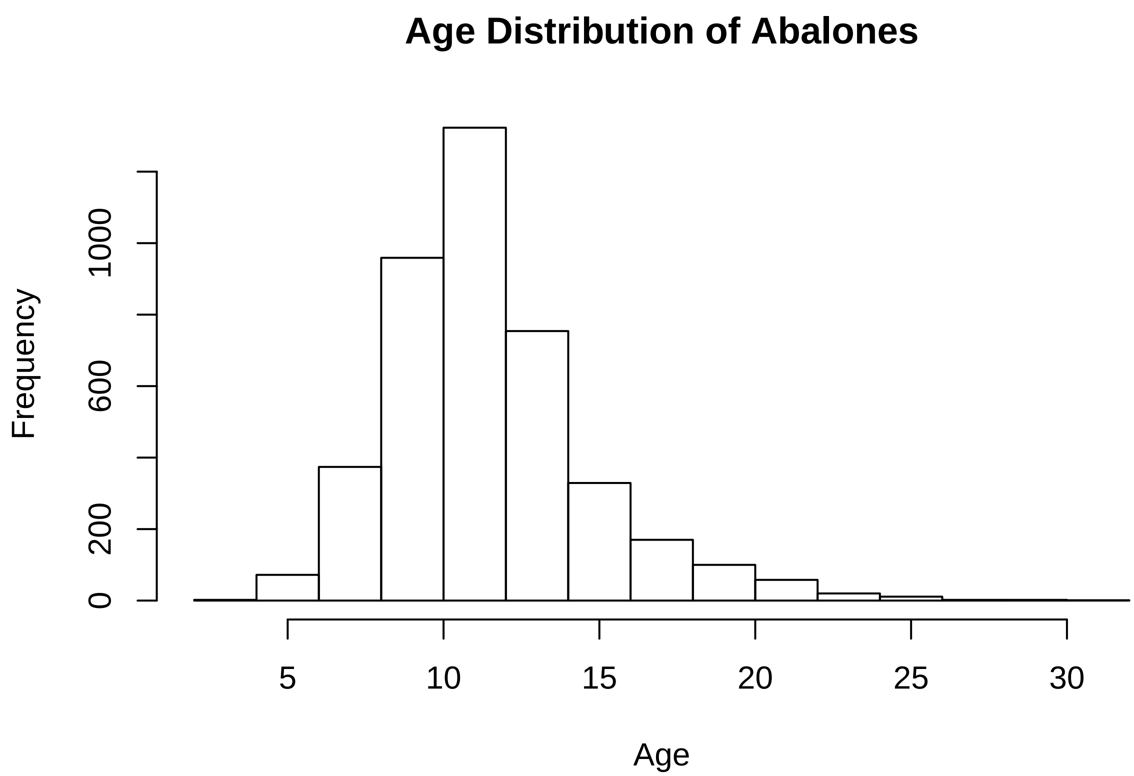
```
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(data)
```

```
## # A tibble: 6 x 9
##   type  longest_shell diameter height whole_weight shuck~1 visce~2 shell~3 rings
##   <chr>         <dbl>    <dbl>  <dbl>        <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1 M             0.455    0.365  0.095        0.514  0.224   0.101    0.15     15
## 2 M             0.35     0.265  0.09         0.226  0.0995  0.0485   0.07      7
## 3 F             0.53     0.42   0.135        0.677  0.256   0.142    0.21      9
## 4 M             0.44     0.365  0.125        0.516  0.216   0.114    0.155    10
## 5 I             0.33     0.255  0.08         0.205  0.0895  0.0395   0.055     7
## 6 I             0.425    0.3    0.095        0.352  0.141   0.0775   0.12      8
## # ... with abbreviated variable names 1: shucked_weight, 2: viscera_weight,
## #   3: shell_weight
```

Question 1

```
newData <- data
newData$age <- data$rings + 1.5
hist(newData$age, xlab = "Age", main = "Age Distribution of Abalones")
```



The abalone ages are normally distributed and right skewed. Most abalones fall between 7 and 15 years old. It is very rare to find an abalone older than 20.

Question 2

```r
set.seed(823)

abalone_split <- initial_split(newData, prop = 0.80, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

We should not use rings to predict age because age is just (rings +1.5) meaning that they would be perfectly correlated.

```r
abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
 step_rm(rings) %>%
  step_dummy(all_nominal_predictors()) %>%
   step_interact(~ starts_with("type"):shucked_weight) %>%
    step_interact(~ longest_shell:diameter) %>%
     step_interact(~ shucked_weight:shell_weight) %>%
      step_normalize(all_predictors())

abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          9
##
## Operations:
##
## Variables removed rings
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight
## Interactions with longest_shell:diameter
## Interactions with shucked_weight:shell_weight
## Centering and scaling for all_predictors()
```

Question 4

```r
lm_model <- linear_reg() %>%
  set_engine("lm")
```

Question 5

```r
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)

lm_fit <- fit(lm_wflow, abalone_train)
```

Question 6

```r
testAbalone <- tibble(type = "F", longest_shell = 0.50, diameter = 0.10, height = 0.30,
                     whole_weight = 4, shucked_weight = 1, viscera_weight = 2,
                     shell_weight = 1, rings = 0)

predict(lm_fit, new_data = testAbalone)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.2
```

Predicted age: 23.22974

Question 7

```r
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1  9.47
## 2  8.11
## 3  9.77
## 4 10.4
## 5 10.1
## 6  6.28
```

```r
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.47   8.5
## 2  8.11   8.5
## 3  9.77   8.5
## 4 10.4    8.5
## 5 10.1    9.5
## 6  6.28   6.5
```

```r
rmse(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.17
```

```r
abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
```

```
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.17
## 2 rsq     standard       0.555
## 3 mae     standard       1.55
```