pstat131-hw2

Sissi Shen

2022-10-15

```
abalone = read.csv(file='abalone.csv')
```

Question 1:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

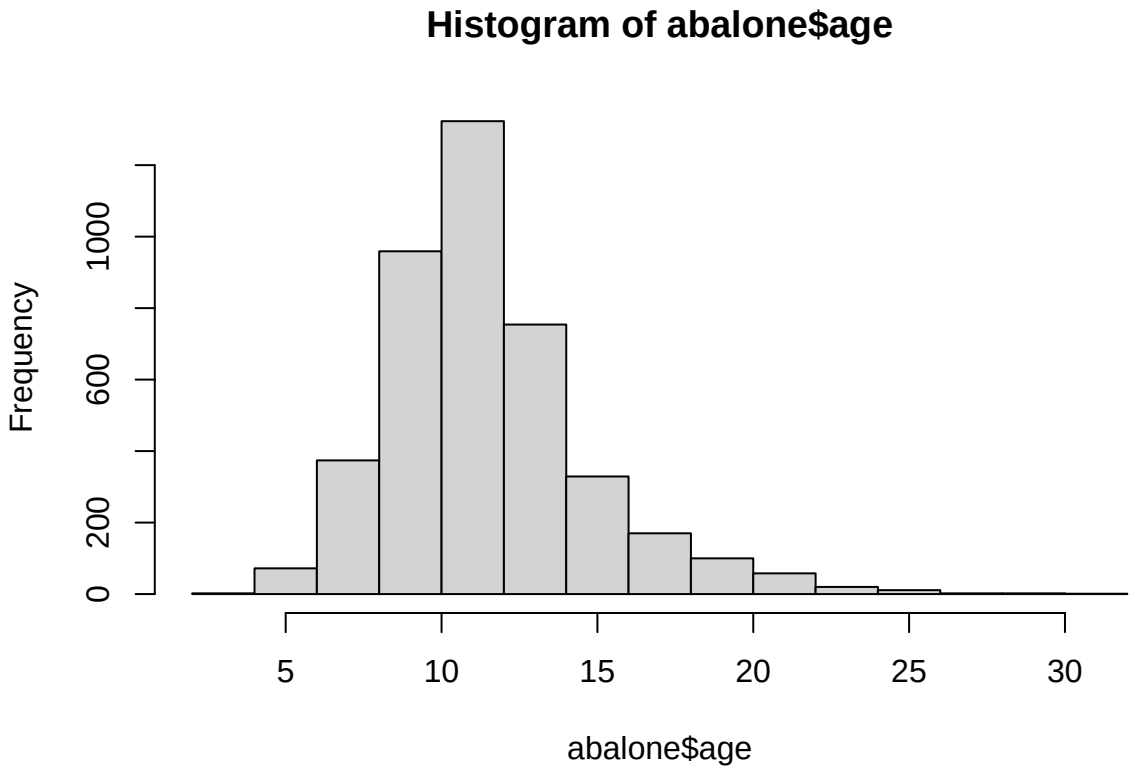
```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --

## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tibble       3.1.8
## v ggplot2      3.3.6     v tidyr        1.2.1
## v infer        1.0.3     v tune         1.0.1
## v modeldata    1.0.1     v workflows    1.1.0
## v parsnip      1.0.2     v workflowsets 1.0.0
## v purrr        0.3.5     v yardstick    1.1.0
## v recipes      1.0.2

## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v readr   2.1.3     v forcats 0.5.2
## v stringr 1.4.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x readr::spec()       masks yardstick::spec()
```

```
abalone <- abalone %>%
  mutate(age = rings + 1.5)
hist(abalone$age)
```



From the histogram, we can tell that the distribution of abalone's age is skewed to the left, which indicates that most abalone in our data sample are relatively young.

Question 2:

```
set.seed(1029)
abalone_split <- initial_split(abalone, prop=0.8, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3:

```
abalone_train <- abalone_train %>% select(-rings)
abalone_test <- abalone_test %>% select(-rings)
```

We should not include the variable "rings" because it is basically the outcome "age" that we want to predict.

```
simple_abalone_recipe <- recipe(age ~ ., data = abalone_train)
simple_abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
```

```
abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
  step_interact(terms = ~ shucked_weight:shell_weight) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors())
```

Question 4:

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

Question 5:

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

Question 6:

```
lm_fit <- fit(lm_wflow, abalone_train)
lm_fit %>%
  extract_fit_parsnip()%>%
  tidy()
```

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     11.4      0.0371    308.   0
##  2 longest_shell                    0.747    0.281       2.66 7.86e- 3
##  3 diameter                         2.07     0.306       6.75 1.74e-11
##  4 height                           0.226    0.0697      3.24 1.19e- 3
```

```
##  5 whole_weight                    5.41     0.414      13.1  3.25e-38
##  6 shucked_weight                 -4.31     0.253     -17.0  1.67e-62
##  7 viscera_weight                 -0.976    0.161      -6.07 1.42e- 9
##  8 shell_weight                    1.58     0.214       7.38 2.00e-13
##  9 type_I                         -1.06     0.114      -9.26 3.72e-20
## 10 type_M                         -0.345    0.102      -3.37 7.54e- 4
## 11 type_I_x_shucked_weight         0.564    0.0869      6.49 9.76e-11
## 12 type_M_x_shucked_weight         0.366    0.109       3.36 7.97e- 4
## 13 longest_shell_x_diameter       -3.35     0.400      -8.36 9.19e-17
## 14 shucked_weight_x_shell_weight  -0.209    0.201      -1.04 2.99e- 1
```

```
H_abalone <- data.frame(type = "F", longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight =
predict(lm_fit, H_abalone)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  25.9
```

From the model's prediction, the age of our hypothetical female abalone should be 25.87 years old.

Question 7:

```
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
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  8.06   8.5
## 2  9.76   8.5
## 3 10.4    8.5
## 4 11.1    9.5
## 5  6.24   6.5
## 6  5.72   6.5
```

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

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.14
```

```
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       2.14
## 2 rsq     standard       0.558
## 3 mae     standard       1.54
```

The model has a root mean squared error (RMSE) of 2.138, an R^2 of 0.5575 and a mean absolute error (MAE) of 1.542. An R^2 value of 0.5575 means that the model can explain 55.75% of the variance of the observed data.