HW2

Hanying Feng

2022-10-13

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## 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      1.0.2     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()
## * Use tidymodels_prefer() to resolve common conflicts.
```

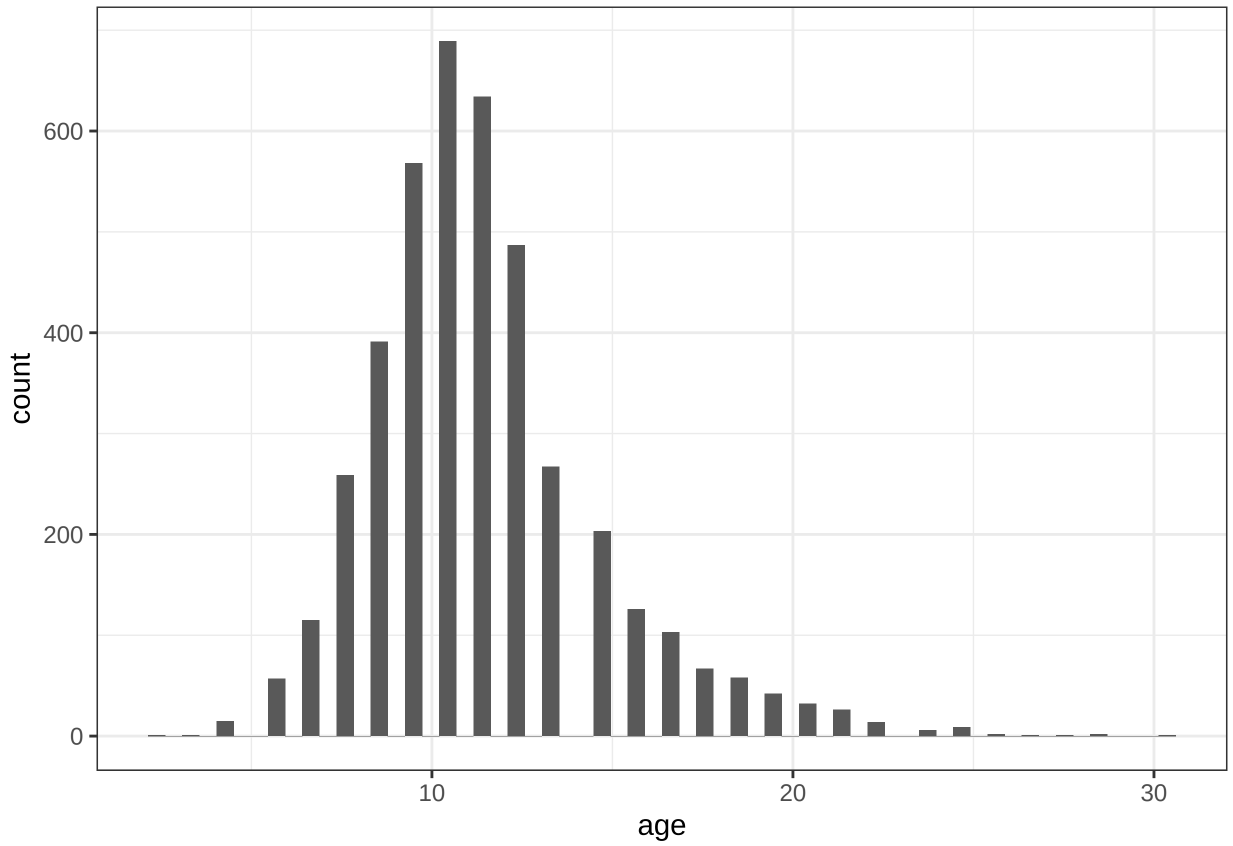
```
data = read.csv('homework-2/data/abalone.csv')
```

Question 1:

We first add a column “age” by adding 1.5 to the column “rings”.

```
data['age'] = data['rings']+1.5
```

```
data %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 60) +
  theme_bw()
```



The ages of most of the abalones in this dataset are from 6 to 14, and the histogram is right-skewed.

Question 2:

We split 80% data as training data and 20% data as testing data.

```
set.seed(3435)

data_split <- initial_split(data, prop = 0.80)
data_train <- training(data_split)
data_test <- testing(data_split)
```

Question 3:

We first create a new dataframe that does not include column "rings". Because the age of abalones can be calculated directly from rings data, and it will be meaningless to use other predictors in the linear regression model.

```
data_train_new = data_train[,c(10, 1, 2, 3, 4, 5, 6, 7, 8)]
head(data_train_new)
```

```
##       age type longest_shell diameter height whole_weight shucked_weight
## 2666  9.5    F         0.575    0.480  0.150       0.8970         0.4235
## 2584  9.5    F         0.530    0.405  0.150       0.8890         0.4055
## 2483 13.5    M         0.520    0.400  0.165       0.8565         0.2745
## 2524 10.5    F         0.565    0.440  0.150       0.9830         0.4475
## 2477 19.5    M         0.645    0.485  0.155       1.4890         0.5915
## 3860 10.5    F         0.570    0.440  0.190       1.0180         0.4470
##      viscera_weight shell_weight
## 2666         0.1905       0.2480
## 2584         0.2275       0.2150
## 2483         0.2010       0.2100
## 2524         0.2355       0.2485
## 2477         0.3120       0.3800
## 3860         0.2070       0.2650
```

```
data_recipe <- recipe(age ~ ., data = data_train_new) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight
                +longest_shell:diameter
                +shucked_weight:shell_weight) %>%
  step_center() %>%
  step_scale()
  #prep()

data_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight + longest_shell...
## Centering for <none>
## Scaling for <none>
```

Question 4:

Create and store a linear regression object.

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

Question 5:

Set up an empty workflow, and add model and recipe.

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

Question 6:

Fit the model.

```
lm_fit <- fit(lm_wflow, data_train_new)
lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy()
```

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                      3.93      0.696     5.64  1.80e- 8
##  2 longest_shell                    2.56      2.45      1.04  2.96e- 1
##  3 diameter                        25.1       3.25      7.74  1.27e-14
##  4 height                           5.24      1.64      3.19  1.42e- 3
##  5 whole_weight                     9.49      0.797    11.9   4.93e-32
##  6 shucked_weight                 -18.1       1.15    -15.7   7.49e-54
##  7 viscera_weight                  -8.01      1.44     -5.57  2.78e- 8
##  8 shell_weight                    12.2       1.57      7.79  9.15e-15
##  9 type_I                          -1.86      0.248    -7.49  8.63e-14
## 10 type_M                          -0.503     0.218    -2.31  2.11e- 2
## 11 type_I_x_shucked_weight          3.93      0.758     5.18  2.34e- 7
## 12 type_M_x_shucked_weight          1.04      0.447     2.34  1.95e- 2
## 13 longest_shell_x_diameter       -31.7       4.26     -7.44  1.23e-13
## 14 shucked_weight_x_shell_weight   -1.44      1.71     -0.840 4.01e- 1
```

Predict the age of a hypothetical female abalone.

```
result <- predict(lm_fit, data.frame(type = 'F', longest_shell=0.50, diameter=0.10,
                                     height=0.30, whole_weight=4, shucked_weight=1,
                                     viscera_weight=2, shell_weight=1))
result
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.3
```

Question 7:

Create a metric set.

```
data_metrics <- metric_set(rmse, rsq, mae)
```

Create a tibble of my model's predicted values from the training data along with the actual observed ages.

```
data_train_res <- predict(lm_fit, new_data = data_train %>% select(-age))
data_train_res <- bind_cols(data_train_res, data_train %>% select(age))
head(data_train_res)
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  11.7   9.5
## 2  11.2   9.5
## 3  13.4  13.5
## 4  11.6  10.5
## 5  14.0  19.5
## 6  12.5  10.5
```

```
data_metrics(data_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       0.550
## 3 mae     standard       1.56
```

The R^2 is 0.5423931, which means about 54.23931% of variance in the age of abalone can be explained by the inpependent variable such as abalone's gender, longest shell, diameter and so on.