PSTAT 131 HW2

Jiacong Wu

2022-10-17

```
library(tidymodels)
```

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

## v broom        1.0.1     v recipes      1.0.1
## v dials        1.0.0     v rsample      1.1.0
## v dplyr        1.0.9     v tibble       3.1.7
## v ggplot2      3.3.6     v tidyr        1.2.0
## v infer        1.0.3     v tune         1.0.0
## v modeldata    1.0.1     v workflows    1.1.0
## v parsnip      1.0.1     v workflowsets 1.0.0
## v purrr        0.3.4     v yardstick    1.1.0

## -- 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()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```
library(tidyverse)
```

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

## v readr   2.1.2     v forcats 0.5.2
## v stringr 1.4.0
## -- 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()
```

```
f = "abalone.csv"
```

```
aba_data = read.csv(file = f)
head(aba_data)
```

```
##   type longest_shell diameter height whole_weight shucked_weight viscera_weight
## 1    M         0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    M         0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    F         0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    M         0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    I         0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    I         0.425    0.300  0.095       0.3515         0.1410         0.0775
##   shell_weight rings
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

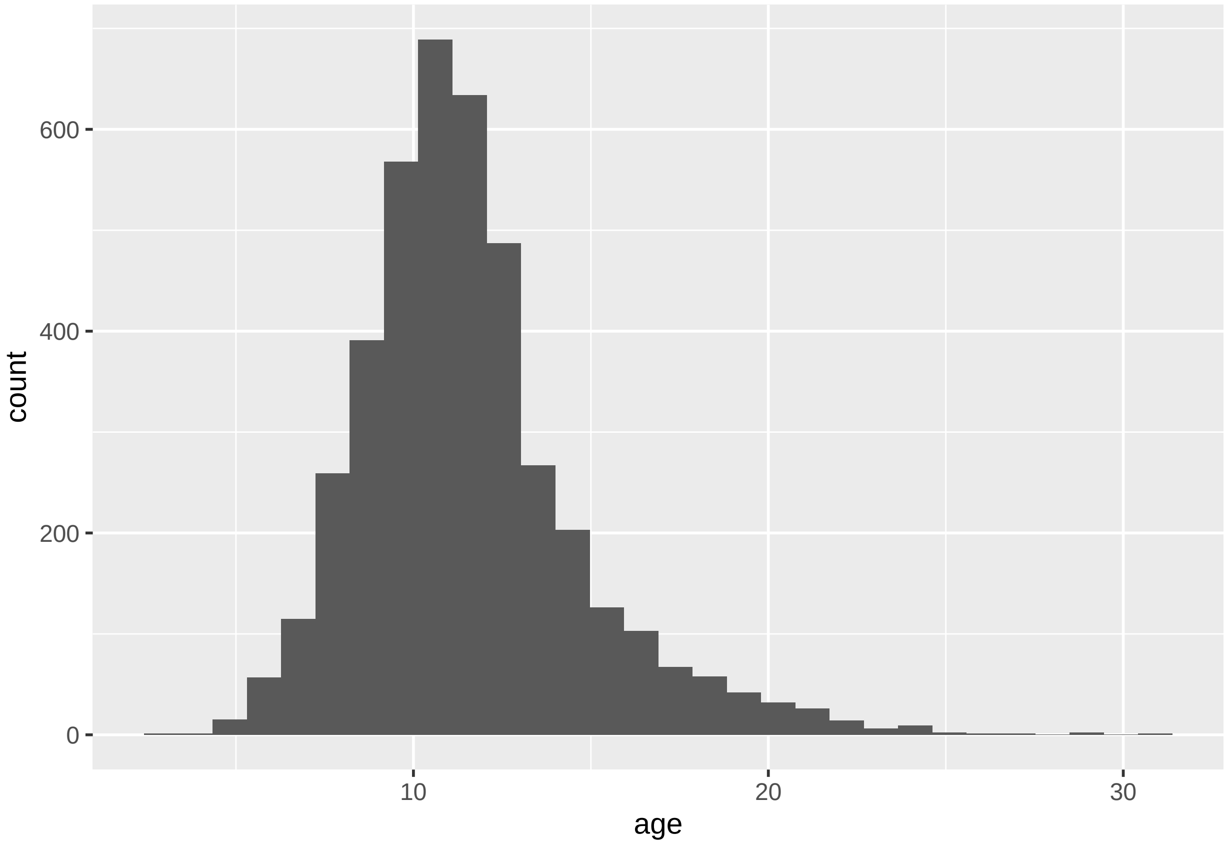
Question 1

```
aba_data$age <-aba_data$rings + 1.5
head(aba_data)
```

```
##   type longest_shell diameter height whole_weight shucked_weight viscera_weight
## 1    M         0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    M         0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    F         0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    M         0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    I         0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    I         0.425    0.300  0.095       0.3515         0.1410         0.0775
##   shell_weight rings  age
## 1        0.150    15 16.5
## 2        0.070     7  8.5
## 3        0.210     9 10.5
## 4        0.155    10 11.5
## 5        0.055     7  8.5
## 6        0.120     8  9.5
```

```
aba_data%>%
  ggplot(aes(x = age)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of age is left skewed, much of the mass of its distribution is at the lower end, majority of the abalones are aged less than 15.

Question 2

```
set.seed(1000)

aba_split <- initial_split(aba_data, prop = 0.75,
                                    strata = age)
aba_train <- training(aba_split)
aba_test <- testing(aba_split)
```

Question 3

Rings should not be included because the age is diredctly calculated from rings. There is a super strong correlation between the two variables. If rings is included in the predictors, the model will be overfit.

```
aba_recipe <-
  recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight + viscera_weight
  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_predictors())%>%
  step_scale(all_predictors())
aba_recipe
```

```
## Recipe
##
```

```
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## 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 for all_predictors()
## Scaling for all_predictors()
```

Question 4

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

Question 5

```
aba_lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(aba_recipe)
```

Question 6

```
lm_fit <- fit(aba_lm_wflow, aba_train)
lm_fit
```

```
## == Workflow [trained] ===========================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor -----------------------------------------------------------------
## 6 Recipe Steps
##
## * step_dummy()
## * step_interact()
## * step_interact()
## * step_interact()
## * step_center()
## * step_scale()
##
## -- Model ------------------------------------------------------------------------
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##                   (Intercept)                  longest_shell
##                      11.42558                        0.27720
##                      diameter                         height
##                       2.26024                        0.24242
##                  whole_weight                 shucked_weight
##                       5.19864                       -4.51219
##                viscera_weight                   shell_weight
##                      -1.06043                        1.43975
##                        type_I                         type_M
##                      -0.95225                       -0.32397
##       type_I_x_shucked_weight        type_M_x_shucked_weight
##                       0.51578                        0.38995
##      longest_shell_x_diameter  shucked_weight_x_shell_weight
##                      -2.56822                       -0.04577
```

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

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  21.1
```

Question 7

```
library(yardstick)
aba_metric = metric_set(rsq,rmse,mae)
aba_train_result = predict(lm_fit,aba_train %>% select(-age,-rings))
aba_train_result =bind_cols(aba_train_result,aba_train %>% select(age))
head(aba_train_result)
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  8.15   8.5
## 2  9.34   9.5
## 3 10.4    8.5
## 4 10.0    9.5
## 5 11.0    9.5
## 6  6.35   6.5
```

```
aba_metrics = aba_metric(aba_train_result,truth =age, estimate = .pred)
aba_metrics
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.558
## 2 rmse    standard       2.15
## 3 mae     standard       1.55
```