

PSTAT 131 HW 2

Raymond Lee

2022-04-07

```
abalone = read.csv("abalone.csv")

library(tidyverse)

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

## v ggplot2 3.3.5      v purrr 0.3.4
## v tibble 3.1.6       v dplyr 1.0.8
## v tidyr 1.2.0        v stringr 1.4.0
## v readr 2.1.2        v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()

library(tidymodels)

## -- Attaching packages ----- tidymodels 0.2.0 --

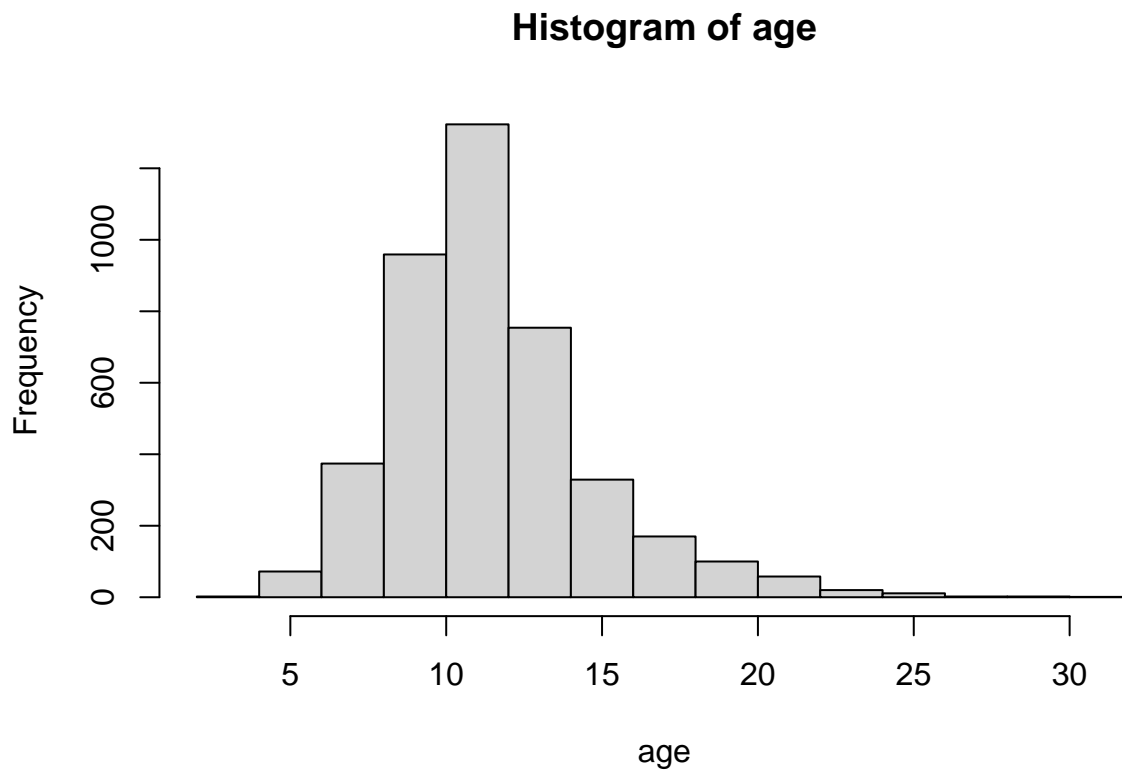
## v broom      0.7.12      v rsample      0.1.1
## v dials      0.1.0       v tune         0.2.0
## v infer      1.0.0       v workflows    0.2.6
## v modeldata  0.1.1       v workflowsets 0.2.1
## v parsnip    0.2.1       v yardstick    0.0.9
## v recipes    0.2.0

## -- 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()
## * Dig deeper into tidy modeling with R at https://www.tmw.org

tidymodels_prefer()
```

1.

```
age = abalone$rings+1.5
abalone["age"] = age
hist(age)
```



Age appears to be slightly skewed right.

2.

```
set.seed(1114)
abalone_split = initial_split(abalone, prop = .80, strata = age)
abalone_train = training(abalone_split)
abalone_test = testing(abalone_split)
```

3. We should not use rings to predict age because age is simply a transformation of rings.

```
abalone_recipe = recipe(age ~ type + longest_shell + diameter + height +
                        whole_weight + shucked_weight + viscera_weight +
                        shell_weight, data = abalone_train) %>%
  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())
```

4.

```
lm_mod = linear_reg() %>%
  set_engine("lm")
```

5.

```
lm_wflow = workflow() %>%
  add_model(lm_mod) %>%
  add_recipe(abalone_recipe)
```

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      0.0373   307.      0
## 2 longest_shell      0.388     0.286     1.36 1.75e- 1
## 3 diameter           1.98     0.316     6.27 3.98e-10
## 4 height             0.550     0.0962    5.72 1.17e- 8
## 5 whole_weight       5.22     0.399    13.1 3.46e-38
## 6 shucked_weight     -4.53     0.262   -17.3 2.84e-64
## 7 viscera_weight     -1.00     0.155    -6.44 1.37e-10
## 8 shell_weight        1.29     0.220     5.87 4.86e- 9
## 9 type_I             -0.898     0.116    -7.74 1.26e-14
##10 type_M             -0.253     0.104    -2.44 1.47e- 2
##11 type_I_x_shucked_weight 0.495     0.0874    5.66 1.63e- 8
##12 type_M_x_shucked_weight 0.304     0.109     2.78 5.47e- 3
##13 longest_shell_x_diameter -2.54     0.404    -6.29 3.66e-10
##14 shucked_weight_x_shell_weight -0.0634    0.205    -0.309 7.57e- 1
```

```
predict_values = data.frame(type = "F", longest_shell = .5, diameter = .1, height = .3, whole_weight = 4,
                             shucked_weight = 1, viscera_weight = 2, shell_weight = 1)
prediction = predict(lm_fit, predict_values)
prediction
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.1
```

The predicted age of the hypothetical abalone is 23.10256.

7.

```
library(yardstick)

set = metric_set(rsq, rmse, mae)

predicted_age = predict(lm_fit, abalone_train)
predicted_actual_ages = tibble(bind_cols(abalone_train["age"], predicted_age))
predicted_actual_ages
```

```
## # A tibble: 3,340 x 2
##   age .pred
##   <dbl> <dbl>
## 1  8.5  9.43
## 2  8.5  8.05
## 3  8.5 10.2
## 4  9.5  9.93
## 5  6.5  6.13
## 6  6.5  5.76
## 7  5.5  5.87
## 8  8.5  8.57
## 9  8.5 11.7
## 10 7.5  7.61
## # ... with 3,330 more rows
```

```
set(predicted_actual_ages, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>         <dbl>
## 1 rsq     standard       0.557
## 2 rmse    standard       2.15
## 3 mae     standard       1.55
```

The R^2 is 0.5570949, the RMSE is 2.1507378, and the MAE is 1.5543719. About 55.7% of the variability in age can be explained using type, longest_shell, diameter, height, whole_weight, shucked_weight, viscera_weight, and shell_weight.