131-hw2

Russell Liu

2022-10-15

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ purrr   0.3.4 
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1 
## ✔ readr   2.1.2      ✔ forcats 0.5.2 
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

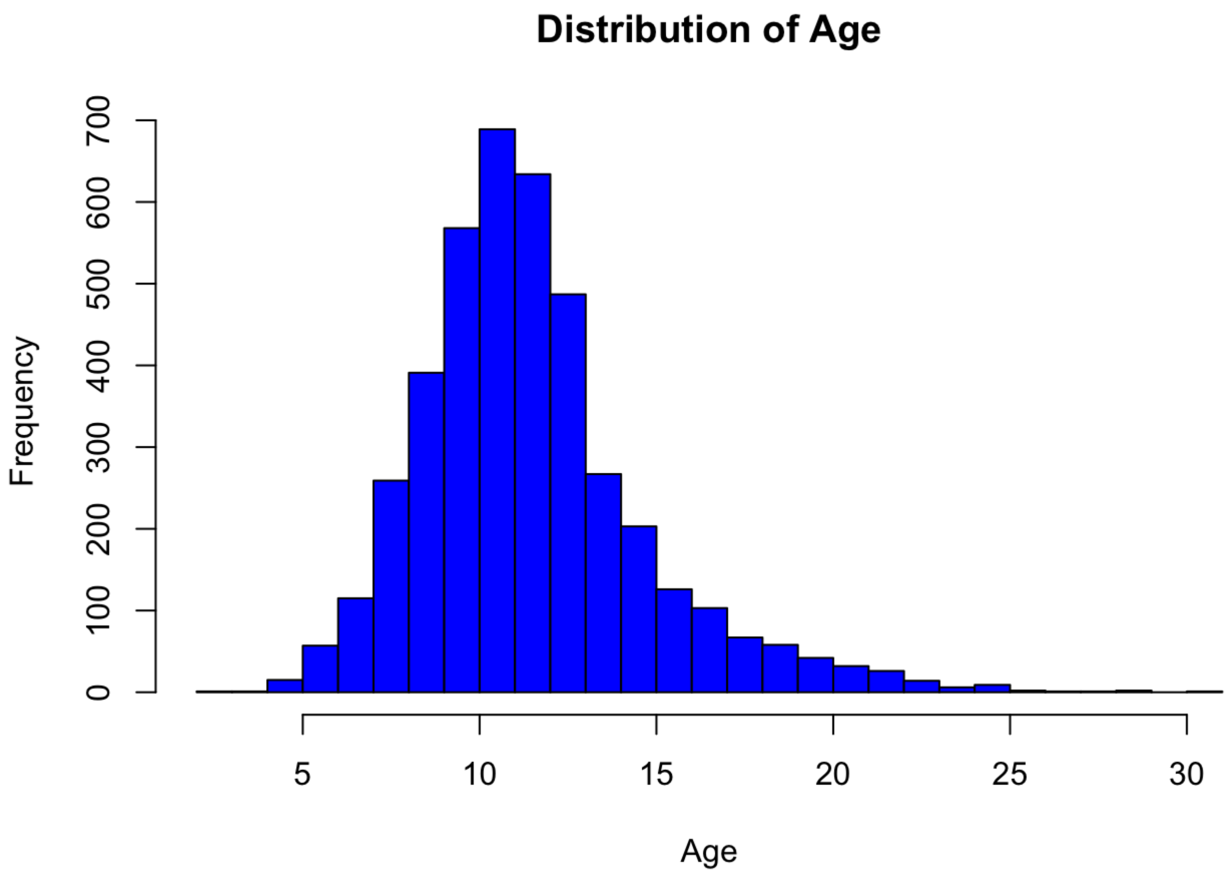
```
## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──
## ✔ broom        1.0.1     ✔ rsample      1.1.0
## ✔ dials        1.0.0     ✔ tune         1.0.0
## ✔ infer        1.0.3     ✔ workflows    1.1.0
## ✔ modeldata    1.0.1     ✔ workflowsets 1.0.0
## ✔ parsnip    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

```
abalone <- read.csv(file = '/Users/liusenyuan/Desktop/PSTAT131/HW/HW2-131/data/abalone.csv')
view(abalone)
```

Questuon 1

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

```
hist(abalone$age, xlab = "Age",breaks =30,
     main = "Distribution of Age", col = 'blue')
```



we can see that it is not evenly distributed while sckewed to left around 10-15.

Question 2

```
set.seed(3435)

abalone_split <- initial_split(abalone, prop = 0.70,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
abalone_recipe <- abalone_train %>%
  recipe(age ~ type+longest_shell+diameter+height+whole_weight+
           shucked_weight+viscera_weight+shell_weight) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight+longest_shell:
                  diameter+shucked_weight:shell_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_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)
abalone_hypo <- data.frame(type = "F",longest_shell = 0.50,
                           diameter = 0.10,
                           height = 0.30,
                           whole_weight = 4,
                           shucked_weight = 1,
                           shell_weight = 1,
                           viscera_weight = 2)
predict(lm_fit, new_data = abalone_hypo)
```

```
## # A tibble: 1 × 1
##   .pred
##   <dbl>
## 1  22.2
```

Question 7

```
#1
abalone_metrics <- metric_set(rsq, rmse, mae)
#2
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 × 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.81   8.5
## 2 10.4    8.5
## 3 10.1    9.5
## 4 11.0    9.5
## 5  5.77   6.5
## 6  5.92   5.5
```

```
#3
abalone_metrics(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 3 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.558
## 2 rmse    standard       2.11
## 3 mae     standard       1.52
```

We can see that R2 score is only 0.55 which we can interpret that the model is not very accurate. RMSE is also a little high for an error and MAE refelcts that there is 1.51 difference between prediction and the true value.