Homework 2

Richard Zhao

2022-10-24

```
library(tidyverse)
library(ggplot2)
library(tidymodels)
```

Question 1:

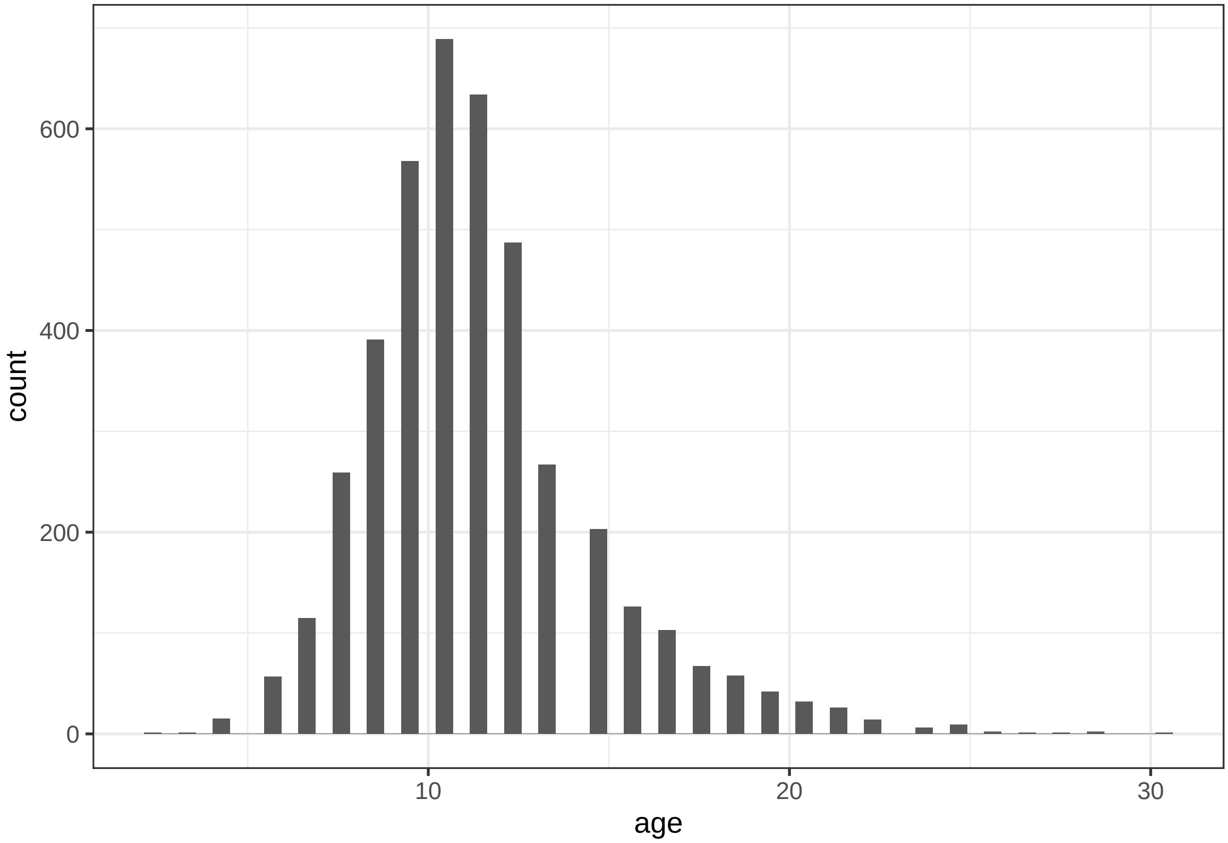
```
abalone <- read_csv("homework-2/data/abalone.csv")
abalone
```

```
## # A tibble: 4,177 x 9
##    type  longest_shell diameter height whole_weight shucked_weight
##    <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>
##  1 M             0.455    0.365  0.095        0.514         0.224
##  2 M             0.35     0.265  0.09         0.226         0.0995
##  3 F             0.53     0.42   0.135        0.677         0.256
##  4 M             0.44     0.365  0.125        0.516         0.216
##  5 I             0.33     0.255  0.08         0.205         0.0895
##  6 I             0.425    0.3    0.095        0.352         0.141
##  7 F             0.53     0.415  0.15         0.778         0.237
##  8 F             0.545    0.425  0.125        0.768         0.294
##  9 M             0.475    0.37   0.125        0.509         0.216
## 10 F             0.55     0.44   0.15         0.894         0.314
## # ... with 4,167 more rows, and 3 more variables: viscera_weight <dbl>,
## #   shell_weight <dbl>, rings <dbl>
```

```
abalone2 <- mutate(abalone, age = rings + 1.5)
abalone2
```

```
## # A tibble: 4,177 x 10
##    type  longest_shell diameter height whole_weight shucked_weight
##    <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>
##  1 M             0.455    0.365  0.095        0.514         0.224
##  2 M             0.35     0.265  0.09         0.226         0.0995
##  3 F             0.53     0.42   0.135        0.677         0.256
##  4 M             0.44     0.365  0.125        0.516         0.216
##  5 I             0.33     0.255  0.08         0.205         0.0895
##  6 I             0.425    0.3    0.095        0.352         0.141
##  7 F             0.53     0.415  0.15         0.778         0.237
##  8 F             0.545    0.425  0.125        0.768         0.294
##  9 M             0.475    0.37   0.125        0.509         0.216
## 10 F             0.55     0.44   0.15         0.894         0.314
## # ... with 4,167 more rows, and 4 more variables: viscera_weight <dbl>,
## #   shell_weight <dbl>, rings <dbl>, age <dbl>
```

```
ggplot(abalone2, aes(x = age)) + geom_histogram(bins = 60) + theme_bw()
```



Age looks slightly skewed right, with a peak at 10.5 years. Most abalones in the data set are younger than 20 years of age.

Question 2:

```
set.seed(1208)
abalone2_split <- initial_split(abalone2, prop = 0.70, strata = age)
abalone2_train <- training(abalone2_split)
abalone2_test <- testing(abalone2_split)
```

Question 3:

```
abalone2_recipe <- recipe(age ~ ., data = abalone2_train) %>%
  step_rm(rings) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight + longest_shell:diameter + shucked_weight:s
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
```

We shouldn’t use rings to predict age because we used the rings variable to create the age variable by adding 1.5 years to the number of rings on an abalone. Including rings to predict age wouldn’t give us any important information.

Question 4:

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

Question 5:

```
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone2_recipe)
```

Question 6:

```
lm_fit <- fit(lm_wflow, abalone2_train)
abalone_example <- tibble(type = "F", longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weigh
predict(lm_fit, new_data = abalone_example)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  24.6
```

We predict the age of this hypothetical female abalone be 24.6 years.

Question 7:

```
abalone2_metric <- metric_set(rmse, rsq, mae)
abalone2_predict <- predict(lm_fit, abalone2_train) %>% bind_cols(abalone2_train %>% select(age))
abalone2_metric(abalone2_predict, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.17
## 2 rsq     standard       0.557
## 3 mae     standard       1.57
```

The R squared value is 0.5560139, which means that only 55.60139% of variation in abalone age is explained by our model. The relationship between the predictors and age isn't very linear, which might explain this R squared value.