# Pstat 131 Homework 2

Yu Tian

Spring 2022-04-10

## Linear Regression

```
# Read the full abalone data set into R using read_csv()
abalone <- read.csv(file = 'abalone.csv')
abalone %>% head()
```

```
##   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

Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no age variable in the data set. Add age to the data set.

Assess and describe the distribution of age.

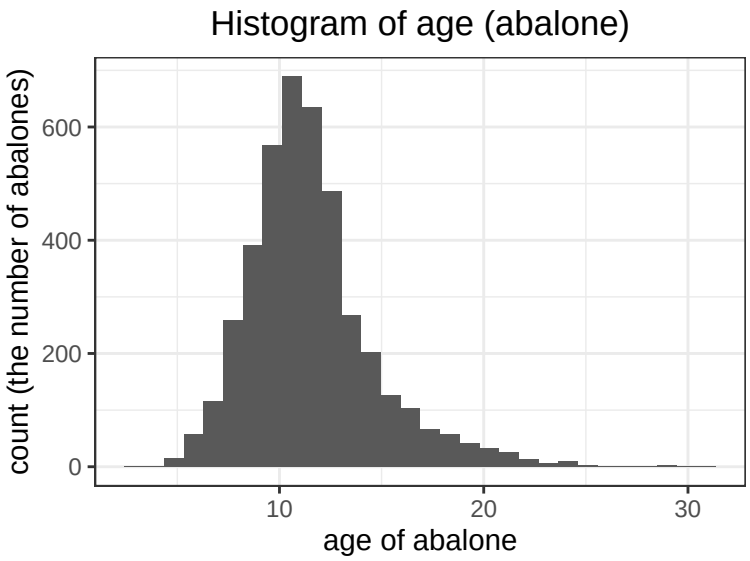
**Answer:** Q1

```
# create the variable age in the data set
abalone["age"] <- abalone$rings + 1.5
head(abalone)
```

```
##   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
```

```
#the distribution of age
abalone %>%
  ggplot(aes(x=age)) +
  geom_histogram(bins=30)+
  labs(title="Histogram of age (abalone)",
       x ="age of abalone", y = "count (the number of abalones)") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```




From the histogram, we can find that the distribution is slight left-skewed, and most abalones' age are distributed around 10. There is a few abalones with age higher than 25.

## Question 2

Split the abalone data into a training set and a testing set. Use stratified sampling. You should decide on appropriate percentages for splitting the data.

Remember that you'll need to set a seed at the beginning of the document to reproduce your results.

**Answer:** Q2

```
# set a seed
set.seed(0623)

# split the abalone data into a training set and a testing set.
abalone_split <- initial_split(abalone, prop = 0.80, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
dim(abalone)
```

```
## [1] 4177   10
```

```
dim(abalone_train)
```

```
## [1] 3340   10
```

```
dim(abalone_test)
```

```
## [1] 837  10
```

## Question 3

Using the training data, create a recipe predicting the outcome variable, age, with all other predictor variables. Note that you should not include rings to predict age. Explain why you shouldn't use rings to predict age.

Steps for your recipe:

dummy code any categorical predictors

create interactions between

type and shucked_weight, longest_shell and diameter, shucked_weight and shell_weight center all predictors, and scale all predictors.

You'll need to investigate the tidymodels documentation to find the appropriate step functions to use.

**Answer:** Q3

We shouldn't use rings to predict age since the new variable "age" is directly calculated from variable "rings" by ring*1.5=age. They are related.

```
# remove variable ring
new_abalone_train <- abalone_train %>% select(-rings)
# confirm remove rings and show the new data set
head(new_abalone_train)
```

```
##    type longest_shell diameter height whole_weight shucked_weight
## 6     I         0.425    0.300  0.095       0.3515         0.1410
## 17    I         0.355    0.280  0.085       0.2905         0.0950
## 19    M         0.365    0.295  0.080       0.2555         0.0970
## 36    M         0.465    0.355  0.105       0.4795         0.2270
## 38    F         0.450    0.355  0.105       0.5225         0.2370
## 43    I         0.240    0.175  0.045       0.0700         0.0315
##    viscera_weight shell_weight age
## 6          0.0775        0.120 9.5
## 17         0.0395        0.115 8.5
## 19         0.0430        0.100 8.5
## 36         0.1240        0.125 9.5
## 38         0.1165        0.145 9.5
## 43         0.0235        0.020 6.5
```

```
abalone_recipe <- recipe(age ~ ., data = new_abalone_train) %>%
  # dummy code any categorical predictors
  step_dummy(all_nominal_predictors()) %>%
  # create interactions between type and shucked_weight
  step_interact(terms = ~starts_with("type"):shucked_weight) %>%
  # create interactions between longest_shell and diameter
  step_interact(terms = ~longest_shell:diameter) %>%
  # create interactions between shucked_weight and shell_weight
  step_interact(terms = ~shucked_weight:shell_weight) %>%
  # center all predictors
  step_center(all_predictors()) %>%
  # scale all predictors
```

```
  step_scale(all_predictors())

abalone_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

Create and store a linear regression object using the "lm" engine.

**Answer:** Q4

```
# create and store a linear regression object
# using the "lm" engine.
lm_model <- linear_reg() %>%
  set_engine("lm")
```

## Question 5

Now:

set up an empty workflow, add the model you created in Question 4, and add the recipe that you created in Question 3.

**Answer:** Q5

```
lm_wflow <- workflow() %>%       # set up an empty workflow
  add_model(lm_model) %>%                # add the model i created in Question 4
  add_recipe(abalone_recipe)             # add the recipe that i created in Question 3
```

## Question 6

Use your fit() object to predict the age of a hypothetical female abalone with longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4, shucked_weight = 1, viscera_weight = 2, shell_weight = 1.

**Answer:** Q6

```
lm_fit <- fit(lm_wflow, new_abalone_train)
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)                     11.4       0.0372   308.    0
##  2 longest_shell                    0.507     0.284      1.79  7.42e- 2
##  3 diameter                         1.90      0.311      6.10  1.15e- 9
##  4 height                           0.251     0.0697     3.60  3.18e- 4
##  5 whole_weight                     5.57      0.399     14.0   3.29e-43
##  6 shucked_weight                  -4.69      0.252    -18.6   6.72e-74
##  7 viscera_weight                  -0.984     0.159     -6.20  6.37e-10
##  8 shell_weight                     1.33      0.212      6.25  4.57e-10
##  9 type_I                          -1.06      0.117     -9.09  1.66e-19
## 10 type_M                          -0.247     0.104     -2.39  1.71e- 2
## 11 type_I_x_shucked_weight          0.620     0.0883     7.02  2.65e-12
## 12 type_M_x_shucked_weight          0.279     0.108      2.57  1.01e- 2
## 13 longest_shell_x_diameter        -2.61      0.397     -6.56  6.09e-11
## 14 shucked_weight_x_shell_weight   -0.0248    0.203     -0.122 9.03e- 1
```

```
predict_abalone <- data.frame(type = 'F', longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_w

predict_abalone_res <- predict(lm_fit, new_data = predict_abalone)
predict_abalone_res
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  24.8
```

the predicted age of a hypothetical female abalone is 24.17389

## Question 7

Now you want to assess your model's performance. To do this, use the yardstick package:

Create a metric set that includes R2, RMSE (root mean squared error), and MAE (mean absolute error). Use predict() and bind_cols() to create a tibble of your model's predicted values from the training data along with the actual observed ages (these are needed to assess your model's performance). Finally, apply your metric set to the tibble, report the results, and interpret the R2 value.

**Answer:** Q7

```
# Use predict() to create a tibble of my model's predicted values from the training data along with the
abalone_train_res <- predict(lm_fit, new_data = new_abalone_train %>% select(-age))
abalone_train_res
```

```
## # A tibble: 3,340 x 1
##    .pred
##    <dbl>
##  1  9.32
##  2  9.69
##  3 10.5
##  4 10.1
```

```
##  5 10.9
##  6  6.28
##  7  5.80
##  8  5.95
##  9  8.44
## 10 11.8
## # ... with 3,330 more rows
```

```
# Use bind_cols() to create a tibble of my model's predicted values from the training data along with t
abalone_train_res <- bind_cols(abalone_train_res, new_abalone_train %>% select(age))
abalone_train_res
```

```
## # A tibble: 3,340 x 2
##    .pred   age
##    <dbl> <dbl>
##  1  9.32   9.5
##  2  9.69   8.5
##  3 10.5    8.5
##  4 10.1    9.5
##  5 10.9    9.5
##  6  6.28   6.5
##  7  5.80   6.5
##  8  5.95   5.5
##  9  8.44   8.5
## 10 11.8    8.5
## # ... with 3,330 more rows
```

```
# create a metric set that includes R2, RMSE (root mean squared error), and MAE (mean absolute error)
abalone_metrics <- metric_set(rsq, rmse, mae)

# apply my metric set to the tibble
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.565
## 2 rmse    standard       2.14
## 3 mae     standard       1.53
```

In my model,

the $R2$ is 0.5584123 (~ 0.56)

the $RMSE$ is 2.1607721 (~2.16)

the $MAE$ is 1.5509346 (~1.55)

Thus, the R2 value means 56 percentage of the variance for response (age) variable that's explained by predictor variables in the linear regression model.