Homework 2

Cyril Wang, PSTAT 131/231

Contents

Linear Regression

For this lab, we will be working with a data set from the UCI (University of California, Irvine) Machine Learning repository (see website here). The full data set consists of $4,177$ observations of abalone in Tasmania. (Fun fact: Tasmania supplies about 25% of the yearly world abalone harvest.)

Figure 1: *Fig 1. Inside of an abalone shell.*

The age of an abalone is typically determined by cutting the shell open and counting the number of rings with a microscope. The purpose of this data set is to determine whether abalone age (**number of rings + 1.5**) can be accurately predicted using other, easier-to-obtain information about the abalone.

The full abalone data set is located in the `\data` subdirectory. Read it into *R* using `read_csv()`. Take a moment to read through the codebook (`abalone_codebook.txt`) and familiarize yourself with the variable definitions.

Make sure you load the `tidyverse` and `tidymodels`!

Question 1

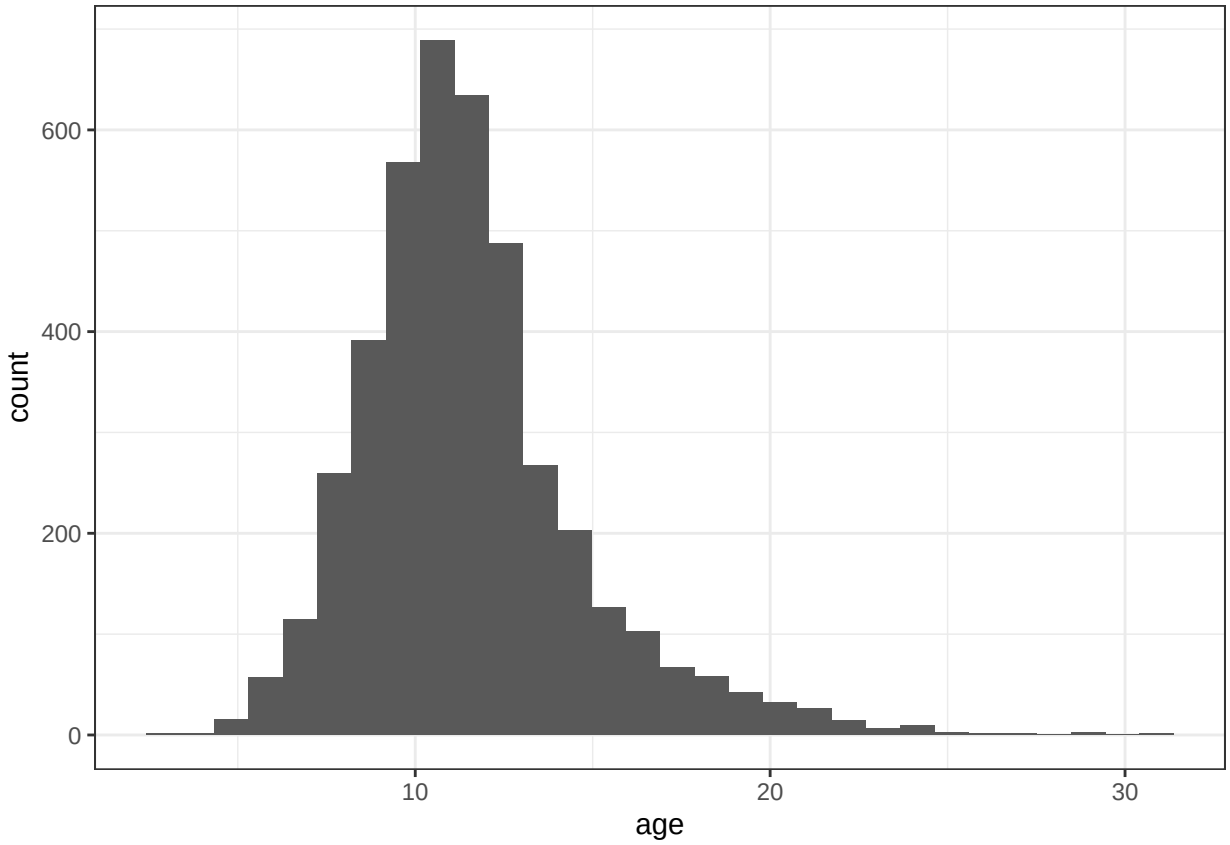
Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no `age` variable in the data set. Add `age` to the data set.

Assess and describe the distribution of `age`.

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

```
# distribution of 'age'
abalone %>% ggplot(aes(x = age)) +
  geom_histogram(bins = 30) +
  theme_bw()
```



The distribution of age seems to be unimodal, and appears to be somewhat symmetric, albeit slightly right skewed (or positively skewed). The distribution is centered around 11 or 12, and most of abalone have an age around 11 or 12, with a few having an age above 20.

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.

```
set.seed(123)
abalone_split <- initial_split(abalone, prop=0.7, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

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:

1. dummy code any categorical predictors
2. create interactions between
    - `type` and `shucked_weight`,
    - `longest_shell` and `diameter`,
    - `shucked_weight` and `shell_weight`
3. center all predictors, and
4. scale all predictors.

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

```
abalone_recipe <- recipe(age ~ type + longest_shell + diameter + height + whole_weight +
                           shucked_weight + viscera_weight + shell_weight, data = abalone) %>%
                  step_dummy(all_nominal_predictors()) %>%
                  step_interact(terms = ~type:shucked_weight) %>%
                  step_interact(terms = ~longest_shell:diameter) %>%
                  step_interact(terms = ~shucked_weight:shell_weight) %>%
                  step_center() %>%
                  step_scale()
```

We shouldn't use rings to predict age because we calculated the age variable using rings + 1.5, so the two variables will be highly correlated (i.e. a correlation of 1), so that can lead to misleading estimates of the association between age and the other predictors.

Question 4

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

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

Question 5

Now:

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

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

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.

```
lm_fit <- fit(lm_wflow, abalone_train)
predict(lm_fit, new_data = data.frame(type='F', longest_shell = 0.50, diameter = 0.10, height = 0.30, wh
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.4
```

Question 7

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

1. Create a metric set that includes R^2, RMSE (root mean squared error), and MAE (mean absolute error).
2. 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).
3. Finally, apply your metric set to the tibble, report the results, and interpret the R^2 value.

```
library(yardstick)
# 1
abalone_metrics <- metric_set(rsq, rmse, mae)

# 2
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-c(rings, age)))

abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
head(abalone_train_res)
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  8.19   8.5
## 2 10.1    8.5
```

```
## 3  9.93   9.5
## 4  6.15   6.5
## 5  5.73   6.5
## 6  5.86   5.5
```

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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.561
## 2 rmse    standard       2.12
## 3 mae     standard       1.53
```

Our R^2 value came out to be 0.5604051, which means that around 0.56 of the variance in the response can be explained by our model.