Homework 2

PSTAT 131 John Wei

Contents

Linear Regression 1

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.)

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`!

```
library(tidyverse)
library(tidymodels)
library(tinytex)
```

```
abalone <- read_csv("abalone.csv")
```

```
summary(abalone)
```

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
```

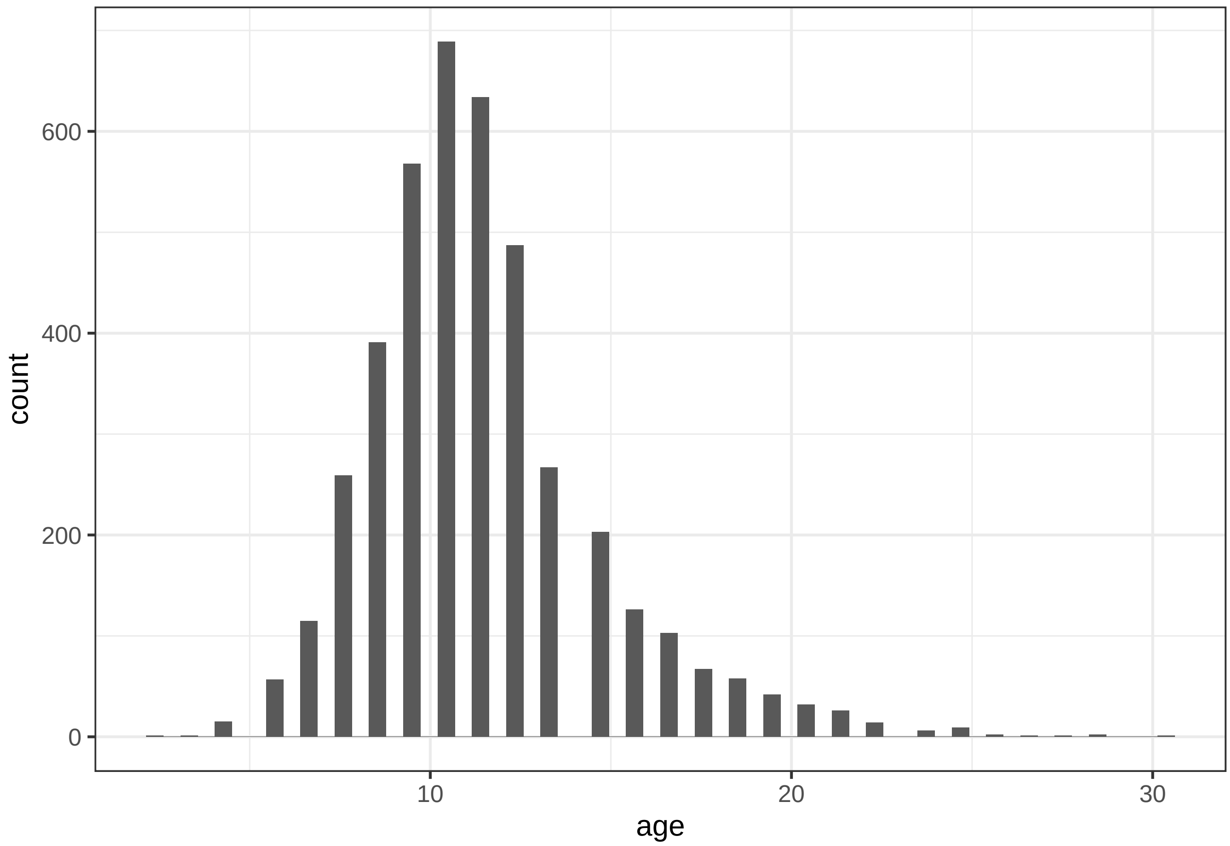
```
## # A tibble: 4,177 x 9
##    type  longest_shell diameter height whole_weight shucked_weight
##    <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>
##  1 M             0.455    0.365  0.095        0.514         0.224
##  2 M             0.35     0.265  0.09         0.226         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>
```

```
abalone$age <- abalone$rings + 1.5
summary(abalone)
```

```
##      type           longest_shell      diameter          height
##  Length:4177        Min.   :0.075   Min.   :0.0550   Min.   :0.0000
##  Class :character   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150
##  Mode  :character   Median :0.545   Median :0.4250   Median :0.1400
##                     Mean   :0.524   Mean   :0.4079   Mean   :0.1395
##                     3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650
##                     Max.   :0.815   Max.   :0.6500   Max.   :1.1300
##   whole_weight    shucked_weight   viscera_weight    shell_weight
##  Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015
##  1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300
##  Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340
##  Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388
##  3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290
##  Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050
##      rings             age
##  Min.   : 1.000   Min.   : 2.50
##  1st Qu.: 8.000   1st Qu.: 9.50
##  Median : 9.000   Median :10.50
##  Mean   : 9.934   Mean   :11.43
##  3rd Qu.:11.000   3rd Qu.:12.50
##  Max.   :29.000   Max.   :30.50
```

```
abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 60) +
  theme_bw()
```



By creating a histogram of age in abalone, we see a normal distribution with a longer tail on the right.

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.

```
set.seed(4167)
```

```
abalone_split <- initial_split(abalone, prop = 0.80,
                               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`.

```
abalone_recipe <-
  recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight + viscera_weigh
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
  step_interact(terms = ~ shucked_weight:shell_weight) %>%
```

```
  step_scale(all_predictors()) %>%
  step_center(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
## Scaling for all_predictors()
## Centering for all_predictors()
```

We should not use rings to predict age because rings is essentially the outcome variable plus an intercept (1.5). We don't want our outcome variable to also be a predictor variable; the model would heavily depend on rings.

Question 4

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

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

Question 5

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)
becka <- data.frame(type = 'F', longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4,
predict(lm_fit, new_data = becka)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.3
```

Using fit, with the given predictor variables we obtain a prediction of 20.87 years.

Question 7

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

```
install.packages("yardstick")
```

```
library('yardstick')
```

1. Create a metric set that includes R^2, RMSE (root mean squared error), and MAE (mean absolute error).

```
abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.15
## 2 rsq     standard       0.561
## 3 mae     standard       1.55
```

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).

```
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 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  8.05   8.5
## 2  9.28   9.5
## 3  9.73   8.5
## 4 10.3    8.5
## 5 10.9    9.5
## 6  6.17   6.5
```

3. Finally, apply your metric set to the tibble, report the results, and interpret the R^2 value.

```
rmse(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.15
```

```
rsq(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.561
```

```
mae(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 mae     standard        1.55
```

The R squared value is approximately .56. We can interpret this as 56% of the data fits the regression model; 56% in variation of age can be described by the model excluding rings.