PSTAT131_HW2

Tao Wang

2022-04-08

```r
# load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.2.0 --
## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```r
# read data by using read_csv()
abalone <- read_csv('/Users/wangtao/Desktop/PSTAT 131/PSTAT131_HW2/data/abalone.csv')
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 1

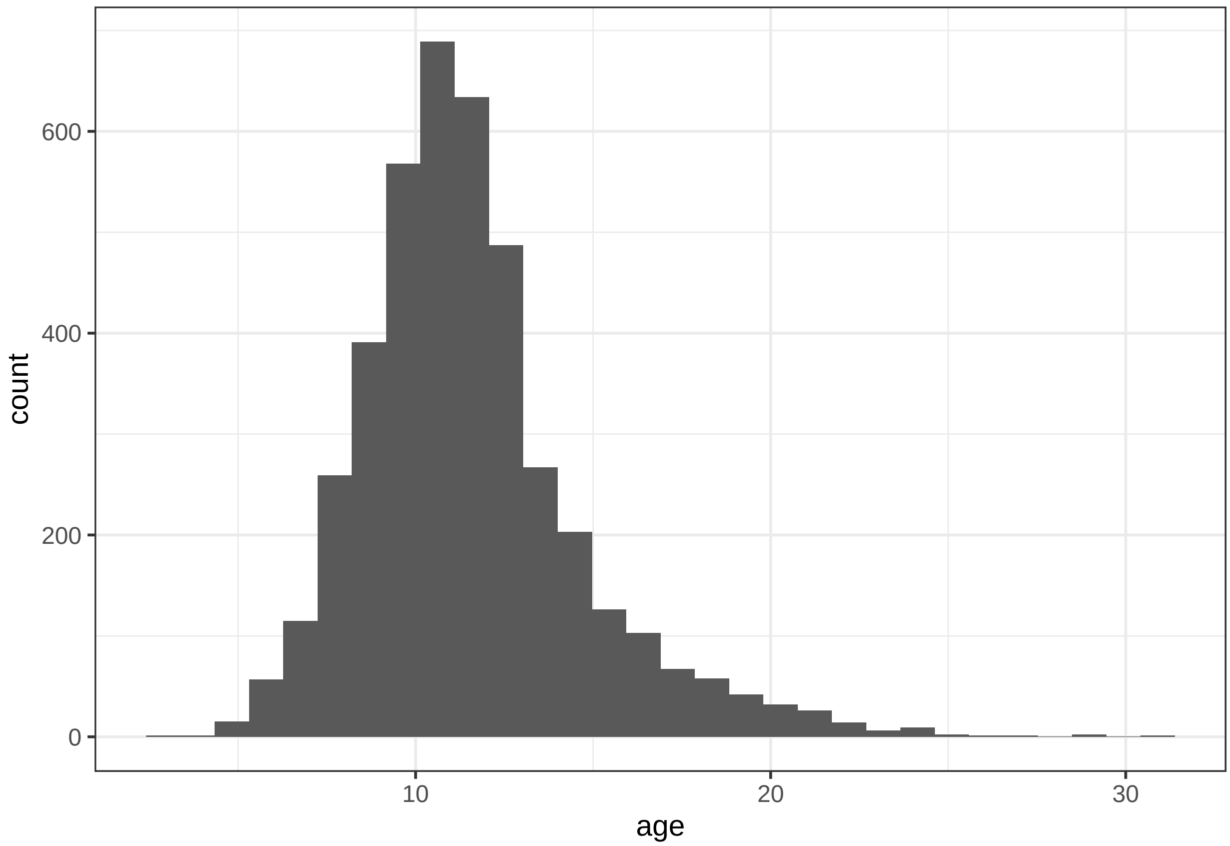
- Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no age variable in the data set. Add age to the data set.
- Assess and describe the distribution of age.

```
# notice: according to the question age = rings + 1.5

# 1.create a new data dataset with variable age

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

# 2.use head() function to check the first few rows of our new dataset
head(new_abalone)
```

```
## # A tibble: 6 x 10
##   type  longest_shell diameter height whole_weight shucked_weight viscera_weight
##   <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M             0.455    0.365  0.095        0.514         0.224          0.101
## 2 M             0.35     0.265  0.09         0.226         0.0995         0.0485
## 3 F             0.53     0.42   0.135        0.677         0.256          0.142
## 4 M             0.44     0.365  0.125        0.516         0.216          0.114
## 5 I             0.33     0.255  0.08         0.205         0.0895         0.0395
## 6 I             0.425    0.3    0.095        0.352         0.141          0.0775
## # ... with 3 more variables: shell_weight <dbl>, rings <dbl>, age <dbl>
```

```
# 3.Assess the the distribution of age by making a histogram
new_abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram()+
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Answer:

- According to the histogram we made here, we can know that the shape of the distribution of age is very similar to the normal distribution. However, they are not the same since the shape distribution of age is a little bit skewed. (notice, we have a small tail to the right)
- The maximum count occurs when age is around 12-13.
- Most age in the data set are less than 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 a seed
set.seed(1234)

# Data Splitting
abalone_split <- initial_split(new_abalone, prop = 0.80,
                                strata = age)
# traning set
abalone_train <- training(abalone_split)

# test set
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.

```
# set up:
# create a recipe predicting the outcome variable, `age`, with all other predictor variables
# notice: we shouldn't use `rings` to predict `age`
abalone_recipe <- recipe(age ~ type +longest_shell + diameter + height +
           whole_weight + shucked_weight + viscera_weight +
           shell_weight,data = abalone_train) %>%
# 1.  dummy code any categorical predictors
  step_dummy(all_nominal_predictors()) %>%
# 2.  create interactions between
# `type` and `shucked_weight`,`longest_shell` and `diameter`,`shucked_weight` and `shell_weight`
step_interact(terms = ~ shucked_weight:starts_with("type")) %>%
  step_interact(terms = ~ longest_shell:diameter )  %>%
  step_interact(terms = ~ shucked_weight:shell_weight ) %>%
# 3.  center all predictors, and
step_center(all_predictors()) %>%
# 4.  scale all predictors.
step_scale(all_predictors())
```

Answer: Explain why you shouldn't use `rings` to predict `age`.

- First, according to the question, 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. Since our goal is to find another information that is easier to obtain than the 'rings' to predict the abalone 'age', then we can know that we shouldn't use `rings` to predict `age`.
- Second, we have already known that 'age' = 'rings' + 1.5, and the variable 'age' is dependent on the variable 'rings'. So, it won't be meaningful if we use `rings` to predict `age`.

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.

```
#  Fit the linear model to the training set:
lm_fit <- fit(lm_wflow, abalone_train)
# create a new data frame
hypo <- data.frame(type = "F", longest_shell = 0.50, diameter = 0.10,
                   height = 0.30,whole_weight = 4, shucked_weight = 1,
                   viscera_weight = 2, shell_weight = 1)
# Use the `fit()` object to predict the age of a hypothetical female abalone
predict(lm_fit,new_data = hypo)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.5
```

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.

Answer:

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

```
# install yardstick package
#install.packages('yardstick')
library('yardstick')

# 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)
```

- 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).

```
# The following code generates predicted values for age for each observation in the training
# set:
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
```

```r
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1  9.33
## 2  9.83
## 3 10.1
## 4  6.32
## 5  5.82
## 6  5.95
```

```r
# Now we attach a column with the actual observed age observations:
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  9.33   9.5
## 2  9.83   8.5
## 3 10.1    9.5
## 4  6.32   6.5
## 5  5.82   6.5
## 6  5.95   5.5
```

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

```r
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       0.558
## 3 mae     standard       1.55
```

- Report the results:
- The R^2 is about 0.558.
- The RMSE (root mean squared error) is about 2.15.
- The MAE (mean absolute error) is about 1.55.
- Interpret the R^2 value:

"The coefficient of determination (commonly denoted R^2) is the proportion of the variance in the response variable that can be explained by the explanatory variables in a regression model." (From https://www.statology.org/r-squared-in-r/)

Since the R^2 of this model is about 0.558, this means 55.8% of the response variable ('age') can be explained by the predictor variables(type,longest_shell,diameter,height,whole_weight, shucked_weight,viscera_weight,shell_weight). By what we learned in PSTAT 126, we know that this model doesn't fit well.