PSTAT 131 Homework 2

Matthew Zhang

Spring 2022

Linear Regression

```
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   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
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()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(ISLR)
library(ggplot2)

df <- read_csv('./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.

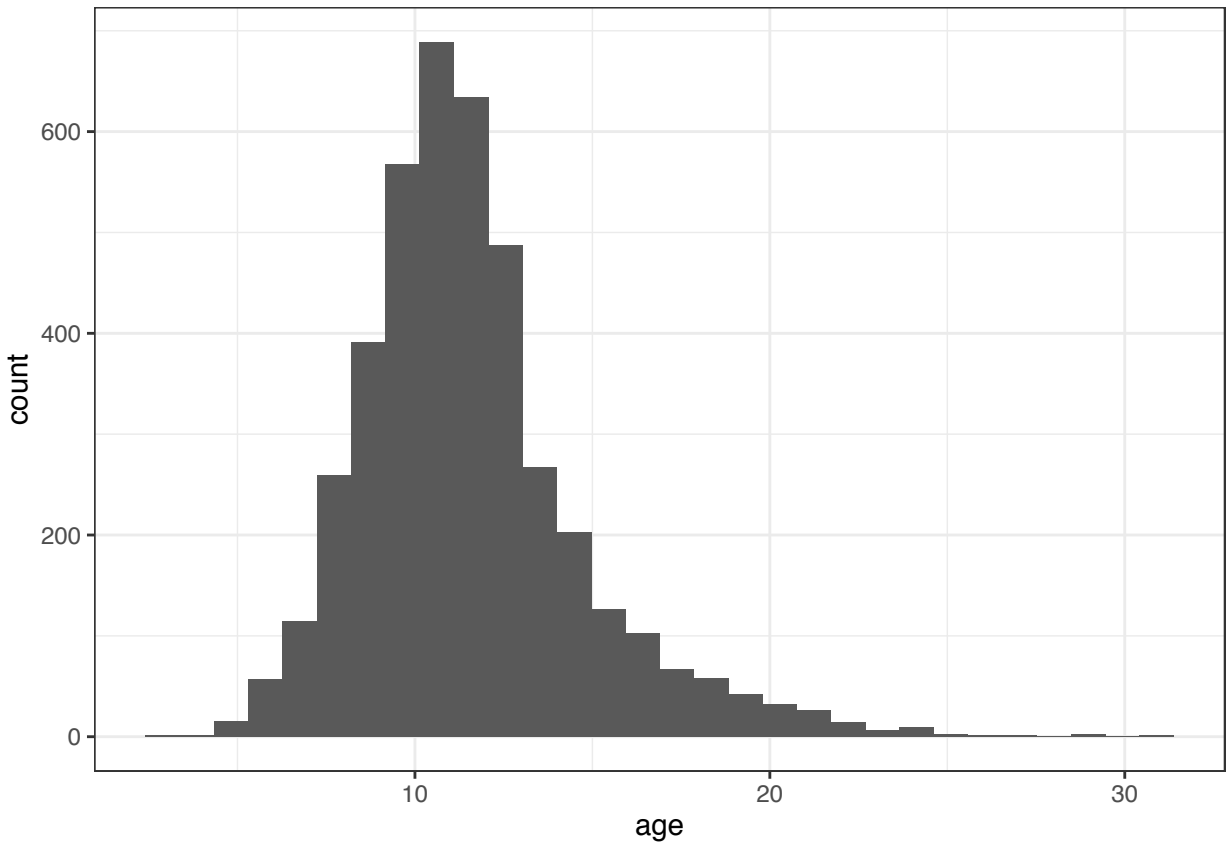
```
df['age'] <- df['rings'] + 1.5
df
```

```
## # 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>
```

Assess and describe the distribution of age.

```
df %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30) +
  theme_bw()
```



After creating the column 'age' based on the number of rings an abalone has plus 1.5, we can note that the distribution is similarly distributed to the 'rings' column. After adjusting the bin size, we can observe that the distribution is approximately normal with a slight tail on the right. Furthermore, a large number of observations have an age around 9-13.

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(3435)

abalone_split <- initial_split(df, prop = 0.80,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
?tidymodels
abalone_train2 = subset(abalone_train, select = -c(rings))
abalone_test2 = subset(abalone_test, select = -c(rings))
```

Once we have removed the rings column to remove multi-collinearity, we create a recipe for the model. Following the steps, we want to turn categorical predictors into dummy variables; center and scale all predictors.

```
abalone_recipe <- recipe(age ~ ., data = abalone_train2) %>%
  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_normalize(all_nominal_predictors())
abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with type:shucked_weight
## Interactions with longest_shell:diameter
## Interactions with shucked_weight:shell_weight
## Centering and scaling for all_nominal_predictors()
```

We should not include 'rings' to predict 'age' because the column 'age' is derived completely from rings and it would increase noise within the model. Age was made from simply adding 1.5 to rings which means they are highly correlated and it would not help improve the accuracy of our model.

Question 4

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

#similar to code in lab2
```

Question 5

```
lm_wf_abalone <- workflow() %>%
  add_model(lm_model_abalone) %>%
  add_recipe(abalone_recipe)

#similar to code in lab2
lm_fit <- fit(lm_wf_abalone, abalone_train2)
```

```
## Warning: Interaction specification failed for: ~type:shucked_weight. No
## interactions will be created.
```

```
lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy()
```

```
## # A tibble: 12 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                      2.16      0.590      3.66  2.52e- 4
##  2 longest_shell                    7.15      2.36       3.03  2.46e- 3
##  3 diameter                        23.9       3.13       7.66  2.51e-14
##  4 height                           5.88      1.64       3.59  3.39e- 4
##  5 whole_weight                     8.65      0.792     10.9   2.80e-27
##  6 shucked_weight                 -16.4       1.08     -15.3   7.28e-51
##  7 viscera_weight                  -7.41      1.45      -5.11  3.44e- 7
##  8 shell_weight                    13.6       1.52       8.92  7.51e-19
##  9 type_I                          -0.708     0.116     -6.11  1.13e- 9
## 10 type_M                           0.0566    0.0930     0.608 5.43e- 1
## 11 longest_shell_x_diameter       -34.7       4.13      -8.40  6.64e-17
## 12 shucked_weight_x_shell_weight   -1.93      1.69      -1.14  2.55e- 1
```

Question 6

```
female_abalone <- data.frame(type = "F",longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_wei

lm_fit_abalone <- fit(lm_wf_abalone, abalone_train2)
```

```
## Warning: Interaction specification failed for: ~type:shucked_weight. No
## interactions will be created.
```

```
lm_fit_abalone %>% extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 12 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                      2.16      0.590      3.66  2.52e- 4
##  2 longest_shell                    7.15      2.36       3.03  2.46e- 3
##  3 diameter                        23.9       3.13       7.66  2.51e-14
##  4 height                           5.88      1.64       3.59  3.39e- 4
##  5 whole_weight                     8.65      0.792     10.9   2.80e-27
##  6 shucked_weight                 -16.4       1.08     -15.3   7.28e-51
##  7 viscera_weight                  -7.41      1.45      -5.11  3.44e- 7
```

```
##  8 shell_weight                   13.6       1.52       8.92  7.51e-19
##  9 type_I                        -0.708     0.116     -6.11  1.13e- 9
## 10 type_M                         0.0566    0.0930     0.608 5.43e- 1
## 11 longest_shell_x_diameter      -34.7       4.13      -8.40  6.64e-17
## 12 shucked_weight_x_shell_weight  -1.93      1.69      -1.14  2.55e- 1
```

```
res <- predict(lm_fit_abalone, new_data = female_abalone)
res %>% head()
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.2
```

Question 7

```
abalone_train_res <- predict(lm_fit_abalone, new_data = abalone_train2)
abalone_train_res
```

```
## # A tibble: 3,340 x 1
##    .pred
##    <dbl>
##  1  8.22
##  2  9.95
##  3 10.3
##  4 10.1
##  5 10.6
##  6  6.35
##  7  5.76
##  8  5.94
##  9  8.87
## 10 11.4
## # ... with 3,330 more rows
```

```
abalone_train_res <- bind_cols(abalone_train_res, abalone_train2 %>% select(age))
abalone_train_res %>% head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  8.22   8.5
## 2  9.95   8.5
## 3 10.3    8.5
## 4 10.1    9.5
## 5 10.6    9.5
## 6  6.35   6.5
```

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

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.18
```

```
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       2.18
## 2 rsq     standard       0.546
## 3 mae     standard       1.57
```

The R-Squared can be interpreted as the proportion of variance explained by the model when using independent variables to predict a dependent variable. In other words, the model explaining a certain percent of variation. Since the R-squared value is 0.5464661, this is interpreted as the model accounting for 54.64661% of the variation in abalone ages.