PSTAT 131 Homework 2

Luke Fields (8385924)

April 6, 2022

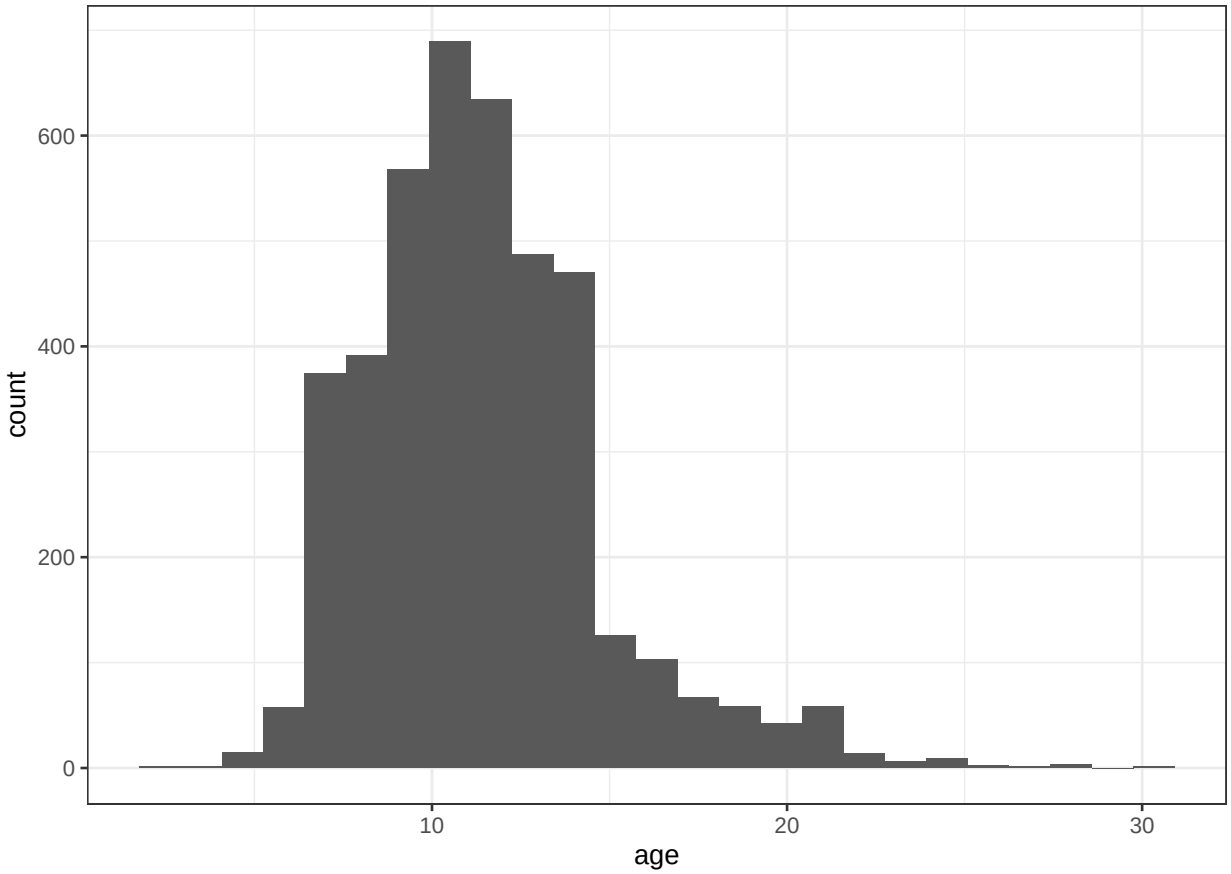
```
## corrplot 0.92 loaded

## -- 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()
## x dplyr::select() masks MASS::select()

## -- 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 dplyr::select()   masks MASS::select()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/

## 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.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.5     9.5    10.5    11.4    12.5    30.5
```

```
abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 25) +
  theme_bw()
```



The distribution of age is a positively skewed distribution, with a mean of 11.4, and median 10.5. THe youngest abalone is about 2.5 years old, and the oldest abalone is about 30.5 years old. Most abalones are somewhere between 7 and 15 years old, on average.

Question 2: Split the abalone data into a training set and a testing set. Use stratified sampling.

```
set.seed(912)
abalone_split <- initial_split(abalone,
                               prop = 0.7, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
abalone_train
```

```
## # A tibble: 2,922 x 10
##    type  longest_shell diameter height whole_weight shucked_weight
##    <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>
##  1 M             0.35     0.265  0.09         0.226         0.0995
##  2 I             0.33     0.255  0.08         0.205         0.0895
##  3 I             0.425    0.3    0.095        0.352         0.141
##  4 I             0.355    0.28   0.085        0.290         0.095
##  5 M             0.365    0.295  0.08         0.256         0.097
##  6 M             0.465    0.355  0.105        0.480         0.227
##  7 I             0.24     0.175  0.045        0.07          0.0315
##  8 I             0.39     0.295  0.095        0.203         0.0875
##  9 I             0.325    0.245  0.07         0.161         0.0755
## 10 I             0.52     0.41   0.12         0.595         0.238
## # ... with 2,912 more rows, and 4 more variables: viscera_weight <dbl>,
## #   shell_weight <dbl>, rings <dbl>, age <dbl>
```

We have 2,922 rows in our training dataset, which is about 75% of our entire abalone dataset's 4,177 rows. We want to predict age, so we will use that as our strata.

Question 3: Using the training data, create a recipe predicting the outcome variable, age, with all other predictor variables. Explain why you 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) %>%
  step_dummy(type, levels = 3) %>%
  step_normalize() %>%
  step_center() %>%
  step_interact(terms = ~longest_shell:diameter) %>%
  step_interact(terms = ~shucked_weight:shell_weight) %>%
  step_interact(terms = ~type_M:shucked_weight) %>%
  step_interact(terms = ~type_F:shucked_weight) %>%
  step_interact(terms = ~type_I:shucked_weight)

abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
```

```
##  predictor          8
##
## Operations:
##
## Dummy variables from type
## Centering and scaling for <none>
## Centering for <none>
## Interactions with longest_shell:diameter
## Interactions with shucked_weight:shell_weight
## Interactions with type_M:shucked_weight
## Interactions with type_F:shucked_weight
## Interactions with type_I:shucked_weight
```

We should not use rings to predict age because if we have rings, then we automatically will know what age will be as it will be the amount of rings + 1.5. The whole purpose of this experiment and model is to see if there is a way to predict abalone age in a way that does not require cutting open the abalone and using microscopic technology, and instead using easier to obtain information. additionally, for our dummy variable (type / gender), female will be our "base" case.

Question 4: Create and store a linear regression object using the "lm" engine.

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

Here we set our engine to "lm" so that we have a linear regression object ready to be used for further questions.

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.

```
abalone_wflow <- workflow() %>%
  add_model(lm_object) %>%
  add_recipe(abalone_recipe)
```

Here we set up our workflow for our abalone model, using the object and recipe from questions 3 and 4, which allows us to fit our model further.

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.

```
abalone_fit <- fit(abalone_wflow, abalone_train)
```

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

```
abalone_pred <- abalone_fit %>%
  extract_fit_parsnip() %>%
  tidy()
hypo_f_attrib <- data.frame(longest_shell = 0.5, diameter = 0.1, height = 0.3,
                        whole_weight = 4, shucked_weight = 1,
                        viscera_weight = 2, shell_weight = 1, type = "F")
hypo_age <- predict(abalone_fit, new_data = hypo_f_attrib)
abalone_pred
```

```
## # A tibble: 14 x 5
##    term                           estimate std.error statistic  p.value
##    <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                       3.68      0.738     4.99  6.42e- 7
##  2 longest_shell                     3.86      2.59      1.49  1.36e- 1
##  3 diameter                         24.5       3.39      7.22  6.55e-13
##  4 height                            4.91      1.67      2.94  3.28e- 3
##  5 whole_weight                     10.5       0.875    12.0   1.51e-32
##  6 shucked_weight                  -20.5       1.25    -16.3   1.52e-57
##  7 viscera_weight                  -11.2       1.55     -7.21  7.32e-13
##  8 shell_weight                      9.96      1.69      5.91  3.86e- 9
##  9 type_I                           -1.83      0.262    -6.96  4.12e-12
## 10 type_M                           -0.405     0.228    -1.78  7.52e- 2
## 11 longest_shell_x_diameter        -28.8       4.47     -6.45  1.31e-10
## 12 shucked_weight_x_shell_weight     1.19      1.84      0.645 5.19e- 1
## 13 type_M_x_shucked_weight           1.10      0.461     2.39  1.71e- 2
## 14 type_I_x_shucked_weight           3.95      0.797     4.96  7.62e- 7
```

```
hypo_age
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  18.5
```

The predicted age for this hypothetical female (let's call her Shelly) will be 18.54. Interpreting this, because Shelly is a female with a longest shell length of 0.5 mm, a diameter of 0.1 mm, a height of 0.3 mm, weighs 4 total g, with a meat weight of 1 g, gut weight of 2 g, and shell weight of 1g, she is expected to have an age of 18.54 years old.

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.

```
abalone_train_resid <- predict(abalone_fit, new_data = abalone_train %>% select(-age))
abalone_train_resid <- bind_cols(abalone_train_resid,
                              abalone_train %>% select(age))
abalone_metrics <- metric_set(rmse, rsq, mae)
```

```
abalone_metrics(abalone_train_resid, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.14
## 2 rsq     standard       0.554
## 3 mae     standard       1.55
```

This R2 value means that about 55.37% of the variability in age can be explained using the predictor variables in our model, which is not that strong. An rmse of 2.145, and a mean square error of 1.546 show that our abalone. model did not perform the greatest.