PSTAT 131 Homework 2

Brett Goldman

10/3/2022

Setup

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

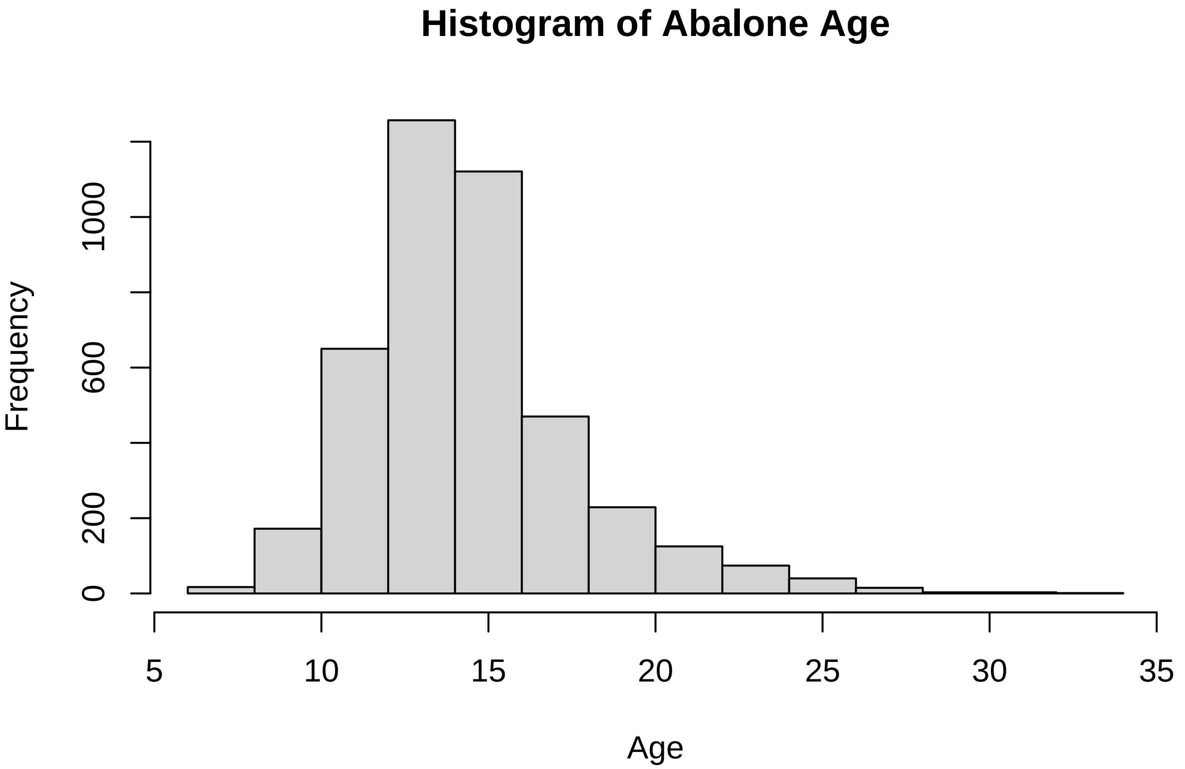
```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.1     v yardstick    1.1.0
## v recipes      1.0.1
## -- 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/
```

```
abalone<-read_csv(file="~/Documents/School/PSTAT 131/homework-2/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:

```
abalone$age<-abalone$rings+5
hist(abalone$age, main="Histogram of Abalone Age", xlab="Age")
```



Looking at the histogram of age, most abalone in this dataset are between 12-15 years old. There are some very young and very old abelone as well, but it's mostly in that 12-15 range.

Question 2:

```
# 80/20 split
set.seed(3005)
abalone_split<-initial_split(abalone, prop=0.8, strata=age)
abalone_train<-training(abalone_split)
abalone_test<-testing(abalone_split)
```

Question 3:

You shouldn't use rings to predict age because age is just rings+5. They have the same distribution just shifted over 5.

```
simple_abalone_recipe<-
  recipe(age~type+longest_shell+diameter+height+whole_weight+
           shucked_weight+viscera_weight+shell_weight, data=abalone_train) %>% # recipe
  step_dummy(all_nominal_predictors())
```

```
abalone_recipe<-simple_abalone_recipe %>%
  step_interact(terms = ~starts_with("type"):shucked_weight + longest_shell:diameter + shucked_weight:sl
abalone_recipe<-abalone_recipe %>%
  step_center(starts_with("type"), longest_shell, diameter, height, whole_weight,
           shucked_weight, viscera_weight, shell_weight)
```

```
abalone_recipe<-abalone_recipe %>%
  step_scale(starts_with("type"), longest_shell, diameter, height, whole_weight,
             shucked_weight, viscera_weight, shell_weight) # scale
```

Question 4:

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

Question 5:

```
abalone_lm_wflow<-workflow() %>%
  add_model(abalone_lm_model) %>%
  add_recipe(abalone_recipe)
```

Question 6:

```
abalone_lm_fit<-fit(abalone_lm_wflow, abalone_train)
Q6Abalone<-data.frame(type="F", longest_shell=.5, diameter=.1, height=.3,
                      whole_weight=4, shucked_weight=1, viscera_weight=2, shell_weight=1)
```

```
(abalonepredict<-predict(abalone_lm_fit, new_data=Q6Abalone))
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  28.5
```

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 is shown above.

Question 7:

```
library(yardstick)
abalone_train_res<-predict(abalone_lm_fit, new_data = abalone_train %>% select(-age))
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1 12.9
## 2 11.6
## 3 13.3
## 4 13.8
## 5 13.6
## 6  9.77
```

```
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 12.9     12
## 2 11.6     12
## 3 13.3     12
## 4 13.8     12
## 5 13.6     13
## 6  9.77    10
```

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

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.16
```

```
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       2.16
## 2 rsq     standard       0.549
## 3 mae     standard       1.55
```

The R squared value is low, so we can say that our model did not do a great job of modeling the true age of the abalone.