PSTAT231 HW2 Cheng Ye

Cheng Ye

2022-10-13

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## 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)
```

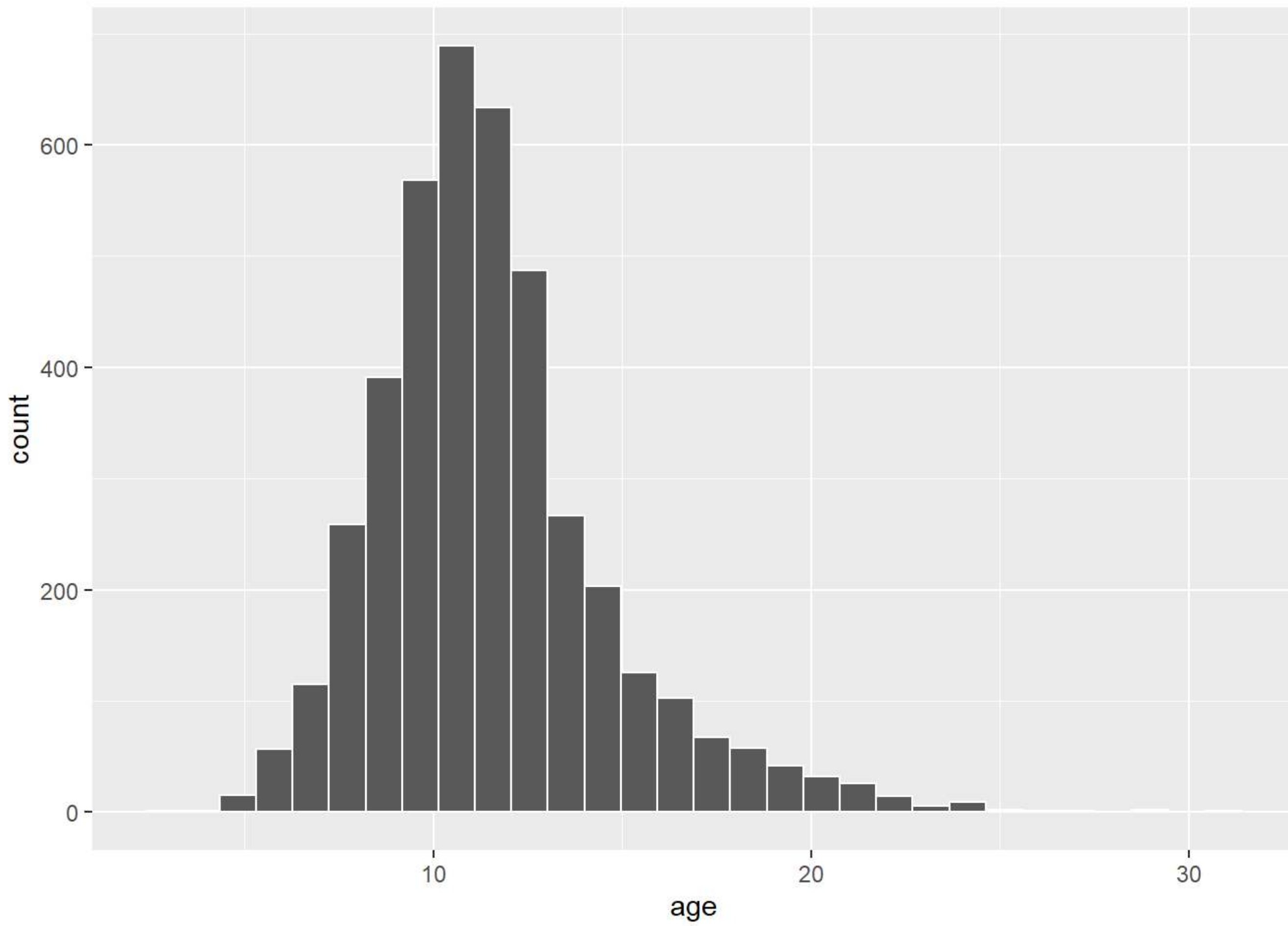
```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.1
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     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()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(ggplot2)
library(readr)
library(workflows)
library(dplyr) # Load required library
```

Question 1

```
abalone <- read.csv("C:/Cheng Ye/UCSB/PSTAT 231/HW/homework-2/homework-2/data/abalone.csv") #loa
d required data set
abalone$age <- abalone$ring+1.5 #define the abalone_age dataset according to the information giv
en
ggplot(data=abalone)+
  geom_histogram(mapping =  aes(x = age ), col = "white") #Plot the histogram of the dataset wit
h updated data
```

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




```
#From the graph, we know that the distribution of age is a right-skewed normal distribution
```

Question 2

```
set.seed(231)
abalone_split<-initial_split(abalone,prop=0.80,strata=age) #Train 80%, test 20%
abalone_train<-training(abalone_split) #Training dataset
abalone_test<-testing(abalone_split) #Testing dataset
```

Question 3

```
abalone_train_data <-subset(abalone_train,select=-rings)
abalone_age_recipe <- recipe(age ~ . , data = abalone_train_data) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_center(all_nominal_predictors()) %>%
  step_scale(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
  step_interact(terms = ~ shucked_weight:shell_weight)
summary(abalone_age_recipe)
```

```
## # A tibble: 9 x 4
##   variable       type    role      source
##   <chr>          <chr>   <chr>     <chr>
## 1 type           nominal predictor original
## 2 longest_shell  numeric predictor original
## 3 diameter       numeric predictor original
## 4 height         numeric predictor original
## 5 whole_weight   numeric predictor original
## 6 shucked_weight numeric predictor original
## 7 viscera_weight numeric predictor original
## 8 shell_weight   numeric predictor original
## 9 age            numeric outcome   original
```

```
#Hence from the result we could observe that we could NOT use rings to predit age as age= rings+
1.5
```

Question 4

```
my_lm_model <- linear_reg()%>%
  set_engine('lm')
print(my_lm_model)
```

```
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

Question 5

```
wkflow <- workflow() %>%
  add_model(my_lm_model) %>%
  add_recipe(abalone_age_recipe)
print(wkflow)
```

```
## == Workflow ================================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 6 Recipe Steps
##
## * step_dummy()
## * step_center()
## * step_scale()
## * step_interact()
## * step_interact()
## * step_interact()
##
## -- Model -----------------------------------------------------------------------
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

Question 6

```
lm_fit_model<-fit(wkflow,abalone_train_data)
female_abalone_age <- 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)
predict(lm_fit_model,new_data=female_abalone_age)
```

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

Question 7

```
library(yardstick)
abalone_metric <- metric_set(rsq,rmse,mae)
abalone_predict_res <- predict(lm_fit_model, new_data = abalone_train_data %>% select(-age))
abalone_predict_res <- bind_cols(abalone_predict_res, abalone_train_data %>% select(age))
abalone_predict_res %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.42   8.5
## 2  8.08   8.5
## 3  9.37   9.5
## 4  9.77   8.5
## 5 10.4    8.5
## 6 10.0    9.5
```

```
abalone_metric(abalone_predict_res, truth=age,
                estimate=.pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.555
## 2 rmse    standard       2.14
## 3 mae     standard       1.53
```

Required for 231 Students

Question 8 Which term(s) in the bias-variance tradeoff above represent the reproducible error? Which term(s) represent the irreducible error?

##In the bias-variance tradeoff above, the term $Var(\hat{f}(x_0))$ represent the repoducible error. ##In the bias-variance tradeoff above, ther term $Var(\epsilon)$ represent the irreducible error.

Question 9 Using the bias-variance tradeoff above, demonstrate that the expected test error is always at least as large as the irreducible error.

##By Lecture Notes 1, Slide 72 we have that $hatf(x_0)$ = $E[Y|X = x_0]$, and then $E[\hat{f}(x+0) - E\hat{f}(x_0))^2]$ = [E[(x_0)]-f(x_0)]^2), hence we know that $E[\hat{f}(x+0) - E\hat{f}(x_0))^2]$ = [E[(x_0)]-f(x_0)]^2) = 0 (Slide 70 of Lecture Notes 1). Then we could get that the reproducible error $Var(\hat{f}(x_0))$ = $[Bias(\hat{f}(x_0))]^2$ = 0. Because that $Var(\epsilon)$ is always greater or equal to 0 so the expected test error is always at least the same as the irreducible error. QED

Question 10 Prove the bias-variance tradeoff.

##Knowing that $Bias(\hat{f}(x_0))$ = $E[\hat{f}(x_0)] - f(x_0)$, we have that $Bias(\hat{f}(x_0))^2$ = $E[\hat{f}(x_0)] - f(x_0)^2$. From Lecture Notes 1 Slide 70 that Y = f(X) +ϵ, then we could deduce that E(ϵ) = 0 and $E(f(X)) = f(X)$.Hence we get that Var(ϵ) = E(ϵ^2). Hence the bias-variance tradeoff could be substituted in the form that $E[(y_0 - \hat{f}(x_0))^2]$ = $E[(f(x_0) - \hat{f}(x_0))^2]$ + Var(ϵ) = $E[(f(x_0) - E[\hat{f}(x_0)] - (\hat{f}(x_0) - E[\hat{f}(x_0)]))^2]$ + Var(ϵ) = $E[(E[\hat{f}(x_0)] - f(x_0))^2] + E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2$ -(2E[f(x_0)]-E[(x_0)])((x_0)-E[(x_0)]) + Var(ϵ) =((E[(x_0)]-f(x_0))^2) + $E[\hat{f}(x_0) - E[\hat{f}(x_0)])^2]$ - $2(f(x_0) - E[\hat{f}(x_0)])E[\hat{f}(x_0) - E[\hat{f}(x_0)]])]$+Var([]) = $(E[\hat{f}(x_0)] - f(x_0))^2$ + $E[\hat{f}(x_0) - E[\hat{f}(x_0)])^2]$ + Var()) = $[Bias(\hat{f}(x_0))]^2$ + $Var(\hat{f}(x_0))$ + Var(ϵ). Therefore, QED [Most of the latex are learned based on http://www.evanlray.com/stat242_f2019/resources/R/MathinRmd.html (http://www.evanlray.com/stat242_f2019/resources/R/MathinRmd.html)]