Pstat231HW2

Zihao Yang

2022-04-10

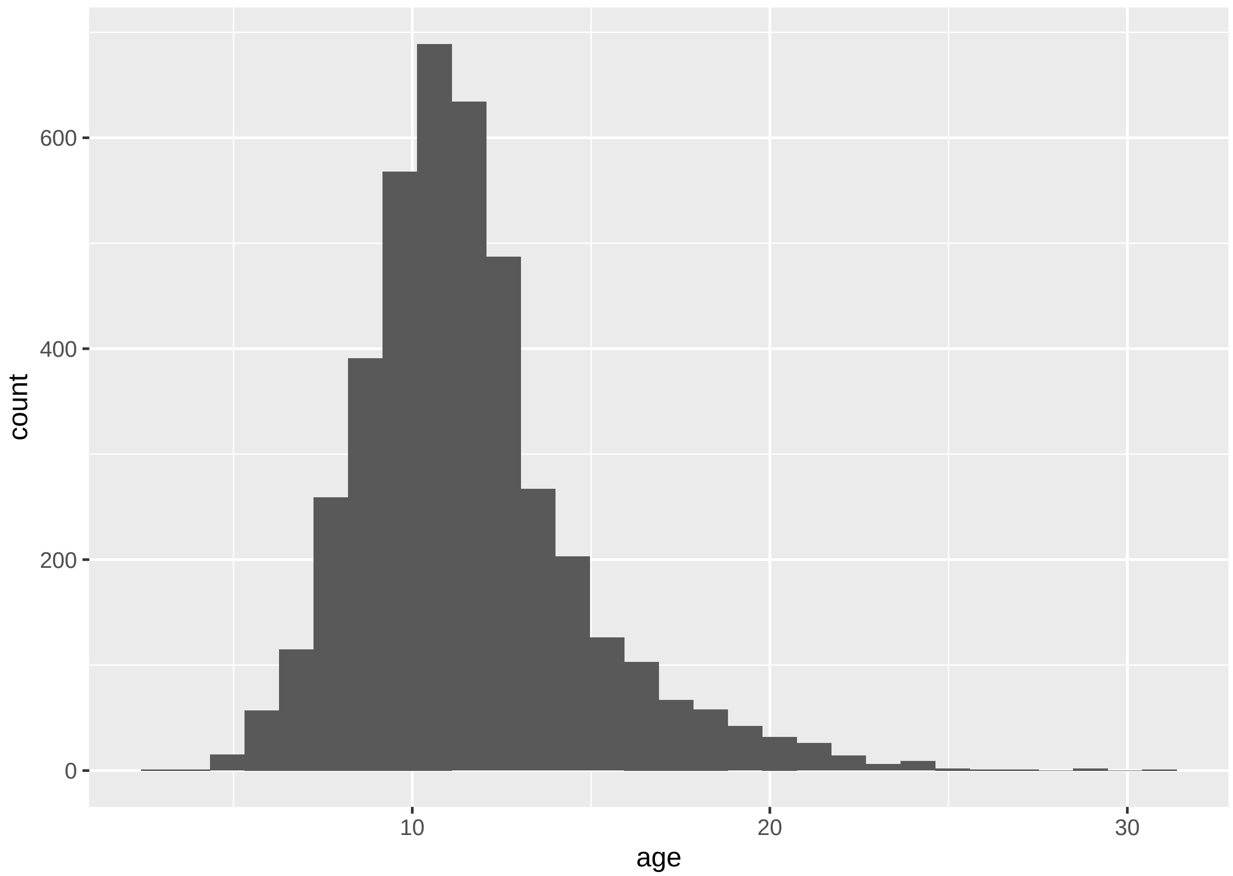
```
#install.packages("tidyverse")
#install.packages("tidymodels")
#install.packages("ISLR")
tinytex::install_tinytex
library(tidyverse)
library(tidymodels)
library(ISLR)
library(ggplot2)
library(corrplot)
library(ggthemes)
library(yardstick)
tidymodels_prefer()
set.seed(100)
```

```
# Get the dataset
abalone <- read.csv("abalone.csv")
head(abalone)
```

```
##   type longest_shell diameter height whole_weight shucked_weight viscera_weight
## 1    M         0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    M         0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    F         0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    M         0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    I         0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    I         0.425    0.300  0.095       0.3515         0.1410         0.0775
##   shell_weight rings
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

Q1

```
# Add age column to the abalone with "rings" + 1.5
abalone["age"] <- abalone["rings"]+1.5
# To assess the distribution of age, we can use histogram to check
abalone %>% ggplot(aes(age))+geom_histogram(bins=30)
```



According to the plot, the distribution of age relatively follows the normal distribution with mean at about 10-12, but it is slightly skewed to the right. The majority of data locates between 4 and 17, however, there exist some extreme outliers around 30.

Q2

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

Q3

```
abtrain_wo_rings <- abalone_train %>% select(-rings)
abalone_recipe <- recipe(age ~ ., data = abtrain_wo_rings) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms= ~ starts_with("type"):shucked_weight+
                longest_shell:diameter+
                shucked_weight:shell_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
```

We can't use rings to predict age, because the age column is just the linear transformation of the rings column, they have exactly the same trend and distribution with shift. Thus, rings cannot be used to predict age.

Q4

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

Q5

```
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
```

Q6

```
lm_fit <- fit(lm_wflow,abalone_train %>% select(-rings))
female_pred <- 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, new_data = female_pred)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  20.5
```

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

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                    11.4       0.0374   306.    0
##  2 longest_shell                   0.0318    0.289      0.110 9.12e- 1
##  3 diameter                        2.15      0.317      6.78  1.46e-11
##  4 height                          0.464     0.0984     4.72  2.49e- 6
##  5 whole_weight                    4.84      0.397     12.2   1.62e-33
##  6 shucked_weight                 -4.03      0.255    -15.8   2.35e-54
##  7 viscera_weight                 -1.06      0.158     -6.70  2.40e-11
##  8 shell_weight                    1.57      0.222      7.06  1.99e-12
##  9 type_I                         -0.915     0.114     -8.01  1.61e-15
## 10 type_M                         -0.171     0.104     -1.64  1.02e- 1
## 11 type_I_x_shucked_weight         0.499     0.0862     5.79  7.70e- 9
## 12 type_M_x_shucked_weight         0.202     0.110      1.84  6.53e- 2
## 13 longest_shell_x_diameter       -2.43      0.409     -5.94  3.08e- 9
## 14 shucked_weight_x_shell_weight  -0.232     0.207     -1.12  2.62e- 1
```

Q7

```
abalone_train_res <- predict(lm_fit, new_data = abtrain_wo_rings %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abtrain_wo_rings %>% select(age))
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.58   8.5
## 2  8.06   8.5
## 3  9.19   9.5
## 4  9.71   8.5
## 5 10.0    9.5
## 6  5.96   5.5
```

```
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       0.554
## 3 mae     standard       1.55
```

We get approximate 0.55437525 for R squared value which indicates that 55.437525% of the data fit the regression model.

231 part:

Q8

reducible error: $var(\hat{f}(x_0)) + [bias(\hat{f}(x_0))]^2$ irreducible error: $Var(\epsilon)$

Q9

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

The best case is letting the reducible error reduce to 0, that is to say, let $\hat{f}(x_0)$ be unbiased, which means it equals $f(x_0)$. Then the first and second term of the right hand side equal to 0. But the irreducible error $Var(\epsilon)$ still exists. Then

$$E[(y_0 - \hat{f}(x_0))^2] = 0 + 0 + Var(\epsilon) = Var(\epsilon)$$

Thus, the expected test error is always at least as large as the irreducible error.

Or we can say

$$E[(y_0 - \hat{f}(x_0))^2] = E[(f(x_0) + \epsilon - \hat{f}(x_0))^2]$$

by underlying model.

Since under the best case,$\hat{f}(x_0)$ is unbiased, and equals to $f(x_0)$. Then we have

$$E[(f(x_0) + \epsilon - \hat{f}(x_0))^2] = E[(f(x_0) + \epsilon - f(x_0))^2] = E[\epsilon^2]$$

Since ϵ is zero-mean random noise term. Then

$$\begin{aligned} E[\epsilon] &= 0 \\ E[\epsilon]^2 &= 0 \\ E[\epsilon^2] &= E[\epsilon^2] - 0 \\ &= E[\epsilon^2] - E[\epsilon]^2 \\ &= Var(\epsilon) \end{aligned}$$

Thus, in the best case $E[(f(x_0) + \epsilon - \hat{f}(x_0))^2] = Var(\epsilon)$.

Q10

$$\begin{aligned} E[(y_0 - \hat{f}(x_0))^2] &= E[(f(x_0) + \epsilon - \hat{f}(x_0))^2] \\ &= E[(f(x_0) - \hat{f}(x_0))^2] + 2E[(f(x_0) - \hat{f}(x_0))\epsilon] + E[\epsilon^2] \\ &= E\left[\left(f(x_0) - E(\hat{f}(x_0)) + E(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2\right] + 2E[(f(x_0) - \hat{f}(x_0))\epsilon] + Var(\epsilon) \\ &= E\left[\left(f(x_0) - E(\hat{f}(x_0)) + E(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2\right] + 2E[(f(x_0) - \hat{f}(x_0))]E[\epsilon] + Var(\epsilon) \\ &= E\left[\left(f(x_0) - E(\hat{f}(x_0)) + E(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2\right] + Var(\epsilon) \\ &= E[(E[\hat{f}(x_0)] - f(x_0))^2] + E[(\hat{f}(x_0) - E[\hat{f}(x_0)]^2] - 2E[(f(x_0) - E[\hat{f}(x_0)])(\hat{f}(x_0) - E[\hat{f}(x_0)])] + Var(\epsilon) \\ &= (E[\hat{f}(x_0)] - f(x_0))^2 + E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2] - 2(f(x_0) - E[\hat{f}(x_0)])E[(\hat{f}(x_0) - E[\hat{f}(x_0)])] + Var(\epsilon) \\ &= [Bias[\hat{f}(x_0)]]^2 + Var(\hat{f}(x_0)) - 2(f(x_0) - E[\hat{f}(x_0)])(E[\hat{f}(x_0)] - E[\hat{f}(x_0)]) + Var(\epsilon) \\ &= [Bias[\hat{f}(x_0)]]^2 + Var(\hat{f}(x_0)) + Var(\epsilon) \\ &= Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon). \end{aligned}$$