HW2_ShuheiKaneko

Shuhei Kaneko

2022/10/04

Question 1

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
```

```
## v broom        1.0.1      v recipes      1.0.1
## v dials        1.0.0      v rsample      1.1.0
## v dplyr        1.0.10     v tibble       3.1.8
## v ggplot2      3.3.6      v tidyr        1.2.1
## v infer        1.0.3      v tune         1.0.0
## v modeldata    1.0.1      v workflows    1.0.0
## v parsnip      1.0.1      v workflowsets 1.0.0
## v purrr        0.3.4      v yardstick    1.1.0
```

```
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v readr   2.1.2     v forcats 0.5.2
## v stringr 1.4.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x readr::spec()       masks yardstick::spec()
```

```
abalone_data <- read_csv("abalone.csv")
```

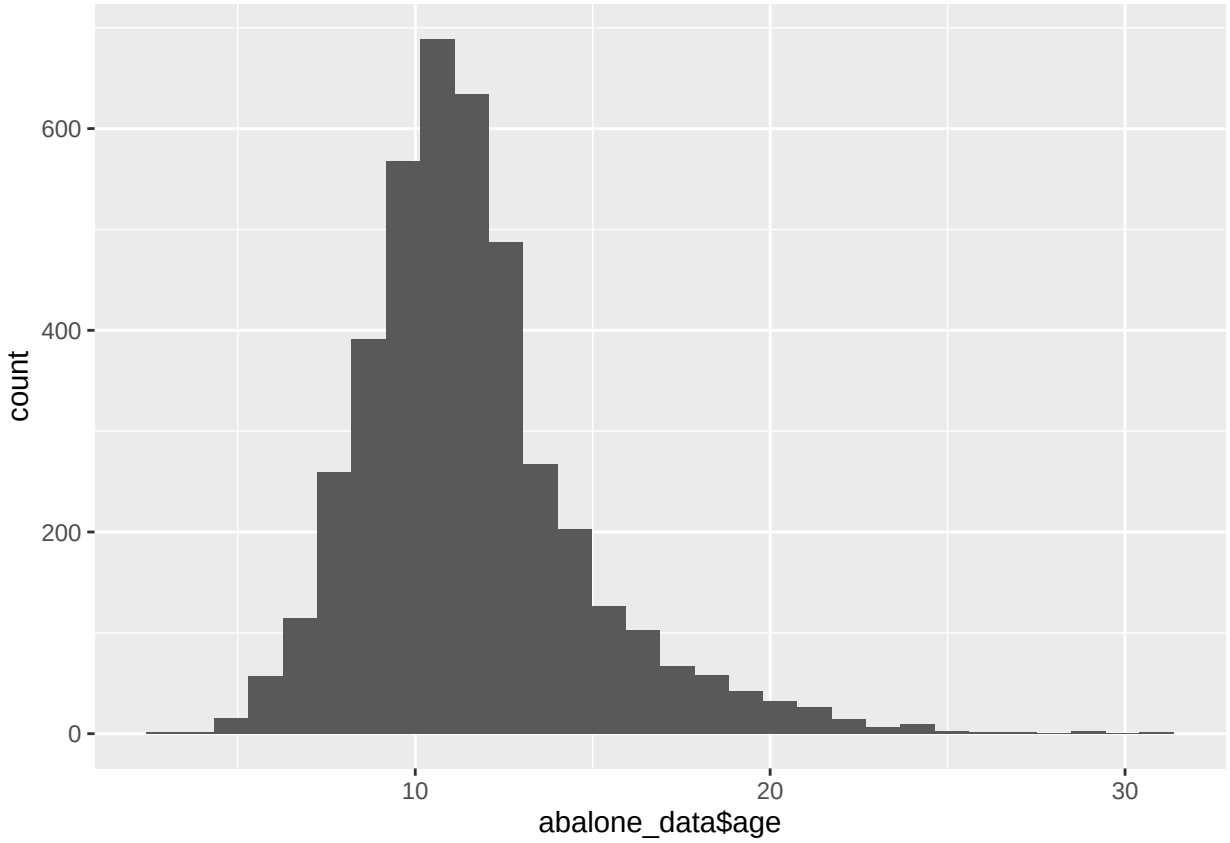
```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
abalone_data <- abalone_data %>%
  mutate(age = rings + 1.5)

age_plot <-   ggplot() +
  geom_histogram(aes(x = abalone_data$age))
plot(age_plot)
```

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



There are a lot of observation around 10-15. The range is approximately 30. The distribution of the age variable is slightly right skewed.

Question 2 - 3

```
set.seed(10)

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

abalone_train <- abalone_train %>% select(-rings)
abalone_test <- abalone_test %>% select(-rings)

abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
```

```
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
  step_interact(terms = ~ shucked_weight:shell_weight) %>%
  step_center() %>%
  step_scale()
```

Because age and rings are lineary dependent by its construction, we cannot include both of them in the same model. Otherwise, we cannot estimate other parameters.

Quesiton 4

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

Question 5

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

Question 6

```
lm_fit <- fit(lm_wflow, abalone_train)

abalone_hypo <- tibble(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 = abalone_hypo)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  24.4
```

The predicted value of age of the hypothesized data is 24.429.

Question 7

```
#Construct metrics
abalone_metrics <- metric_set(rmse, rsq, mae)

# Prediction based on training data
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))

# Add a column of real value
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  9.45   8.5
## 2  8.01   8.5
## 3  9.27   9.5
## 4  9.66   8.5
## 5 10.3    8.5
## 6 10.9    9.5
```

```
# Report the performance measure
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       0.560
## 3 mae     standard       1.55
```

The interpretation of R-square: How much of the variance in the outcome variable can be explained by the model. However, we should be careful to use R-square because R-square can be improved if we add variables (even if they are totally useless for the prediction.).

Question 8

The first term ($Var(\hat{f}(x_0))$) and the second term ($Bias(\hat{f}(x_0))^2$) are reproducible error. The final term ($Var(\epsilon)$) is the irreducible error.

Question 9

Noting that the first term and second term are always positive (more precisely, they become zero if $\hat{f}(x) = \mathbb{E}(y|x)$.), the expected test error is always at least as large as the irreducible error (i.e. $Var(\epsilon)$).

Question 10

$$\begin{aligned}
\mathbb{E}[(y_0 - \hat{f}(x_0))^2] &= \mathbb{E}[(y_0 - f(x_0) + f(x_0) - \hat{f}(x_0))^2] \\
&= \mathbb{E}[(y_0 - f(x_0))^2] + \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + 2\mathbb{E}[\{y_0 - f(x_0)\}\{f(x_0) - \hat{f}(x_0)\}] \\
&= \mathbb{E}[\epsilon^2] + \mathbb{E}[(f(x_0) - \hat{f}(x_0))]^2 + \mathbb{V}((f(x_0) - \hat{f}(x_0)) + 0 \\
&= \mathbb{V}(\epsilon) + Bias(\hat{f}(x_0))^2 + \mathbb{V}(\hat{f}(x_0))
\end{aligned}$$

where the first equality holds by adding and subtracting $f(x_0)$, the third equality holds by $y_0 = f(x_0) + \epsilon$ and $\mathbb{E}(y_0) = f(x_0)$, and the final equality holds by $V(\epsilon) = \mathbb{E}(\epsilon^2) - \mathbb{E}(\epsilon)^2$ and the definition of the $Bias(.)$.