Homework02

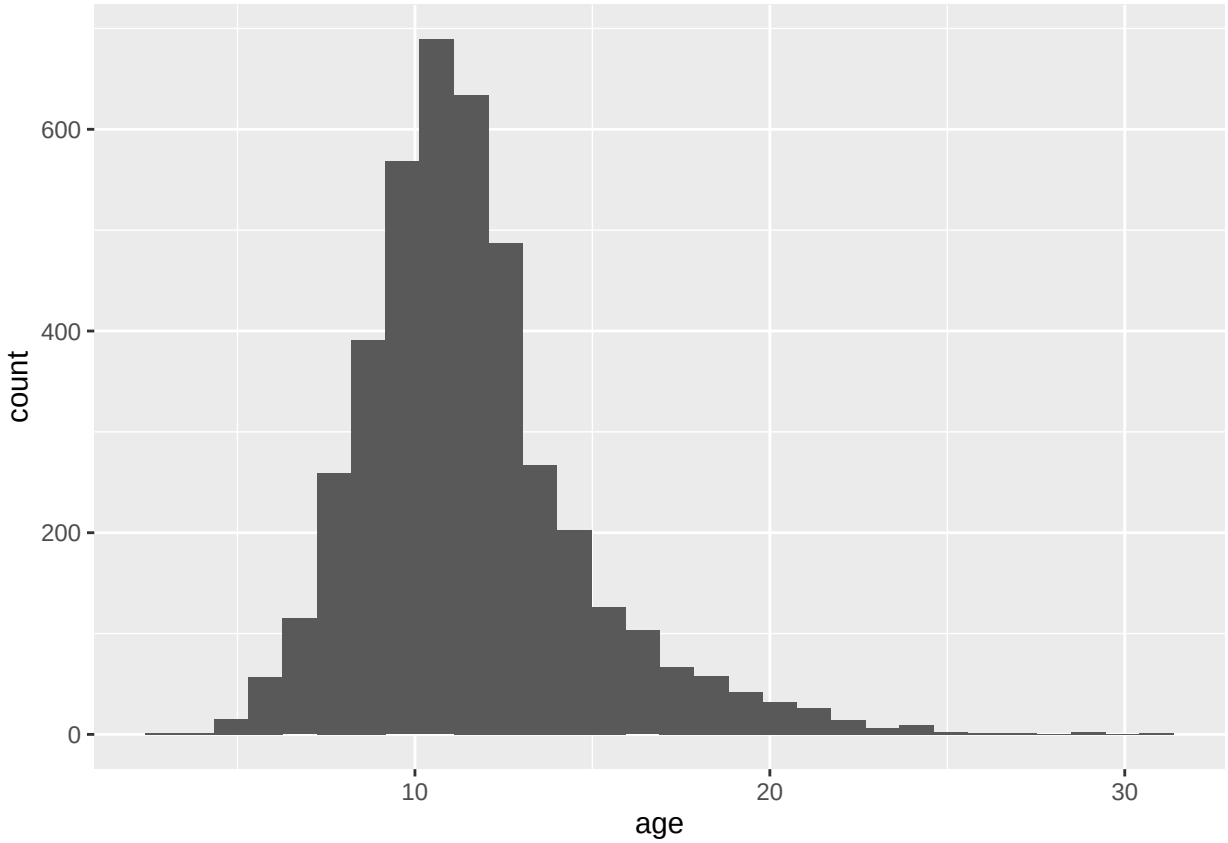
Contents

```
set.seed(231)
```

Question 1

```
abalone = read.csv(file="data/abalone.csv")
abalone[, 'age'] = abalone$rings + 1.5
ggplot(abalone, aes(x = age)) +
        geom_histogram()
```

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



The ages are distributed around 11. Most of values are around 10 and 11.

Question 2

```
abalone2 = subset(abalone, select = -rings)
```

```
abalone_split <- initial_split(abalone2, prop = 0.80,
                               strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```
abalone_recipe = recipe(age ~ ., data = abalone_train) %>%
        step_dummy(all_nominal_predictors())
```

1. dummy code any categorical predictors

```
abalone_recipe = abalone_recipe %>%
        step_interact(terms = ~ starts_with('type'):shucked_weight +
                              longest_shell:diameter +
                              shucked_weight:shell_weight)
```

2. create interactions between

```
abalone_recipe = abalone_recipe %>%
        step_center(all_predictors())
```

3. center all predictors, and

```
abalone_recipe = abalone_recipe %>%
        step_scale(all_predictors())
```

```
cor(abalone$rings, abalone$age)
```

4. scale all predictors.

```
## [1] 1
```

There is a strong positive correlation between 'rings' and 'age', because 'age' is 'rings' + 1.5. It indicates those two variables have exactly same role/effect.

Question 4

```
lm_model = linear_reg() %>%
        set_engine('lm')
```

Question 5

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

```
lm_fit = fit(lm_wflow, abalone_train)
p1 = 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, p1)
```

Question 6

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

Question 7

```
abalone_metrics <- metric_set(rmse, rsq, mae)
```

1

```
train_result = predict(lm_fit,
                       new_data = abalone_train %>%
                               select(-age))

train_result = bind_cols(train_result, abalone_train %>% select(age))
```

2

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

3

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

R^2 is 0.55. It means the model could explain the data approximately 50%. It is quite a disappointing result.

Question 8

$Var(\hat{f}(x_o))$, $[Bias(\hat{f}(x_0))]^2$ are the reproducible error and $Var(\epsilon)$ is the irreducible error.

Question 9

The expected test MSE, the left term, is decomposed to Variance + Biasˆ2 + Irreducible error. The expected values of Variance and Bias terms would be zero by several methods such as reducing predictor or using several models. So that, the expected test MSE is 0 + Irreducible error. Therefore, the irreducible error is the minimum value of the expected test MSE.

Question 10

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

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

$\bullet E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2] = 0$ $\bullet [E[\hat{f}(x_0)] - f(x_0)]^2 = 0$

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