Homework Assignment 2

Vasyl Ostapenko (774 970 8)

April 06, 2022

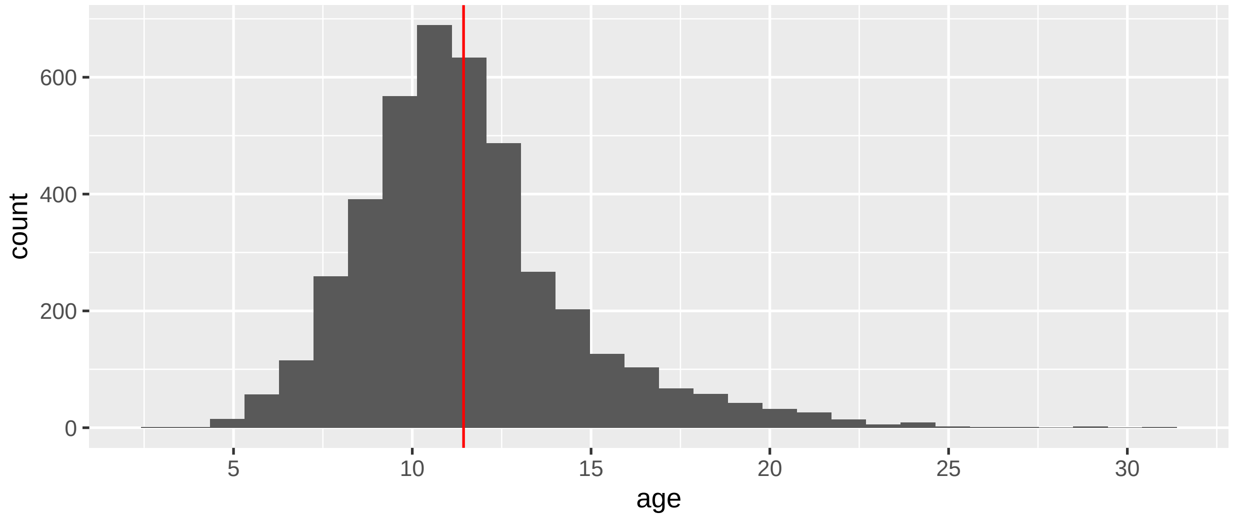
Load Data

```
DATA_FOLDER = "./data"
ABALONE_FNAME = file.path(DATA_FOLDER, "abalone.csv")
data = read.csv(ABALONE_FNAME)
```

Question 1

```
data$age = data$rings + 1.5
data = data[ , !(colnames(data) == "rings")] %>% copy() # deselect for later
```

```
ggplot(data, aes(x=age)) +
  geom_histogram(bins=30) +
  scale_x_continuous(breaks=seq(0, 35, 5)) +
  geom_vline(aes(xintercept=mean(age)), col='red')
```



The variable age displays a right-skewed distribution with a mean around 12 years and standard deviation of around 3 years.

Question 2

Stratified random sample (using types as strata) with a .80 / .20 training and test split.

```
abalone_split = data %>%
  initial_split(prop=0.8, strata="type")
```

```
abalone_train = training(abalone_split)
abalone_test = testing(abalone_split)
```

Question 3

```
abalone_recipe = recipe(age ~ ., data=abalone_train) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact( ~ starts_with("type"):shucked_weight) %>%
  step_interact( ~ longest_shell:diameter) %>%
  step_interact( ~ shucked_weight:shell_weight) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors())
```

We shouldn't use rings in predicting age because the two variables are perfectly correlated. We will get a model which will only use rings as the predictor and this will not be useful to us at all.

Question 4

```
lm_model = linear_reg() %>%
  set_engine("lm") %>%
  set_mode("regression")
```

Question 5

```
abalone_workflow = workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
```

Question 6

```
abalone_fit = abalone_workflow %>%
  fit(abalone_train)
```

```
single_sample = list(longest_shell=0.50,
                     diameter=0.10,
                     height=0.30,
                     whole_weight=4,
                     shucked_weight=1,
                     viscera_weight=2,
                     shell_weight=1,
                     type="F")
single_sample = as.data.frame(single_sample)
predict(abalone_fit, single_sample) %>% unlist() %>% unname()
```

```
## [1] 23.45
```

Question 7

```
multi_metric = metric_set(rsq, rmse, mae)

bound_test_data = bind_cols(predict(abalone_fit, abalone_test),
                            abalone_test$age)
```

```
## New names:
## * `` -> ...2
```

```
colnames(bound_test_data) = c("Predicted Age", "True Age")

multi_metric(data=bound_test_data,
             truth="True Age",
             estimate="Predicted Age")
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.522
## 2 rmse    standard       2.30
## 3 mae     standard       1.62
```

Evaluated on the test data, our model performs quite poorly based on the R-squared criterion. At an R-squared of about .52, we have that 52% of the variability in the response is explained by the predictors. To note, we also have a RMSE of 2.30 and MAE of 1.62. As these are absolute measures of model performance, there needs to be more context to the numbers for their proper evaluation.