HW2

Peter Chu

10/16/2022

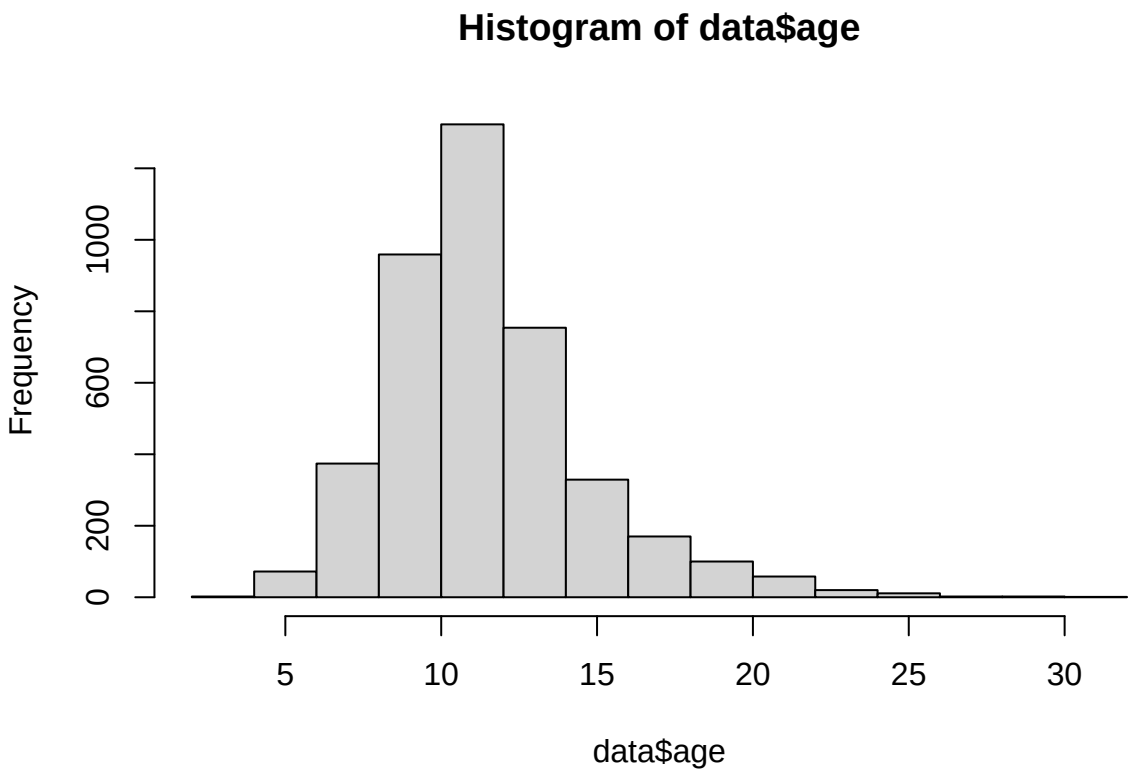
Question 1.

```
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.
```

```
data$age <- data$rings + 1.5

hist(data$age)
```




```
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.50    9.50   10.50   11.43   12.50   30.50
```

The distribution of ages among the abalones appears to be normal, but is right skewed a bit. The max age is 30.5, the min age is 2.5, and the average age is 11.43. Thus our summary statistics confirm that the graph is right skewed as more points lie closer to the min than the max.

Question 2.

```
set.seed(100)

data_split <- initial_split(data, prop = 0.8, strata = age)
data_train <- training(data_split)
data_tests <- testing(data_split)
```

Question 3.

We shouldn't use rings to predict age, because we calculated age from rings. Thus age is dependent on the value of rings, so rings will most likely be able to predict age. It is clear that there is no point in checking if rings is a predictor since age is dependent on it.

```
data_train_recipe <- recipe(age~ type + diameter + height + whole_weight + shucked_weight + viscera_wei
  step_dummy(all_nominal_predictors()) %>%
```

```
  step_interact(terms = ~ starts_with('type'):shucked_weight) %>%
  step_interact(terms = ~ shell_weight:shucked_weight) %>%
  step_interact(terms = ~ longest_shell:diameter) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())

data_train_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight
## Interactions with shell_weight:shucked_weight
## Interactions with longest_shell:diameter
## Centering for all_predictors()
## Scaling for all_predictors()
```

Question 4.

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

Question 5.

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

Question 6.

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

prediction <- predict(lm_fit, new_data = data.frame(longest_shell = 0.5, diameter = 0.1, height = 0.3, 

prediction
```

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

Using our model, the predicted age for the female abalone with given traits is 20.5 years old

Question 7.

```
data_train_res <- predict(lm_fit, new_data = data_train %>% select(-age))
data_train_res <- bind_cols(data_train_res, data_train %>% select(age))
data_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
```

```
data_metric <- metric_set(rmse,rsq,mae)
data_metric(data_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
```

Our R^2 value is 0.554 which means that about 55.4% of the variance in age, can be explained by the other independent variables. The MAE and RMSE are both 1.55 and 2.15, respectively,which shows that on average out model was about 1.5 years off when taking the absolute mean error and 2.55 years off when taking the quradtic mean error. Our prediction is about 1 ring off from actual. Thus, our R^2, MAE, and RMSE show that our model is not very good at predicting the actual age of abalone eggs.

Question 8.

The reproducible error terms are $Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2$. This is because by having a more accurate $\hat{f}$, we can miniize its value. The terms which are irreducible error is $Var(\epsilon)$. This is because this is inherent to the data and can not be eliminated due to a choice of function.

Question 9.

We know that Test Error $= E[(y_0 - \hat{f}(x_0))^2]$, $Var(\epsilon) \geq 0$ and $[Bias(\hat{f}(x_0))]^2 \geq 0$. Thus Test Error $= E[(y_0 - \hat{f}(x_0))^2] \geq 0 + 0 + Var(\epsilon) = Var(\epsilon)$. Thus we know that the Test Error is always greater than or equal to the irreducible error.

Question 10.

By construction of the problem we have that $E[\epsilon] = 0, Var(\epsilon) = E[\epsilon^2]$. Thus we have that $E[(y_0 - \hat{f}(x_0))^2] = E[(f(x_0) + \epsilon - \hat{f}(x_0))^2] = E[(f(x_0) - \hat{f}(x_0))^2] + E[\epsilon^2] + 2E[(f(x_0) - \hat{f}(x_0)) * \epsilon] = E[(f(x_0) - \hat{f}(x_0))^2] + Var(\epsilon) + 2E(f(x_0) - \hat{f}(x_0)) * 0 = E[(f(x_0) - \hat{f}(x_0))^2] + Var(\epsilon)$

We also have that $E[(f(x_0) - \hat{f}(x_0))^2] = E[((f(x_0) - E[\hat{f}(x_0)]) - (\hat{f}(x_0) - E[\hat{f}(x_0)]))^2] = E[(E[\hat{f}(x_0)] - f(x_0))^2] + E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2] - 2E[(f(x) - E[\hat{f}(x_0)]) * (\hat{f}(x_0) - E[\hat{f}(x_0)])] = Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) - 2E[(f(x) - E[\hat{f}(x_0)]) * (\hat{f}(x_0) - E[\hat{f}(x_0)]) = Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0))$

Thus when we combined these two equations we have that $E[(y_0 - \hat{f}(x_0))^2] = E[(f(x_0) - \hat{f}(x_0))^2] + Var(\epsilon) \rightarrow E[(y_0 - \hat{f}(x_0))^2] = Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) + Var(\epsilon)$