131-hw-2

Tonia Wu

4/8/2022

Q1

Looks like we have a right-skewed distribution with a mean of 11.43 years.

```
abalone['age'] = abalone$rings + 1.5
summary(abalone$age)
```

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

Q2

The data are split with 80% in the training set and 20% in the testing set.

```
# set seed
set.seed(286)

# split
aba_split <- initial_split(abalone, prop = 0.80, strata = age)
aba_train <- training(aba_split)
aba_test <- testing(aba_split)
```

Q3

Rings is directly used to calculate age, so there is no use using it to predict age.

```
aba_recipe <- recipe(age~type + longest_shell + diameter + height + whole_weight + shucked_weight + vis
  step_dummy(all_nominal_predictors()) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors()) %>%
  step_interact(~ type:shucked_weight) %>%
  step_interact(~ longest_shell:diameter) %>%
  step_interact(~shucked_weight:skull_weight)
```

Q4

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

Q5

```
# set up a workflow
lm_wflow <- workflow() %>%
  add_model(lm_model) %>%
  add_recipe(aba_recipe)

# fit linear model to training set
lm_fit <- fit(lm_wflow, aba_train)
```

```
## Warning: Interaction specification failed for: ~type:shucked_weight. No
## interactions will be created.
```

```
## Warning: Interaction specification failed for: ~shucked_weight:skull_weight. No
## interactions will be created.
```

```
# results
lm_fit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 11 x 5
##    term                      estimate std.error statistic  p.value
##    <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)               11.9       0.0589   202.    0
##  2 longest_shell             -0.687     0.252     -2.73  6.33e- 3
##  3 diameter                   0.529     0.244      2.17  3.01e- 2
##  4 height                     0.241     0.0691     3.49  4.91e- 4
##  5 whole_weight               5.48      0.415     13.2   7.19e-39
##  6 shucked_weight            -4.50      0.211    -21.4   4.78e-95
##  7 viscera_weight            -0.988     0.160     -6.18  7.09e-10
##  8 shell_weight               1.30      0.178      7.29  3.82e-13
##  9 type_I                    -0.378     0.0526    -7.20  7.44e-13
## 10 type_M                    -0.0132    0.0442    -0.298 7.65e- 1
## 11 longest_shell_x_diameter  -0.446     0.0463    -9.64  1.09e-21
```

Q6

The predicted age is 23.68:

```
hypo_f_aba <- data.frame(type = 'F', longest_shell = 0.5, diameter = 0.1, height = 0.3, whole_weight = 4

# results
predict(lm_fit, new_data = hypo_f_aba)
```

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

Q7

As the r-squared value is only about 56%, our model only explains 56% of the variation in abalone age.

```
# get training rmse
aba_train_rmse <- predict(lm_fit, new_data = aba_train %>% select(-age))
aba_train_rmse
```

```
## # A tibble: 3,340 x 1
##    .pred
##    <dbl>
##  1  9.31
##  2  8.27
##  3  9.99
##  4 10.3
##  5 10.1
##  6 10.7
##  7  6.30
##  8  5.63
##  9  5.84
## 10  8.92
## # ... with 3,330 more rows
```

```
# attach a column with observed ages
aba_train_rmse <- bind_cols(aba_train_rmse, aba_train %>% select(age))
aba_train_rmse %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.31   8.5
## 2  8.27   8.5
## 3  9.99   8.5
## 4 10.3    8.5
## 5 10.1    9.5
## 6 10.7    9.5
```

```r
# get rmse
rmse(aba_train_rmse, truth = age, estimate = .pred)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.14
```

```r
# create metric set
aba_metrics <- metric_set(rmse, rsq, mae)
aba_metrics(aba_train_rmse, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.14
## 2 rsq     standard       0.556
## 3 mae     standard       1.55
```