hw2

Sabrina Lem

4/10/2022

Question 1:

```
abalone <-read.csv("abalone.csv")
# adding age variable
abalone['age']<-c(abalone$rings+1.5)
head(abalone)
```

```
##   type longest_shell diameter height whole_weight shucked_weight viscera_weight
## 1    M         0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    M         0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    F         0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    M         0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    I         0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    I         0.425    0.300  0.095       0.3515         0.1410         0.0775
##   shell_weight rings  age
## 1        0.150    15 16.5
## 2        0.070     7  8.5
## 3        0.210     9 10.5
## 4        0.155    10 11.5
## 5        0.055     7  8.5
## 6        0.120     8  9.5
```

The distribution of age is the same as 'rings', shifted right by 1.5 units.
Question 2:

```
set.seed(1027)
abalone_split <- initial_split(abalone, prop = 0.7, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3:
we should not use rings to predict age because age is a function of rings. rings would be a perfect predictor for age, creating bias towards the entire regression and the rest of the estimator coefficients.

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

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

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

Question 4:

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

Question 5:

```
lm_wfl<-workflow() %>%
  add_model(lm_mod) %>%
  add_recipe(abalone_recipe)
```

Question 6:

```
lm_fit <-fit(lm_wfl, abalone_train)
lm_fit %>% extract_fit_parsnip() %>% tidy
```

```
## # A tibble: 10 x 5
##    term           estimate std.error statistic  p.value
##    <chr>             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)       5.44     0.344     15.8   4.26e-54
```

```
##  2 typeI            -0.832    0.120     -6.91  5.91e-12
##  3 typeM             0.171    0.0985     1.73  8.33e- 2
##  4 longest_shell     0.649    2.10       0.309 7.58e- 1
##  5 diameter         10.2      2.60       3.94  8.41e- 5
##  6 height            8.79     1.65       5.32  1.14e- 7
##  7 whole_weight     10.2      0.939     10.8   8.69e-27
##  8 shucked_weight  -21.4      1.06     -20.3   2.04e-85
##  9 viscera_weight  -12.0      1.60      -7.46  1.13e-13
## 10 shell_weight      7.97     1.42       5.59  2.45e- 8
```

```
new <-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, new)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  12.7
```

The predicted age for this abalone is 12.72 years old.

Question 7:

```
abalone_train_res<- predict(lm_fit, new_data = abalone_train %>%select(-age))
abalone_train_res<-bind_cols(abalone_train_res, abalone_train %>% select(
  age
))
abalone_metrics<-metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate=.pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.17
## 2 rsq     standard       0.539
## 3 mae     standard       1.56
```

$R^2 = 0.5394$. This means that about 53.94% of the variation in age can be explained by the predictor variables in the regression created above.