PSTAT 131 Homework 2

Tammy Truong

2022-04-06

Linear Regression

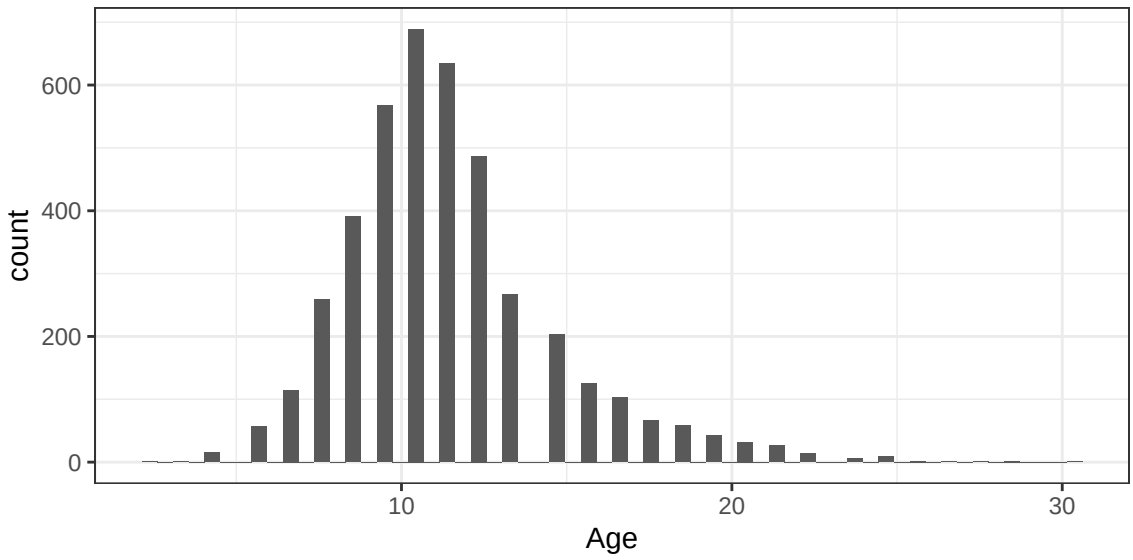
Question 1:

Predicting abalone age by computing `number of rings` + 1.5 and adding the `age` variable into the data set. I renamed the variable `Class Number of Rings` to `Rings` for simplicity.

```
abalone <- rename(abalone, Rings = Class_number_of_rings)
abalone <- mutate(abalone, Age = Rings + 1.5)
```

Here, we will assess and describe the distribution of `Age`.

```
abalone %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 60) +
  theme_bw()
```



From the histogram above, we see that is a spike at Age of 10 and it is skewed right.

Question 2:

Splitting the abalone data into a training set and a testing set using stratified sampling with a proportion of 80%.

```
set.seed(1004)

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

Question 3:

Using the **training** data, I create a recipe predicting the outcome variable, `age`, with all other predictor variables. *We do not include `rings` to predict `age` because the variable `age` stemmed from `rings`, and is only `rings` + 1.5.

Steps for the recipe:

1. dummy code any categorical predictors
2. create interactions between
   - `type` and `shucked_weight`
   - `longest_shell` and `diameter`
   - `shucked_weight` and `shell_weight`
3. center all predictors,
4. scale all predictors.

```
# creating dummy code for the recipe
abalone_recipe <- recipe(Age ~ Sex + Length + Diameter + Height + Whole_weight +
                           Shucked_weight + Viscera_weight + Shell_weight,
                         data = abalone_train) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("Sex"):Shucked_weight) %>%
  step_interact( ~ Length:Diameter) %>%
  step_interact(~ Shucked_weight:Shell_weight) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
```

Question 4:

Creating and storing a linear regression object using the `lm` engine.

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

Question 5:

Now, we create a workflow by 1. setting up an empty workflow,
2. adding the model created in Question 4, and
3. adding the recipe created in Question 3

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

Question 6:

Using fit() object to predict the age of a hypothetical female abalone with `length` = 0.50, `diameter` = 0.10, `height` = 0.30, `whole_weight` = 4, `shucked_weight` = 1, `viscera_weight` = 2, `shell_weight` = 1.

```
# fitting the linear model
lm_fit <- fit(lm_wflow, abalone_train)

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

```
## # A tibble: 14 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     11.5      0.0376    304.   0
##  2 Length                           0.499    0.288       1.73 8.36e- 2
##  3 Diameter                         2.08     0.313       6.63 3.90e-11
##  4 Height                           0.245    0.0695      3.53 4.27e- 4
##  5 Whole_weight                     5.20     0.399      13.1  5.34e-38
##  6 Shucked_weight                  -4.38     0.258     -17.0  5.80e-62
##  7 Viscera_weight                  -0.995    0.160      -6.23 5.10e-10
##  8 Shell_weight                     1.70     0.217       7.86 5.03e-15
##  9 Sex_I                           -0.925    0.115      -8.08 8.79e-16
## 10 Sex_M                           -0.329    0.104      -3.15 1.67e- 3
## 11 Sex_I_x_Shucked_weight           0.506    0.0868      5.83 6.05e- 9
## 12 Sex_M_x_Shucked_weight           0.399    0.111       3.59 3.35e- 4
## 13 Length_x_Diameter               -2.85     0.407      -6.99 3.23e-12
## 14 Shucked_weight_x_Shell_weight   -0.264    0.210      -1.26 2.08e- 1
```

```
predicted_abalone <- predict(lm_fit, data.frame(Sex = 'F', Length = 0.50,
                                                Diameter = 0.10, Height = 0.30,
                                                Whole_weight = 4, Shucked_weight = 1,
                                                Viscera_weight = 2, Shell_weight = 1))
predicted_abalone
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.3
```

The predicted age of the specifications above is ≈ 23.32552.

Question 7:

Now I assess the model's performance using the `yardstick` package:

1. Create a metric set that includes R^2, RMSE (root mean squared error), and MAE (mean absolute error).
2. Use predict() and bind_cols() to create a tibble of the model's predicted values from the training data along with the actual observed ages (these are needed to assess your model's performance).
3. Finally, apply the metric set to the tibble, report the results, and interpret the R^2 value.

```
# creating a metric set
abalone_metrics <- metric_set(rmse, rsq, mae)

# creating a tibble of predicted values
abalone_train_res <- predict(lm_fit, new_data = abalone_train  %>% select(-Age, -Rings))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(Age))

# applying the metric set to the tibble
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       0.553
## 3 mae     standard       1.56
```

From above, we see that the RMSE ≈ 2.1706344, $R^2 \approx 0.5534585$, MAE ≈ 1.5611926. The estimate of R^2 implies that 53.50% of the variability in the response is explained by the predictors.