Homework 2

PSTAT 131/231

Linear Regression

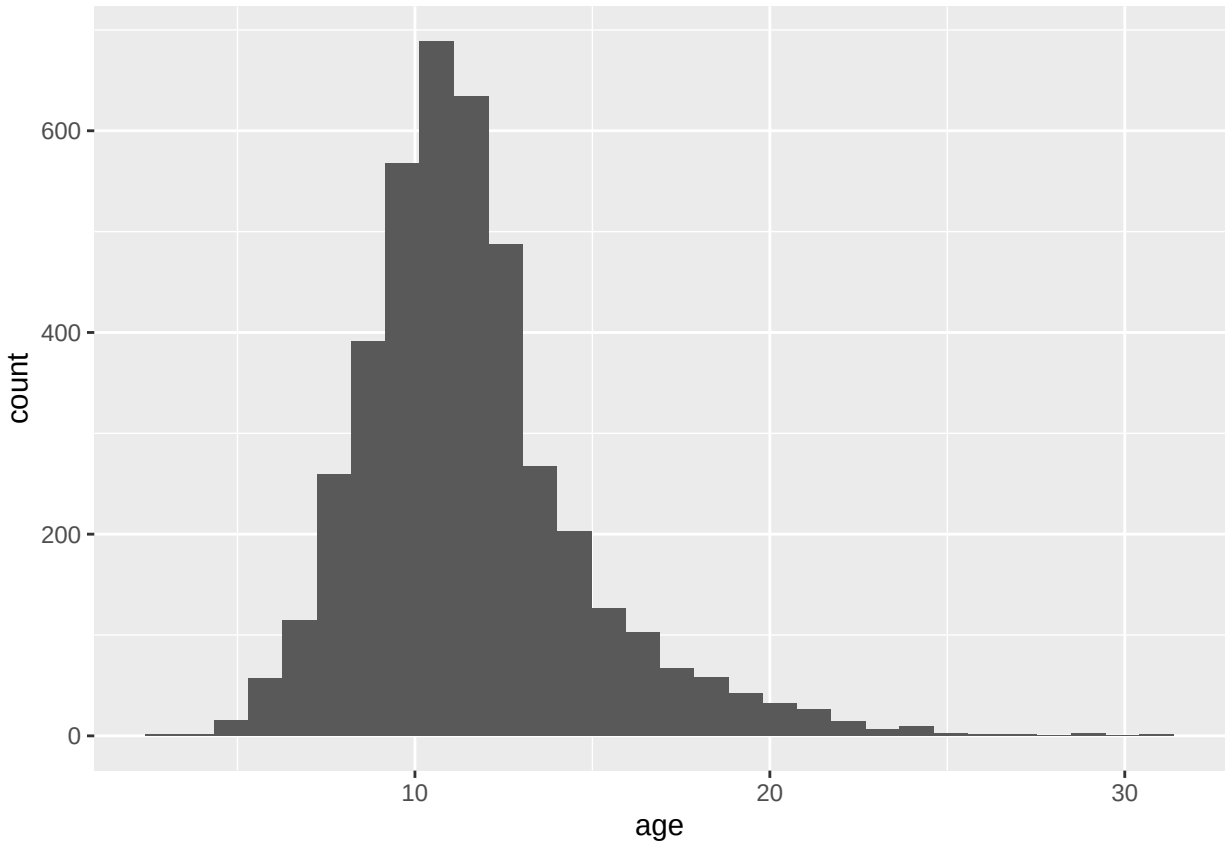
Question 1

Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no `age` variable in the data set. Add `age` to the data set.

Assess and describe the distribution of `age`.

```
# load packages
library(tidyverse)
library(tidymodels)
```

```
# load dataset
abalone <- read.csv("abalone.csv")
abalone <- abalone %>%
  mutate(age = rings + 1.5)
abalone %>%
  ggplot(aes(x = age)) +
  geom_histogram()
```



Solution: The distribution of age is positively skewed, meaning that much of the mass of its distribution is at the lower end, with a long tail to the right

Question 2

Split the abalone data into a training set and a testing set. Use stratified sampling. You should decide on appropriate percentages for splitting the data.

Remember that you'll need to set a seed at the beginning of the document to reproduce your results.

```
set.seed(2022)
abalone_split <- initial_split(abalone, prop = 0.8,
                               strate = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

Using the **training** data, create a recipe predicting the outcome variable, `age`, with all other predictor variables. Note that you should not include `rings` to predict `age`. Explain why you shouldn't use `rings` to predict `age`.

Solution: Because 'rings' and 'age' are linear dependent.

Steps for your 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, and
4. scale all predictors.

You'll need to investigate the `tidymodels` documentation to find the appropriate step functions to use.

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

Question 4

Create and store a linear regression object using the `"lm"` engine.

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

Question 5

Now:

1. set up an empty workflow,
2. add the model you created in Question 4, and
3. add the recipe that you created in Question 3.

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

Question 6

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

```
lm_fit <- fit(lm_wflow, abalone_train)
predict(lm_fit, new_data = 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))
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  23.1
```

Question 7

Now you want to assess your model's performance. To do this, use 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 your 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 your metric set to the tibble, report the results, and interpret the R^2 value.

```
# 1.
abalone_metrics <- metric_set(rsq, rmse, mae)

# 2.
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-rings))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))

# 3.
abalone_metrics(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.554
## 2 rmse    standard       2.14
## 3 mae     standard       1.53
```

Solution: 55.44% of the variation of the age can be explained by the linear model.