PSTAT 131 HW 2

Luke Todd

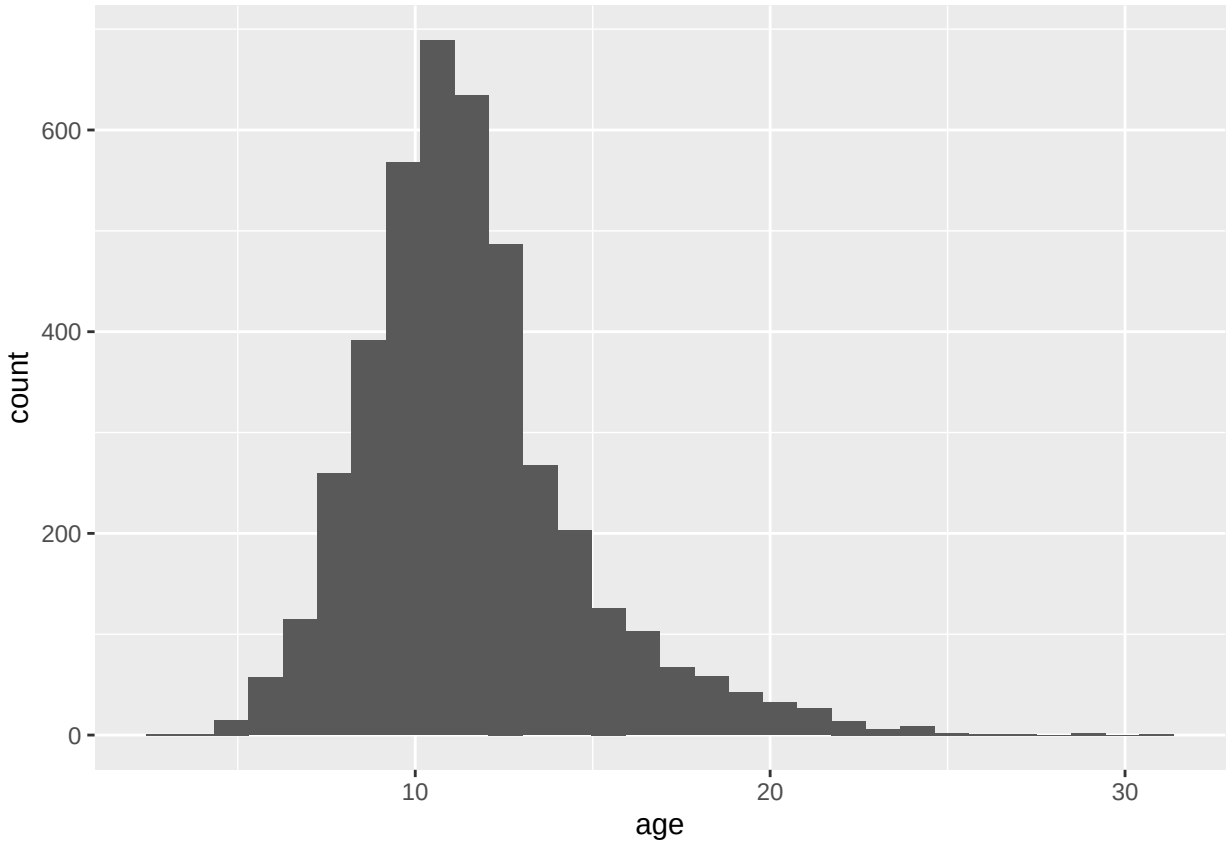
4/10/2022

Question 1

```
abalone <- abalone %>%
  mutate(age = rings + 1.5)

abalone %>%
  ggplot(aes(x = age)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
# hist(abalone$age)
```

The age of abalones appears to have a normal distribution.

Question 2

```r
set.seed(3068)

abalone_split <- initial_split(abalone, prop = 0.80,
                               strata = age)

abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Question 3

```r
# abalone_recipe <- recipe(age ~ type +
#                            longest_shell +
#                            diameter +
#                            height +
#                            whole_weight +
#                            shucked_weight +
#                            viscera_weight +
#                            shell_weight,
#                          data = abalone_train) %>%
#   step_dummy(all_nominal_predictors()) %>%
#   step_interact(terms = ~ type:shucked_weight +
#                   longest_shell:diameter +
#                   shucked_weight:shell_weight) %>%
#   step_center() %>%
#   step_scale()

abalone_recipe <- recipe(age ~ . , data = abalone_train) %>%
  step_rm(rings) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("type"):shucked_weight +
                  longest_shell:diameter +
                  shucked_weight:shell_weight) %>%
  step_normalize(all_predictors())

# step_normalize does step_center() and step_scale() in the same line of code
```

We should use rings to predict age because we are trying to figure out a better way to predict age that is not through the rings. Rings are logistically hard to measure.

Questions 4, 5, 6

```r
# Question 4

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

# Question 5

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

lm_fit <- fit(lm_wflow, abalone_train)

tib <- lm_fit %>%
  extract_fit_parsnip() %>%
  tidy()

# Question 6

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,
                       rings = 0)

predict(lm_fit, new_data = new_data)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  24.0
```

The hypothetical female abalone age would be around 13.6 years old.

Question 7

```r
metrics <- metric_set(rmse, rsq, mae)

abalone_train_res <- predict(lm_fit, new_data = abalone_train)

abalone_train_res %>% head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1  8.06
## 2  9.32
## 3 10.5
## 4 10.9
## 5  6.27
## 6  5.79
```

```r
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))

abalone_train_res %>% head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  8.06   8.5
## 2  9.32   9.5
## 3 10.5    8.5
## 4 10.9    9.5
## 5  6.27   6.5
## 6  5.79   6.5
```

```r
metrics(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.15
## 2 rsq     standard       0.558
## 3 mae     standard       1.54
```

```r
# multi_metric <- metric_set(rmse, rsq, mae)
# abalone_predict <- predict(abalone_fit, abalone_train) %>%
#   bind_cols(abalone_train %>% select(age))
# multi_metric(abalone_predict, truth = age, estimate = .pred)
```