Homework 2

PSTAT 131/231

Contents

Linear Regression 1

Linear Regression

```
library(ggplot2)
library(tidyverse)
library(tidymodels)
library(corrplot)
library(ggthemes)
library(yardstick)
tidymodels_prefer()
```

Question 1

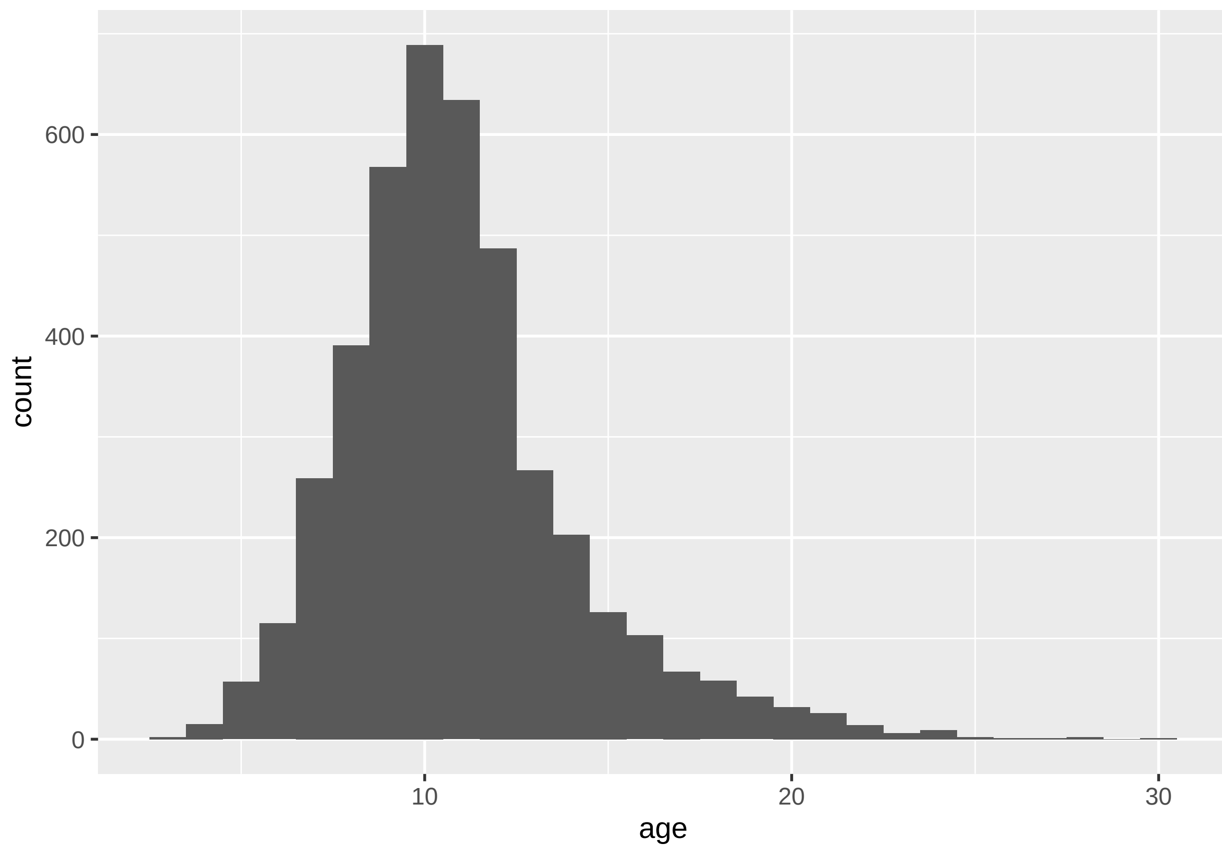
Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no `age` variable in the data set. Add `age` to the data set.

Assess and describe the distribution of `age`.

The age histogram is right skewed means that more of the abalone has age less than 15.

```
abalone <- read.csv(file = 'abalone.csv')
abalone['age']=abalone['rings']+1.5

ggplot(abalone, aes(x=age)) + geom_histogram(binwidth=1)
```



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.

```
set.seed(3435)

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


a <- c(16.5,8.5,10.5,11.5,8.5,9.5,21.5,17.5,10.5,20.5)


b <- scale(a,center=TRUE,scale=TRUE)
```

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

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`.

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.

```
# we are not include 'rings' to predict age because age is directly calculated from rings


abalone_recipe <- recipe(age ~ ., data = abalone_train) %>%
  step_rm('rings')%>%
  step_dummy(all_nominal_predictors())%>%
  step_interact(terms = ~type:shucked_weight)%>%
  step_interact(terms = ~longest_shell:diameter)%>%
  step_interact(terms = ~shucked_weight:shell_weight)%>%
  step_center(all_predictors())%>%
  step_scale(all_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.

```
myWorkflow <- 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.

```r
lm_fit <- fit(myWorkflow,abalone_train)

lm_fit %>%
  # This returns the parsnip object:
  extract_fit_parsnip() %>%
  # Now tidy the linear model object:
  tidy()
```

```
## # A tibble: 12 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                    11.4       0.0377   303.    0
##  2 longest_shell                   0.859     0.284      3.03  2.46e- 3
##  3 diameter                        2.37      0.310      7.66  2.51e-14
##  4 height                          0.251     0.0699     3.59  3.39e- 4
##  5 whole_weight                    4.24      0.389     10.9   2.80e-27
##  6 shucked_weight                 -3.66      0.240    -15.3   7.28e-51
##  7 viscera_weight                 -0.813     0.159     -5.11  3.44e- 7
##  8 shell_weight                    1.88      0.211      8.92  7.51e-19
##  9 type_I                         -0.332     0.0544    -6.11  1.13e- 9
## 10 type_M                          0.0272    0.0447     0.608 5.43e- 1
## 11 longest_shell_x_diameter       -3.26      0.389     -8.40  6.64e-17
## 12 shucked_weight_x_shell_weight  -0.229     0.201     -1.14  2.55e- 1
```

```r
predicted_age <- 11.42335329+0.5*0.85928463+0.1*2.37253150+0.3*0.25076110+
  4*4.24434255+1*-3.65543589+2*-0.81318578+1*1.88006719-0.33199999+0.02721407-
  3.26447671-0.22901120
```

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.

```r
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-c(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       2.18
## 2 rsq     standard       0.546
## 3 mae     standard       1.57
```

```
# In our model,r-squared is 0.5464 reveals that 54% of the data fit the regression model.
```

Required for 231 Students

In lecture, we presented the general bias-variance tradeoff, which takes the form:

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

where the underlying model $Y = f(X) + \epsilon$ satisfies the following:

- ϵ is a zero-mean random noise term and X is non-random (all randomness in Y comes from ϵ);
- (x_0, y_0) represents a test observation, independent of the training set, drawn from the same model;
- $\hat{f}(.)$ is the estimate of f obtained from the training set.

Question 8 Which term(s) in the bias-variance tradeoff above represent the reproducible error? Which term(s) represent the irreducible error?

Question 9 Using the bias-variance tradeoff above, demonstrate that the expected test error is always at least as large as the irreducible error.

Question 10 Prove the bias-variance tradeoff.

Hints:

- use the definition of $Bias(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$;
- reorganize terms in the expected test error by adding and subtracting $E[\hat{f}(x_0)]$