Code

Homework 2

PSTAT 131/231 Teo Zeng

Linear Regression

For this lab, we will be working with a data set from the UCI (University of California, Irvine) Machine Learning repository (see website here). The full data set consists of $4,177$ observations of abalone in Tasmania. (Fun fact: Tasmania supplies about 25% of the yearly world abalone harvest.)

Fig 1. Inside of an abalone shell.

The age of an abalone is typically determined by cutting the shell open and counting the number of rings with a microscope. The purpose of this data set is to determine whether abalone age (**number of rings + 1.5**) can be accurately predicted using other, easier-to-obtain information about the abalone.

The full abalone data set is located in the `\data` subdirectory. Read it into *R* using `read_csv()`. Take a moment to read through the codebook (`abalone_codebook.txt`) and familiarize yourself with the variable definitions.

Make sure you load the `tidyverse` and `tidymodels`!

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

setwd("~/Desktop/PSTAT 131/pstat131-hw2")
set.seed(125)
```

Question 1

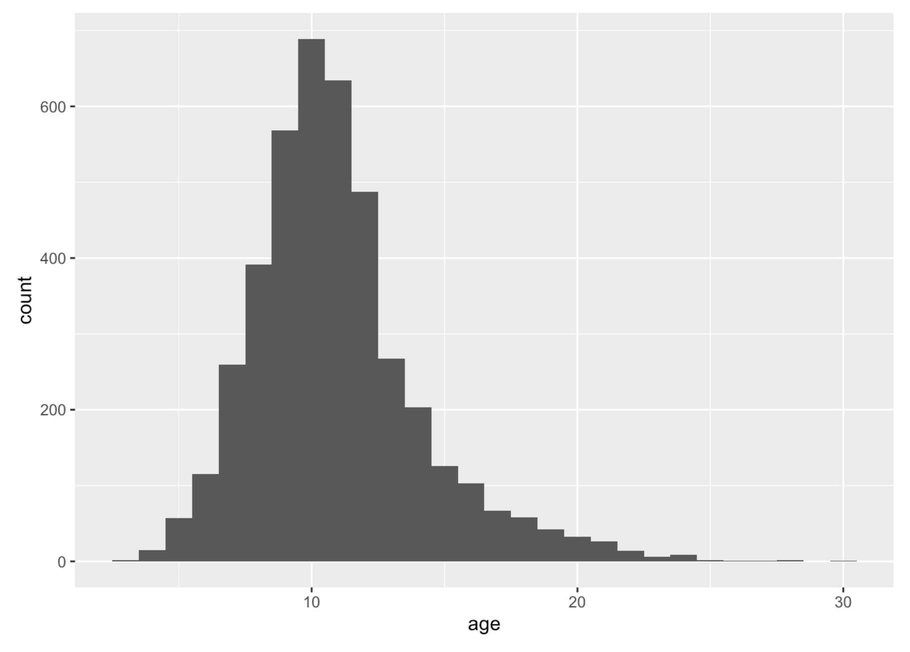
Your goal is to predict abalone age, which is calculated as the number of rings plus 1.5. Notice there currently is no `age` variable in the data set. Add `age` to the data set.

Assess and describe the distribution of `age`.

```r
abalone <- read.csv(file = 'data/abalone.csv')
abalone %>% head()
```

```
##   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
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

```r
abalone["age"] <- abalone["rings"] + 1.5
ggplot(abalone, aes(x=age)) + geom_histogram(binwidth = 1)
```



From the histogram we see that age has a left-skewed distribution, and it has its peak at the age of ~11

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.

```r
abalone2 <- subset(abalone, select = -rings)
abalone_split <- initial_split(abalone2, prop = 0.80,strata = 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`.

we should not use rings to predict age because age is dependent on rings (rings + 1.5 = age). If rings is included, then age can be 100 percent explained by rings.

Steps for your recipe:

1. dummy code any categorical predictors

```r
abalone_recipe <- recipe(age ~ ., data = abalone_train) %>% step_dummy(all_nominal_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.

```r
int_mod <- abalone_recipe %>%
           step_interact(terms = ~ starts_with("type"):shucked_weight +
                                             longest_shell:diameter +
                                             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.

```r
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.

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

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(lm_wflow, abalone_train)
hypothetical_female <- data.frame(longest_shell = 0.50,
                                  diameter = 0.10,
                                  height = 0.30,
                                  whole_weight = 4,
                                  shucked_weight = 1,
                                  viscera_weight = 2,
                                  shell_weight = 1,
                                  type = "F")
predict(lm_fit, hypothetical_female)
```

```
## # A tibble: 1 × 1
##   .pred
##   <dbl>
## 1  24.7
```

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.

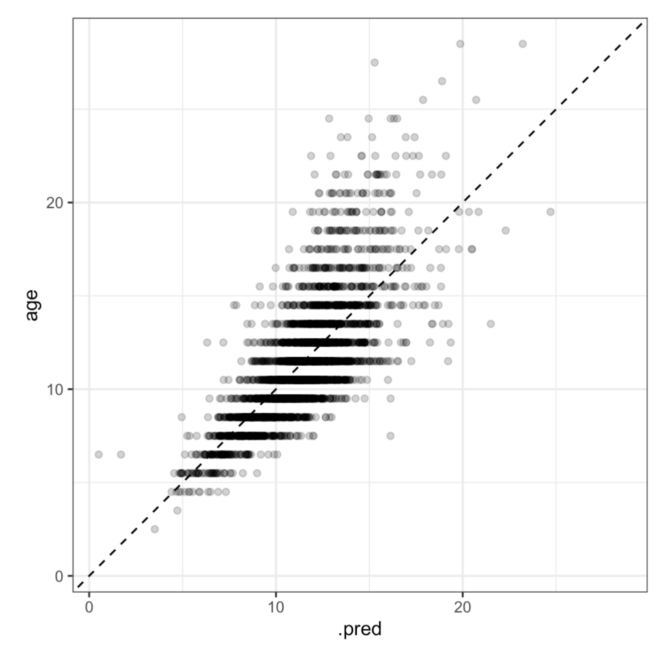
```r
abalone_train_res <- predict(lm_fit, new_data = abalone_train %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abalone_train %>% select(age))
rmse(abalone_train_res, truth = age, estimate = .pred)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.12
```

```r
abalone_metrics <- metric_set(rmse, rsq, mae)
abalone_metrics(abalone_train_res, truth = age,
                estimate = .pred)
```

```
## # A tibble: 3 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.12
## 2 rsq     standard       0.556
## 3 mae     standard       1.52
```

```r
abalone_train_res %>%
  ggplot(aes(x = .pred, y = age)) +
  geom_point(alpha = 0.2) +
  geom_abline(lty = 2) +
  theme_bw() +
  coord_obs_pred()
```



Evaluated on the test data, our model performs moderately based on the R-squared criterion. At an R-squared of about .556, we have that 55.6% of the variability in the response is explained by the predictors, which is a moderate correlation. We have a RMSE of 2.12 and MAE of 1.52, which are both small and acceptable. So this model can make relatively good prediction on age.

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?

the **reducible error** terms are $[Bias(\hat{f}(x_0))]^2$ and $Var(\hat{f}(x_0))$, and the **irreducible error** is $Var(\epsilon)$

Question 9

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

with the formula above, we have

$$E[(y_0-\hat{f}(x_0))^2] = \underbrace{Var(\hat{f}(x_0))}_{\geq 0} + \underbrace{[Bias(\hat{f}(x_0))]^2}_{\geq 0} + Var(\epsilon)$$

so we have 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 $E[\hat{f}(x_0)]$

First, recall that, by definition, for any random variable X, we have

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2.$$

Rearranging, we get:

$$\mathrm{E}\left[X^2\right] = \mathrm{Var}[X] + \mathrm{E}[X]^2$$

Since f is deterministic,

$$\mathrm{E}[f] = f.$$

Thus, given $y = f + \varepsilon$ and $\mathrm{E}[\varepsilon] = 0$ (because ε is noise), implies $\mathrm{E}[y] = \mathrm{E}[f+\varepsilon] = \mathrm{E}[f] = f$. Also, since $\mathrm{Var}[\varepsilon] = \sigma^2$,

$$\mathrm{Var}[y] = \mathrm{E}\left[(y-\mathrm{E}[y])^2\right] = \mathrm{E}\left[(y-f)^2\right] = \mathrm{E}\left[(f+\varepsilon-f)^2\right] = \mathrm{E}\left[\varepsilon^2\right] = \mathrm{Var}[\varepsilon] + \mathrm{E}[\varepsilon]^2 = \sigma^2 + 0^2 = \sigma^2$$

Thus, since ε and $\hat{f}$ are independent, we can write

$$\begin{aligned}
\mathrm{E}\left[(y-\hat{f})^2\right] &= \mathrm{E}\left[(f+\varepsilon-\hat{f})^2\right] \\
&= \mathrm{E}\left[(f+\varepsilon-\hat{f}+\mathrm{E}[\hat{f}]-\mathrm{E}[\hat{f}])^2\right] \\
&= \mathrm{E}\left[(f-\mathrm{E}[\hat{f}])^2\right] + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] + 2\mathrm{E}[(f-\mathrm{E}[\hat{f}])\varepsilon] + 2\mathrm{E}[\varepsilon(\mathrm{E}[\hat{f}]-\hat{f})] + 2\mathrm{E}[(\mathrm{E}[\hat{f}]-\hat{f})(f-\mathrm{E}[\hat{f}])] \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] + 2(f-\mathrm{E}[\hat{f}])\mathrm{E}[\varepsilon] + 2\mathrm{E}[\varepsilon]\mathrm{E}[\mathrm{E}[\hat{f}]-\hat{f}] + 2\mathrm{E}[\mathrm{E}[\hat{f}]-\hat{f}](f-\mathrm{E}[\hat{f}]) \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}[\hat{f}] \\
&= \mathrm{Bias}[\hat{f}]^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}[\hat{f}] \\
&= \mathrm{Bias}[\hat{f}]^2 + \sigma^2 + \mathrm{Var}[\hat{f}]
\end{aligned}$$