# 231-hw2

Rita Han

2022-10-12

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

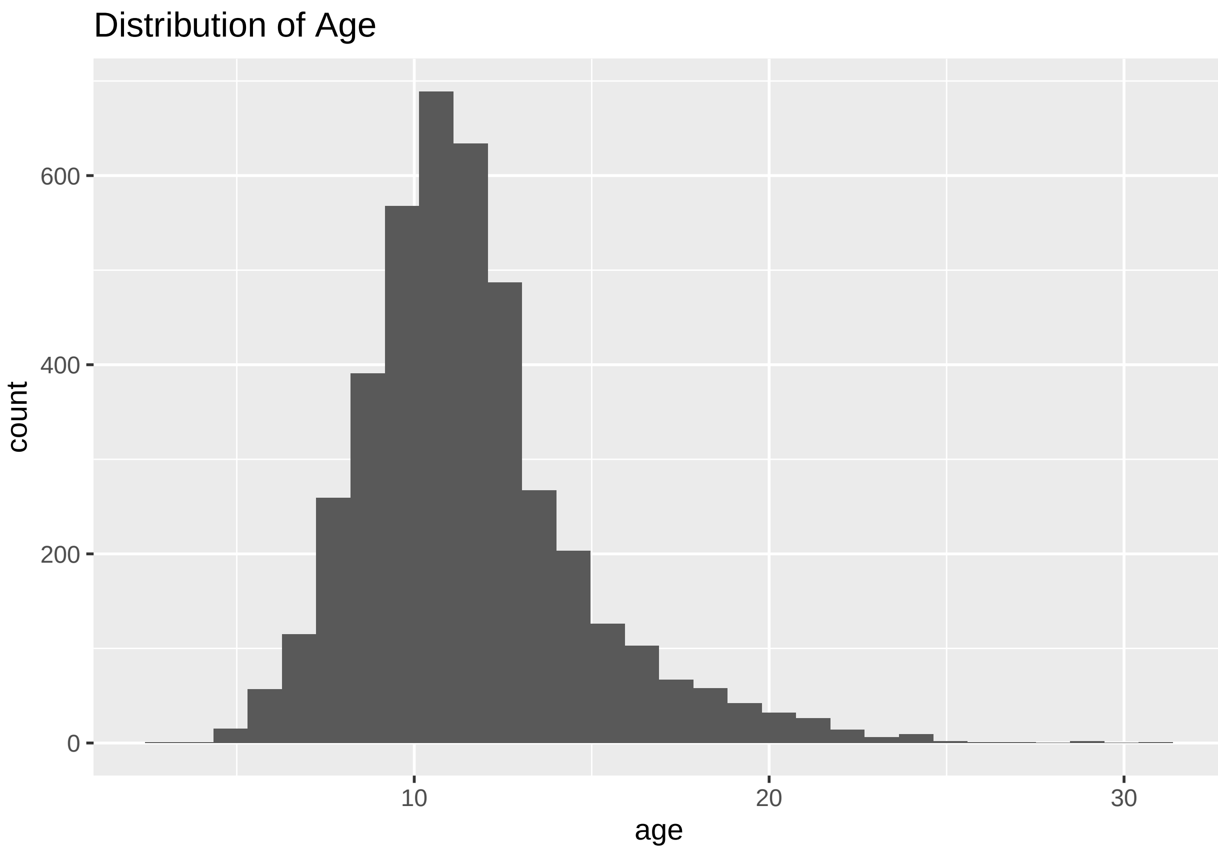
```r
abalone = read_csv(file= "/Users/ritahan/Desktop/pstat131/gauchospace/homework-2/data/abalone.csv")
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): type
## dbl (8): longest_shell, diameter, height, whole_weight, shucked_weight, visc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#Q1
abalone['age']=abalone$rings+1.5
```

```r
abalone %>%
  ggplot(aes(x=age))+geom_histogram()+labs(title='Distribution of Age')
```

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




The plot shows a positively skewed normally distribution. The majority of abalone are between 9-13 years old.

```r
#Q2
set.seed(4500)
abalone_split <- initial_split(abalone, prop = 0.80, strata = age)
abalone_train <- training(abalone_split)
abalone_test <- testing(abalone_split)
```

Q3. We should not use rings to predict age, because age is calculated directly by rings.

```r
#Q3
update_abalone_train=abalone_train %>%
  select(-rings)
abalone_recipe <- recipe(age ~ . , data = update_abalone_train) %>%
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>%
  step_interact(~ starts_with("type"):shucked_weight+
```

```
                  longest_shell:diameter
                +shucked_weight:shell_weight) %>%
  step_normalize(all_predictors())
abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("type"):shucked_weight + longest_shell...
## Centering and scaling for all_predictors()
```

```
#Q4
lm_model=linear_reg() %>%
  set_engine("lm")
```

```
#Q5
lm_wflow= workflow() %>%
  add_model(lm_model) %>%
  add_recipe(abalone_recipe)
```

```
abalone_fit=fit(lm_wflow, update_abalone_train)
abalone_fit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 16 x 5
##    term                         estimate std.error statistic   p.value
##    <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                    11.4      0.0372   308.      0
##  2 longest_shell                   0.566    0.286      1.98    4.80e- 2
##  3 diameter                        1.97     0.314      6.27    4.17e-10
##  4 height                          0.270    0.0694     3.89    1.03e- 4
##  5 whole_weight                    5.14     0.395     13.0     1.03e-37
##  6 shucked_weight                 -4.07     0.252    -16.1     2.24e-56
##  7 viscera_weight                 -1.02     0.157     -6.50    9.17e-11
##  8 shell_weight                    1.44     0.220      6.54    7.20e-11
##  9 type_F                          0.361    0.0991     3.64    2.72e- 4
## 10 type_I                         -0.654    0.0991    -6.60    4.74e-11
## 11 type_M                         NA       NA         NA      NA
## 12 type_F_x_shucked_weight        -0.357    0.103     -3.47    5.22e- 4
## 13 type_I_x_shucked_weight         0.362    0.0807     4.48    7.55e- 6
## 14 type_M_x_shucked_weight        NA       NA         NA      NA
## 15 longest_shell_x_diameter       -2.64     0.407     -6.47    1.09e-10
## 16 shucked_weight_x_shell_weight  -0.128    0.206     -0.621   5.35e- 1
```

```r
#Q6
abalone_fit=fit(lm_wflow, update_abalone_train)
tibble_abalone=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)
predict(abalone_fit, new_data=tibble_abalone)
```

```
## Warning in predict.lm(object = object$fit, newdata = new_data, type =
## "response"): prediction from a rank-deficient fit may be misleading

## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  22.0
```

The predicted age of a hypothetical female abalone is 22.

```r
#Q7
metrics=metric_set(rsq, rmse, mae)
abalone_predict=predict(abalone_fit, update_abalone_train)
```

```
## Warning in predict.lm(object = object$fit, newdata = new_data, type =
## "response"): prediction from a rank-deficient fit may be misleading
```

```r
abalone_predict_result=bind_cols(abalone_predict, update_abalone_train %>% select(age))
abalone_predict_result
```

```
## # A tibble: 3,340 x 2
##     .pred   age
##     <dbl> <dbl>
##  1   9.44   8.5
##  2   8.11   8.5
##  3   9.39   9.5
##  4  10.3    8.5
##  5   6.30   6.5
##  6   5.96   5.5
##  7   8.58   8.5
##  8  11.9    8.5
##  9   7.73   7.5
## 10  11.2    9.5
## # ... with 3,330 more rows
```

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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.552
## 2 rmse    standard       2.15
## 3 mae     standard       1.54
```

The rmse is 2.15, the mae is 1.54. The $R^2$ is 0.552, this means 55.2% of variable fit the model and can be explained by the predictors.

Question 8:
$Var(\epsilon)$ represent the irreducible error and $Var(\hat{f}(x_0))$ and $[Bias(\hat{f}(x_0))]^2$ represent the reproducible error.
Question 9:
$Var(\epsilon)$ is the minimum lower bound for the LHS, which is irreducible, so that $E\left[\left(y_0 - \hat{f}\left(x_0\right)\right)^2\right]$ can not be less than $Var(\epsilon)$. In other words, the expected test error is always at least as large as the irreducible error.
Question 10:

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

---

$$E[(y_0 - \hat{f}(x_0))^2] = E[(f(x) + \varepsilon - \hat{f}(x_0))^2] \quad \because y_0 = f(x) + \varepsilon$$

$$= E[(f(x_0) - \hat{f}(x_0))^2] + E[\varepsilon^2] + 2E[(f(x_0) - \hat{f}(x_0))\varepsilon]$$

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

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

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

$$- \cancel{2E[(f(x_0) - E[\hat{f}(x_0)])(E[\hat{f}(x_0)] - E[\hat{f}(x_0)])]} + Var(\varepsilon)$$

$$= \underbrace{(E[\hat{f}(x_0)] - f(x_0))^2}_{bias(\hat{f}(x_0))^2} + \underbrace{E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2]}_{Var(\hat{f}(x_0))} + Var(\varepsilon)$$

Figure 1: picture