160A HW2

Tianyi Li

2022-04-08

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 0.2.0 --
```

```
## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0
```

```
## -- 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/
```

```
library(readr)
library(dplyr)
```

```
abalone <- read_csv("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.
```

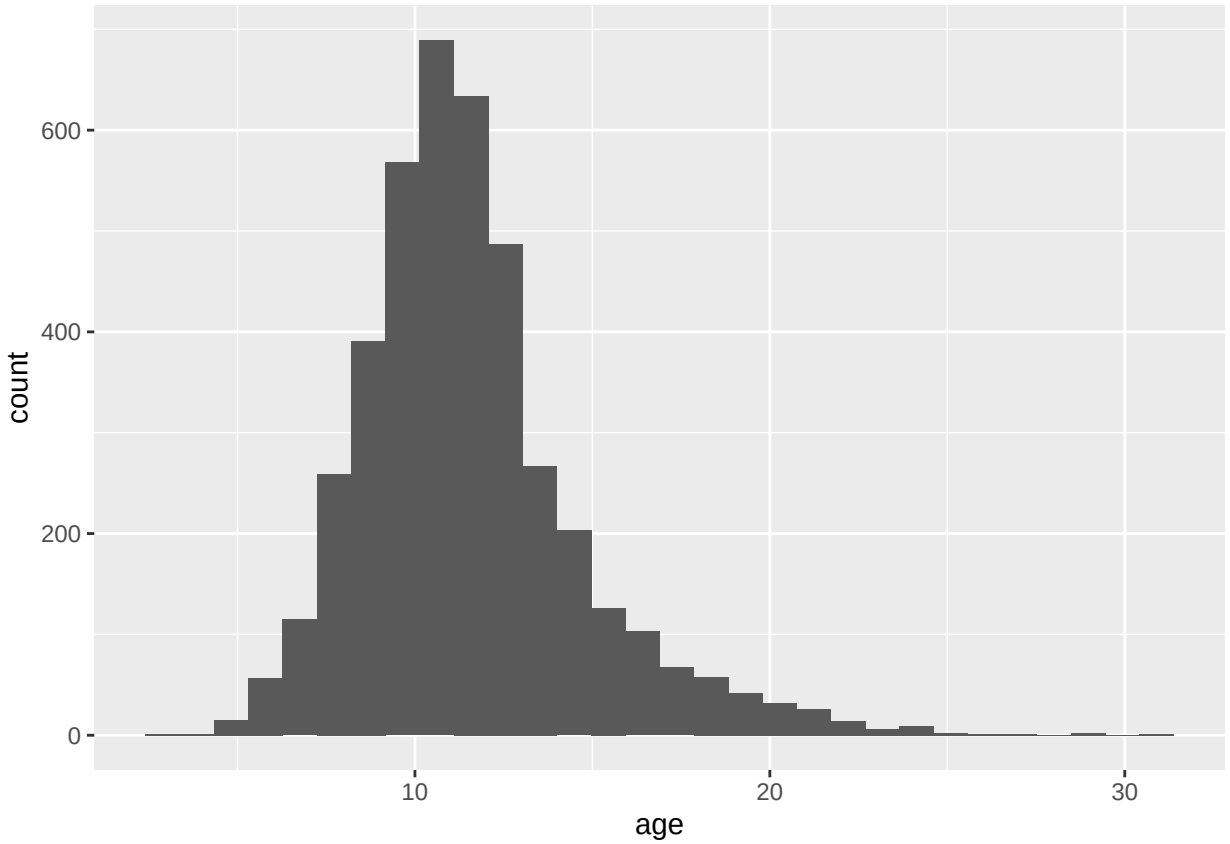
```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(abalone)
```

```
## # A tibble: 6 x 9
##   type  longest_shell diameter height whole_weight shucked_weight viscera_weight
##   <chr>         <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M             0.455    0.365  0.095        0.514         0.224          0.101
## 2 M             0.35     0.265  0.09         0.226         0.0995         0.0485
## 3 F             0.53     0.42   0.135        0.677         0.256          0.142
## 4 M             0.44     0.365  0.125        0.516         0.216          0.114
## 5 I             0.33     0.255  0.08         0.205         0.0895         0.0395
## 6 I             0.425    0.3    0.095        0.352         0.141          0.0775
## # ... with 2 more variables: shell_weight <dbl>, rings <dbl>
```

Question 1

```
abalone["age"] <- abalone["rings"]+1.5
ggplot(abalone,aes(x=age))+geom_histogram(bins=30)
```



The age presents a normal distribution, with a peak appears about age at 11.

Question 2

```
set.seed(1202)
abalone_split<-initial_split(abalone,prop=0.80,
                             strata = age )
abalone_train<-training(abalone_split)
abalone_test<-testing(abalone_split)
```

Question 3

```
abalone_training<- abalone_train %>% select(-rings)
abalone_recipe<-recipe(age ~ ., data=abalone_training) %>%
  step_dummy(all_nominal_predictors())
```

```
abalone_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())
abalone_mod
```

```
## 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 for all_predictors()
## Scaling for all_predictors()
```

Question 4

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

Question 5

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

Question 6

```
lm_fit<-fit(lm_wflow,abalone_training)
female_abalone_age<-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(lm_fit,new_data=female_abalone_age)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  14.1
```

```
lm_fit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 10 x 5
##    term           estimate std.error statistic  p.value
```

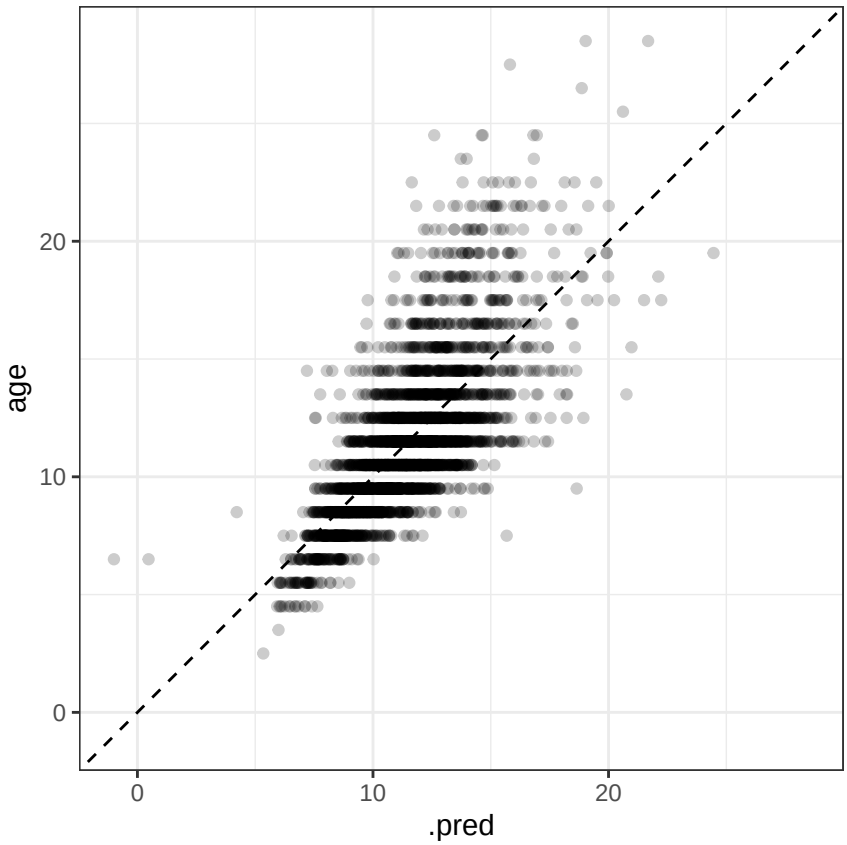
```
##    <chr>             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)      5.53      0.328     16.9   2.31e-61
##  2 longest_shell   -0.700     2.00      -0.350 7.27e- 1
##  3 diameter        11.5       2.45       4.70  2.66e- 6
##  4 height           9.40      1.64       5.72  1.17e- 8
##  5 whole_weight     8.22      0.782     10.5   1.75e-25
##  6 shucked_weight -19.1       0.892    -21.5   7.43e-96
##  7 viscera_weight  -9.12      1.41      -6.47  1.11e-10
##  8 shell_weight     9.44      1.21       7.77  1.01e-14
##  9 type_I          -0.865     0.114     -7.60  3.85e-14
## 10 type_M           0.0266    0.0930     0.286 7.75e- 1
```

Question 7

```
library(yardstick)
abalone_train_res <- predict(lm_fit, new_data = abalone_training %>% select(-age))
abalone_train_res <- bind_cols(abalone_train_res, abalone_training %>% select(age))
abalone_train_res %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.37   8.5
## 2  8.25   8.5
## 3  9.33   9.5
## 4  9.73   8.5
## 5 10.2    8.5
## 6  9.95   9.5
```

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




```
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       2.19
## 2 rsq     standard       0.535
## 3 mae     standard       1.59
```

The R-squared of approximately 53% shows that 53% of the data fit the regression model.