Contents

Pstat 131 Hw 2

Noe Arambula ID: 3561131

4/5/22

Libraries

```
# install.packages('ggthemes')
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()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggthemes)
tidymodels_prefer()
library(readr)
abalone <- read_csv("abalone.csv")
```

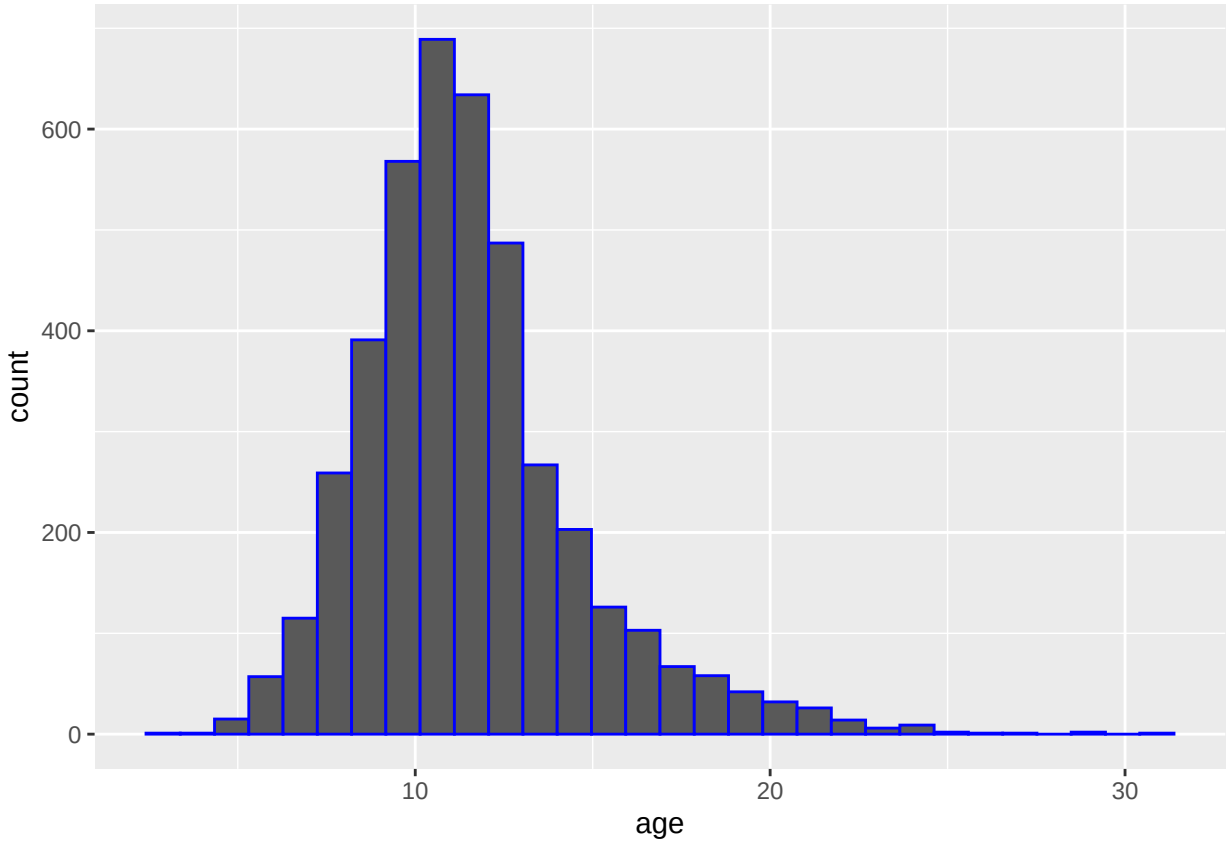
```
## Rows: 4177 Columns: 9

## -- 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.
```

Question 1

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

  ## Distribution of age
ggplot(new_abalone, aes(x = age)) + geom_histogram(color = 'blue', bins = 30)
```



The distribution of age is a little positively skewed most of the points are centered around age 11

Question 2

```
set.seed(777)

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

Question 3

```
simple_abalone_recipe = recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_

simple_abalone_recipe
```

```
## Recipe
##
## Inputs:
```

```
##
##       role #variables
##    outcome          1
##  predictor          8
```

```
  ## we did not include rings to predict age because age is just rings + 1.5 this would scew the data
abalone_recipe <- recipe(age ~ type + longest_shell + diameter + height + whole_weight + shucked_weight
  step_dummy(all_nominal_predictors()) %>%   # creates dummy variables
  step_normalize(all_predictors()) %>%    # centers and scales all predictors
  step_interact(terms = type ~ shucked_weight) %>%
  step_interact(terms = longest_shell ~ diameter) %>%
  step_interact(terms = shucked_weight ~ shell_weight)

abalone_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Centering and scaling for all_predictors()
## Interactions with type, shucked_weight
## Interactions with longest_shell, diameter
## Interactions with shucked_weight, shell_weight
```

Question 4

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

Question 5

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

lm_fit <- fit(lm_wflow, abalone_train)
```

```
## Warning: Interaction specification failed for: type ~ shucked_weight. No
## interactions will be created.
```

```
head(lm_fit)
```

```
## $pre
## $actions
## $actions$recipe
## $recipe
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Dummy variables from all_nominal_predictors()
## Centering and scaling for all_predictors()
## Interactions with type, shucked_weight
## Interactions with longest_shell, diameter
## Interactions with shucked_weight, shell_weight
##
## $blueprint
## Recipe blueprint:
##
## # Predictors: 0
##   # Outcomes: 0
##    Intercept: FALSE
## Novel Levels: FALSE
##  Composition: tibble
##
## attr(,"class")
## [1] "action_recipe" "action_pre"    "action"
##
##
## $mold
## $mold$predictors
## # A tibble: 3,340 x 9
##    longest_shell diameter height whole_weight shucked_weight viscera_weight
##            <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
##  1        -1.45    -1.44  -1.17        -1.23          -1.17          -1.21
##  2        -1.62    -1.54  -1.40        -1.27          -1.22          -1.29
##  3        -1.41    -1.29  -1.29        -1.10          -1.19          -1.29
##  4        -0.492   -0.535 -0.814       -0.714         -0.596         -0.515
##  5        -0.617   -0.535 -0.814       -0.626         -0.551         -0.583
##  6        -2.37    -2.35  -2.23        -1.55          -1.48          -1.43
##  7        -2.66    -2.60  -1.99        -1.61          -1.51          -1.51
##  8        -2.62    -2.60  -2.11        -1.61          -1.54          -1.53
##  9        -1.12    -1.14  -1.05        -1.28          -1.23          -1.24
## 10        -0.534   -0.333 -0.460       -0.752         -0.820         -0.643
## # ... with 3,330 more rows, and 3 more variables: shell_weight <dbl>,
## #   type_I <dbl>, type_M <dbl>
##
```

```
## $mold$outcomes
## # A tibble: 3,340 x 1
##      age
##    <dbl>
##  1   8.5
##  2   8.5
##  3   8.5
##  4   9.5
##  5   9.5
##  6   6.5
##  7   6.5
##  8   5.5
##  9   8.5
## 10   8.5
## # ... with 3,330 more rows
##
## $mold$blueprint
## Recipe blueprint:
##
## # Predictors: 8
##   # Outcomes: 1
##    Intercept: FALSE
## Novel Levels: FALSE
##  Composition: tibble
##
## $mold$extras
## $mold$extras$roles
## NULL
##
##
##
## attr(,"class")
## [1] "stage_pre" "stage"
##
## $fit
## $actions
## $actions$model
## $spec
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
##
##
## $formula
## NULL
##
## attr(,"class")
## [1] "action_model" "action_fit"   "action"
##
##
## $fit
## parsnip model object
##
##
```

```
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##    (Intercept)   longest_shell        diameter          height    whole_weight
##        11.4476         -0.2798          1.3061          0.4166          4.7863
## shucked_weight  viscera_weight    shell_weight          type_I          type_M
##        -4.5335         -1.2754          1.1633         -0.3547          0.0510
##
##
## attr(,"class")
## [1] "stage_fit" "stage"
##
## $post
## $actions
## list()
##
## attr(,"class")
## [1] "stage_post" "stage"
##
## $trained
## [1] TRUE
```

Question 6

```
 # variables represents the new data we will be using to predict an age
variables = data.frame(type = 'F',longest_shell = 0.50, diameter = 0.10, height = 0.30, whole_weight = 4
predict(lm_fit, new_data = variables)
```

```
## # A tibble: 1 x 1
##   .pred
##   <dbl>
## 1  12.2
```

Question 7

```
  ## Creating a metric set
abalone_metrics = metric_set(rmse, rsq, mae)

  ## My predictions of the data
abalone_pred <- predict(lm_fit, new_data = abalone_train %>% select(-age))

abalone_pred = bind_cols(abalone_pred, abalone_train %>% select(age))

abalone_pred %>%
  head()
```

```
## # A tibble: 6 x 2
```

```
##   .pred   age
##   <dbl> <dbl>
## 1  9.33   8.5
## 2  8.26   8.5
## 3  9.81   8.5
## 4  9.85   9.5
## 5 10.2    9.5
## 6  6.84   6.5
```

```
  ## apply your metric set to the tibble, report the results, and interpret the R^2 value.
abalone_metrics(abalone_pred, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2.18
## 2 rsq     standard       0.548
## 3 mae     standard       1.57
```

The r-squared value is the goodness of fit measure and measures the relationship between my model and the dependent variable, in this case age.Here the r-squared value is 0.548 for my model meaning 54.8% of the variance of age was explained by my model. This suggests that this model was not that great at estimating values, although still some correlation.