Homework 2

Mason Wong

2022-10-16

```
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   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("tidymodels")
```

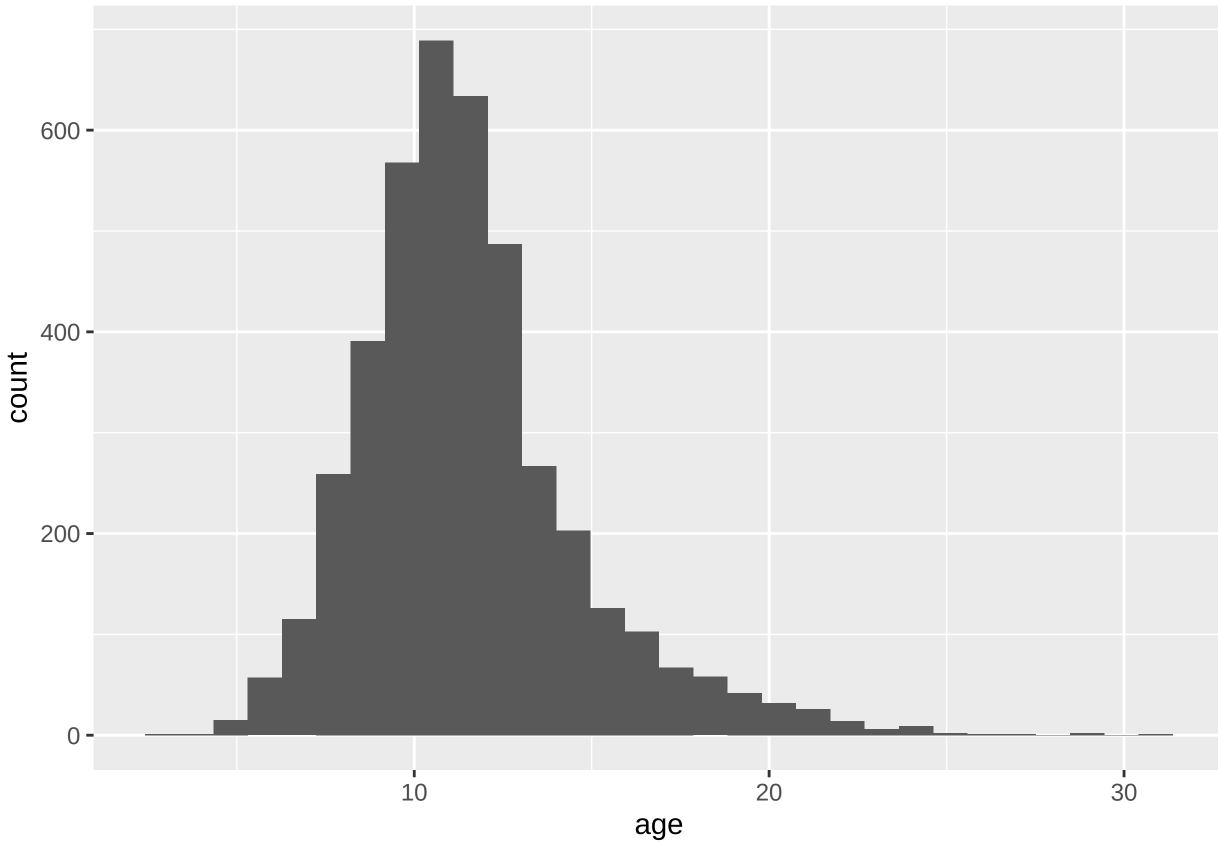
```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      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()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library("yardstick")
library("ggplot2")

abs <- read.csv("abalone.csv")
```

Question 1

```
age <- abs$rings + 1.5
```

```r
abs$age <- age

ggplot(abs, aes(x=age)) + geom_histogram(bins=30)
```



Since age is calculated using rings + 1.5, we can say that age is definitely dependent on ring's distribution. Looking at the above data, age follows a normal distribution

Question 2

```r
set.seed(100)

abs_split <- initial_split(abs, prop = 0.7)

abs_train <- training(abs_split)

abs_test <- testing(abs_split)
```

Question 3

```r
abs_recipe <- recipe(age~ type + longest_shell + diameter + height + whole_weight + shucked_weight + vi
  step_dummy(all_nominal_predictors()) %>%
```

```
  step_interact(~ starts_with("type"):shucked_weight) %>%
  step_interact(~ longest_shell:diameter) %>%
  step_interact(~ shucked_weight:shell_weight)


abs_recipe <- step_center(recipe = abs_recipe, longest_shell, diameter, height, whole_weight, shucked_w

abs_recipe <- step_scale(recipe = abs_recipe, longest_shell, diameter, height, whole_weight, shucked_we
```

Since age is already dependent on rings, we shouldn't use rings as a predictor variable, or else the other predictor variables would just be equal to the 1.5 that is added to rings to make age.

Question 4

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

Question 5

```
abs_wrkflw <- workflow() %>%
  add_model(abs_lm) %>%
  add_recipe(abs_recipe)
```

Question 6

```
abs_fit <- fit(abs_wrkflw, abs_train)

abs_fit
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 6 Recipe Steps
##
## * step_dummy()
## * step_interact()
## * step_interact()
## * step_interact()
## * step_center()
## * step_scale()
##
## -- Model -----------------------------------------------------------------------
##
## Call:
```

```
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##                 (Intercept)                 longest_shell
##                     19.5740                        0.8466
##                    diameter                        height
##                      2.3050                        0.2253
##                whole_weight                shucked_weight
##                      5.2033                       -4.5681
##              viscera_weight                  shell_weight
##                     -1.0500                        1.4841
##                      type_I                        type_M
##                     -2.0436                       -0.6928
##    type_I_x_shucked_weight       type_M_x_shucked_weight
##                      4.3375                        1.6181
##   longest_shell_x_diameter  shucked_weight_x_shell_weight
##                    -34.9207                        0.4532
```

Our linear regression equation is now

$Y = 0.8466B_{LShell} + 2.305B_{diam} + 0.2253B_{height} + 5.2033B_{Wweight} - 4.5681B_{Sweight} - 1.05B_{Vweight} + 1.4841B_{ShellWeight} - 2.0436$

Using the given values to predict age we get:

```r
Y = (0.8466 * 0.5) + (2.305 * 0.1) + (0.2253 * 0.3) + (5.2033 * 4) - (4.5618 * 1) - (1.05 * 2) + (1.484

Y
```

```
## [1] 34.63806
```

Question 7

```r
abs_metrics <- metric_set(rsq, rmse, mae)

abs_RMSE <- predict(abs_fit, new_data = abs_train %>% select(-age))

abs_RMSE %>%
  head()
```

```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1 12.8
## 2 15.7
## 3  7.16
## 4 12.5
## 5 11.4
## 6 12.2
```

```r
abs_RMSE <- bind_cols(abs_RMSE, abs_train %>% select(age))

abs_RMSE %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1 12.8   10.5
## 2 15.7   16.5
## 3  7.16   7.5
## 4 12.5   11.5
## 5 11.4   10.5
## 6 12.2   10.5
```

```r
abs_metrics(abs_RMSE, truth = age, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.553
## 2 rmse    standard       2.17
## 3 mae     standard       1.55
```

Our model as an RMSE of 2.17, an R^2 value of 0.553 and an MAE of 1.554

An R^2 value of 55.3% means that our model explains a little more than half of the variability of the response data around the mean

Question 8

In the equation

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

The reducible error is represented by the first two parts of the equation being $Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))^2]$

While the irreducible error is represented by $Var(\epsilon)$

Question 9

Irreducible error for a model is a constant value as long as we are sampling from the same data set with the same model. Given that the Bias-Variance tradeoff equation is:

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

We can see that we always add $Var(\epsilon)$ or irreducible error. In the case of a model that is a perfect fit, both Variance and Bias are 0. This would give us the lowest possible value for our Bias-Variance tradeoff and yet, the lowest will alwasy be equal to $Var(\epsilon)$

Question 10

Hint way

To prove that the equation holds we can start with $E[y - \hat{f}(x_0)] = E[(f(x_0) + \epsilon - \hat{f}(x_0))^2]$

This results in $E[(f(x_0) - \hat{f}(x_0))^2] - \epsilon$ if we expand the equation

Ignoring Epsilon we want to add abd subtract $E[\hat{f}(x_0)]$ from both elements of the expected value function and thus get

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

which yields

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

Definition of Bias is going back to the definition of bias we know that:

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

and Variance can be translated as:

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

Thus we have created a function of $Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$ with the leftover values being assumed as ϵ

My dumb way

We begin with the equation

$$Bias(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$$

Taking into account that we are dealing with $Bias^2$ we have

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

If we expand this we get

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

As we can see in the above equation, the larger $E[\hat{f}(x_0)]^2$ is the larger Bias can be, since Bias is a factor of $E[\hat{f}(x_0)], f(x_0), and - 2E[\hat{f}(x_0)]f(x_0)$

Thus we can conclude that as $f(x_0)$ increases, so does $E[\hat{f}(x_0)]$.

However as we take a look at the Variance equation $E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2$ we can see that a larger value for $E[\hat{f}(x_0)]^2$ decreases the overall output of variance as it subtracts more and more from $E[\hat{f}(x_0)^2]$.

Thus we can conclude that as Bias increases in value, Variance decreases