Rene Guerra - Homework 2

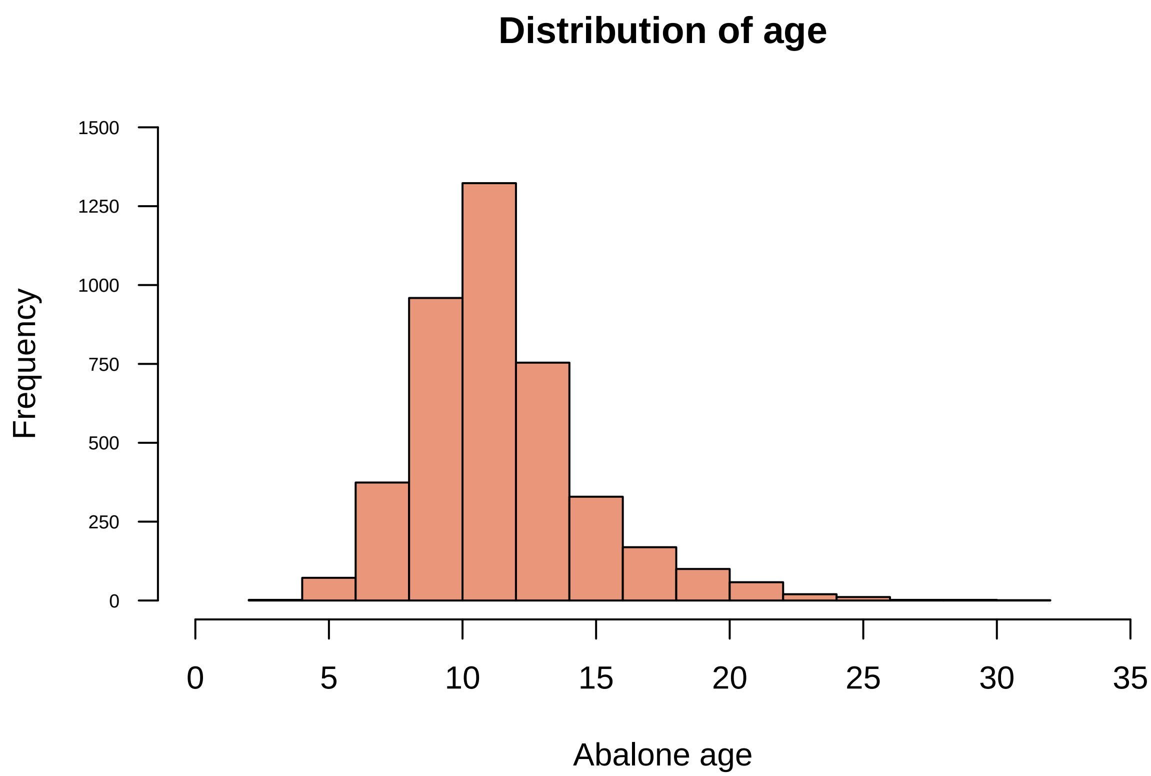
Sunday, October 16th 2022

1.

```
Abalone_ <- read.csv("abalone.data")
Abalone_$age <- Abalone_$X15 + 1.5

Abalone <- Abalone_
Abalone <- transform(Abalone_, age= X15 + 1.5)
```

```
distr <- Abalone$age
hist(distr, main= "Distribution of age", xlab= "Abalone age",
     ylim= c(0,1500), xlim= c(0,35), col= "darksalmon", yaxt= "n")
axis(side= 2, at= seq(0, 1500, by=250), cex.axis= 0.6, las= 1)
```



Most of the abalone in the data set has an age in the range 10-12 years, the youngest abalone is 2.5 years, and the oldest abalone is 30.5 years.

2.

```r
set.seed(2022)

Abalone_split <- initial_split(Abalone, prop= 0.80, strata= age)

Abalone_train <- training(Abalone_split)

Abalone_test <- testing(Abalone_split)
```

3.

```r
Abalone_recipe <- recipe(age ~ M + X0.455 + X0.365 + X0.095 + X0.514 + X0.2245
                         + X0.101 + X0.15, data= Abalone_train)

summary(Abalone_recipe)
```

```
## # A tibble: 9 x 4
##   variable type    role      source
##   <chr>    <chr>   <chr>     <chr>
## 1 M        nominal predictor original
## 2 X0.455   numeric predictor original
## 3 X0.365   numeric predictor original
## 4 X0.095   numeric predictor original
## 5 X0.514   numeric predictor original
## 6 X0.2245  numeric predictor original
## 7 X0.101   numeric predictor original
## 8 X0.15    numeric predictor original
## 9 age      numeric outcome   original
```

```r
Abalone_recipe_steps <- Abalone_recipe %>%
  step_impute_mean(all_numeric()) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors()) %>%
  step_nzv(all_predictors())

Abalone_recipe_steps
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Operations:
##
## Mean imputation for all_numeric()
## Dummy variables from all_nominal_predictors()
## Centering for all_predictors()
## Scaling for all_predictors()
## Sparse, unbalanced variable filter on all_predictors()
```

```r
Abalone_recipe_prep <- prep(Abalone_recipe_steps, training = Abalone_train)
Abalone_recipe_prep
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          8
##
## Training data contained 3339 data points and no missing data.
##
## Operations:
##
## Mean imputation for X0.455, X0.365, X0.095, X0.514, X0.2245, X0.101... [trained]
## Dummy variables from M [trained]
## Centering for X0.455, X0.365, X0.095, X0.514, X0.2245, X0.101... [trained]
## Scaling for X0.455, X0.365, X0.095, X0.514, X0.2245, X0.101... [trained]
## Sparse, unbalanced variable filter removed <none> [trained]
```

```
Abalone_recipe_final <- bake(Abalone_recipe_prep, Abalone_train)
Abalone_recipe_final
```

```
## # A tibble: 3,339 x 10
##    X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101  X0.15   age    M_I    M_M
##     <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>
##  1 -1.47  -1.46  -1.19  -1.24   -1.18  -1.22  -1.22    8.5 -0.683  1.30
##  2 -1.64  -1.56  -1.43  -1.28   -1.23  -1.30  -1.33    8.5  1.46  -0.767
##  3 -0.840 -1.10  -1.07  -0.984  -0.996 -0.952 -0.861   9.5  1.46  -0.767
##  4 -1.34  -1.15  -1.43  -1.18   -1.20  -1.27  -1.00    8.5 -0.683  1.30
##  5 -0.504 -0.545 -0.832 -0.721  -0.606 -0.526 -0.824   9.5 -0.683  1.30
##  6 -0.630 -0.545 -0.832 -0.633  -0.561 -0.594 -0.680   9.5 -0.683 -0.767
##  7 -2.39  -2.37  -2.27  -1.56   -1.49  -1.45  -1.58    6.5  1.46  -0.767
##  8 -2.69  -2.63  -2.03  -1.62   -1.52  -1.53  -1.64    6.5  1.46  -0.767
##  9 -2.64  -2.63  -2.15  -1.62   -1.56  -1.55  -1.62    5.5  1.46  -0.767
## 10 -1.68  -1.66  -1.67  -1.38   -1.29  -1.43  -1.40    7.5  1.46  -0.767
## # ... with 3,329 more rows
```

```
Abalone_recipe_test <- bake(Abalone_recipe_prep, Abalone_test)
Abalone_recipe_test
```

```
## # A tibble: 837 x 10
##    X0.455  X0.365   X0.095 X0.514 X0.2245 X0.101  X0.15   age    M_I    M_M
##     <dbl>   <dbl>    <dbl>  <dbl>   <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>
##  1 -0.798 -0.595  -0.712   -0.872 -0.876   -0.920 -0.752  11.5 -0.683  1.30
##  2 -1.43  -1.31   -1.31    -1.11  -1.20    -1.30  -0.897   8.5  1.46  -0.767
##  3 -0.714 -0.697  -0.952   -0.780 -0.783   -0.865 -0.788  11.5 -0.683 -0.767
##  4  0.545  0.369   0.00773  0.206 -0.0217   0.484  0.294  13.5 -0.683  1.30
##  5  0.126  0.674   0.368    0.794  0.769    1.16   0.727  17.5 -0.683 -0.767
##  6 -1.43  -1.20   -1.19    -1.03  -1.03    -0.874 -1.08   10.5 -0.683  1.30
##  7 -1.13  -1.15   -1.07    -1.29  -1.24    -1.25  -1.19    8.5  1.46  -0.767
##  8 -0.546 -0.342  -0.472   -0.760 -0.831   -0.654 -0.644   8.5 -0.683 -0.767
##  9 -1.05  -0.900  -1.07    -1.08  -1.03    -1.11  -1.00    8.5 -0.683  1.30
## 10 -0.168 -0.0369 -0.712   -0.422 -0.253   -0.195 -0.464  10.5 -0.683  1.30
## # ... with 827 more rows
```

```
Interaction1 <- lm(age ~ M + X0.2245, data= Abalone)
Interaction1
```

```
##
## Call:
## lm(formula = age ~ M + X0.2245, data = Abalone)
##
## Coefficients:
## (Intercept)           MI           MM      X0.2245
##     10.9146      -2.2583      -0.3763       3.8430
```

```
Interaction2 <- lm(age ~ X0.455 + X0.365, data= Abalone)
Interaction2
```

```
##
## Call:
## lm(formula = age ~ X0.455 + X0.365, data = Abalone)
##
## Coefficients:
## (Intercept)       X0.455       X0.365
##       4.204      -10.552       31.277
```

```
Interaction3 <- lm(age ~ X0.2245 + X0.15, data= Abalone)
Interaction3
```

```
##
## Call:
## lm(formula = age ~ X0.2245 + X0.15, data = Abalone)
##
## Coefficients:
## (Intercept)      X0.2245        X0.15
##       8.162       -8.742       26.846
```

The variable rings is proportional to the age of the abalone, however taking it into consideration to predict the age of the abalone can lead to overfitting. Rings is not exclusive and other variables can alter the final prediction.

4.

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

```
## Linear Regression Model Specification (regression)
##
## Computational engine: lm
```

5.

```
lm_Abaflow<- workflow() %>%
  add_model(lm_Abalone) %>%
  add_recipe(Abalone_recipe)
```

6.

```
lm_Abafit <- fit(lm_Abaflow, Abalone_train)

FitModel <- lm(age ~ M + X0.455+ X0.365+ X0.095+ X0.514+ X0.2245 + X0.15,
               data= Abalone)
summary(FitModel)
```

```
##
```

```
## Call:
## lm(formula = age ~ M + X0.455 + X0.365 + X0.095 + X0.514 + X0.2245 +
##     X0.15, data = Abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8443 -1.3261 -0.3342  0.9016 14.8567
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.59791    0.29252  19.137  < 2e-16 ***
## MI           -0.75941    0.10284  -7.384 1.84e-13 ***
## MM            0.08160    0.08388   0.973    0.331
## X0.455       -1.80180    1.81400  -0.993    0.321
## X0.365       11.77939    2.24129   5.256 1.55e-07 ***
## X0.095       10.29366    1.54598   6.658 3.13e-11 ***
## X0.514        5.33402    0.57660   9.251  < 2e-16 ***
## X0.2245     -18.03710    0.79450 -22.702  < 2e-16 ***
## X0.15        11.77955    1.06933  11.016  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.21 on 4167 degrees of freedom
## Multiple R-squared:  0.531,  Adjusted R-squared:  0.5301
## F-statistic: 589.7 on 8 and 4167 DF,  p-value: < 2.2e-16
```

```
Hypothetical <- data.frame(M= c('F'), X0.455= c(0.50), X0.365= c(0.10),
                           X0.095= c(0.30), X0.514= c(4), X0.2245= c(1),
                           X0.101= c(2), X0.15= c(1))

predict(FitModel, newdata= Hypothetical)
```

```
##        1
## 24.04157
```

The hypothetical abalone is approximately 24 years of age.

```
lm_Abafit %>%
  extract_fit_parsnip() %>%
  tidy()
```

```
## # A tibble: 10 x 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)   5.49      0.330    16.6    1.12e-59
##  2 MI           -0.791     0.115    -6.89   6.51e-12
##  3 MM            0.0970    0.0931    1.04   2.98e- 1
##  4 X0.455       -0.149     2.02     -0.0738 9.41e- 1
##  5 X0.365       10.3       2.47      4.18   2.93e- 5
##  6 X0.095       10.2       1.68      6.08   1.38e- 9
##  7 X0.514        8.71      0.810    10.8    1.42e-26
##  8 X0.2245     -19.3       0.904   -21.3    1.26e-94
##  9 X0.101      -10.1       1.44     -7.02   2.60e-12
## 10 X0.15         8.97      1.25      7.18   8.84e-13
```

```r
PredAbalone <- predict(lm_Abafit, new_data= Abalone_train %>% select(-age) )
PredAbalone %>%
  head()
```

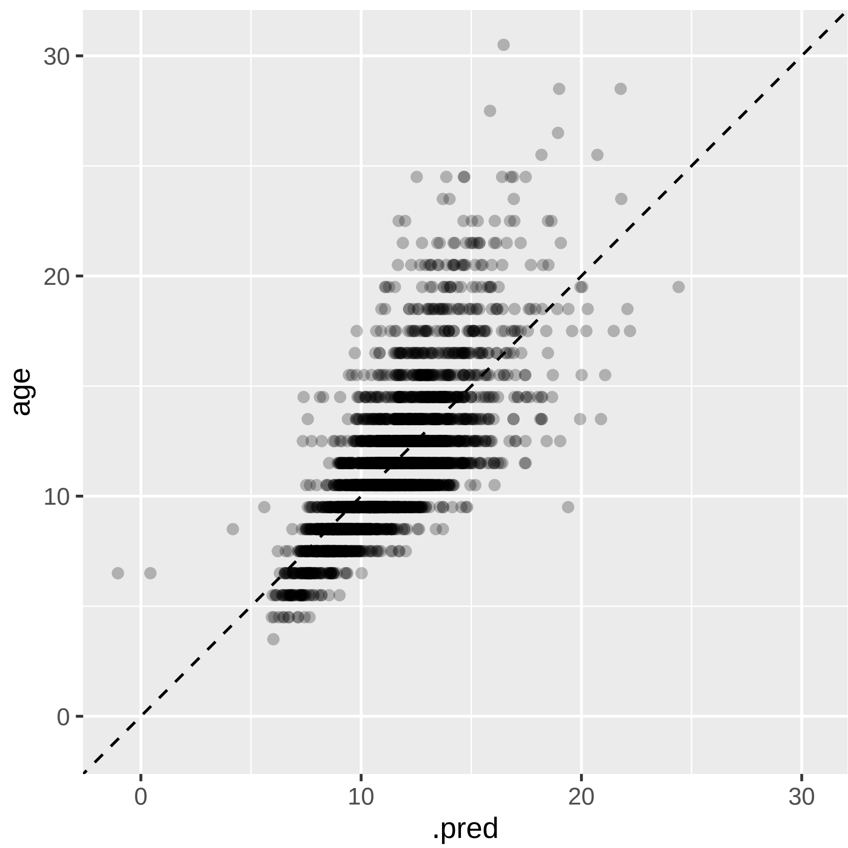
```
## # A tibble: 6 x 1
##   .pred
##   <dbl>
## 1  9.38
## 2  8.26
## 3  9.34
## 4 10.2
## 5  9.92
## 6 10.3
```

7.

```r
PredAbalone <- bind_cols(PredAbalone, Abalone_train %>% select(age))
PredAbalone %>%
  head()
```

```
## # A tibble: 6 x 2
##   .pred   age
##   <dbl> <dbl>
## 1  9.38   8.5
## 2  8.26   8.5
## 3  9.34   9.5
## 4 10.2    8.5
## 5  9.92   9.5
## 6 10.3    9.5
```

```r
PredAbalone %>%
  ggplot(aes(x= .pred, y= age)) +
  geom_point(alpha= 0.25) +
  geom_abline(lty= 2) +
  coord_obs_pred()
```




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

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.531
## 2 rmse    standard       2.19
## 3 mae     standard       1.58
```

The R^2 value demonstrates that approximately 53.11% of the variance of dependent variables is explained by the variance of the independent variable.