WU #11 - Cross Validation

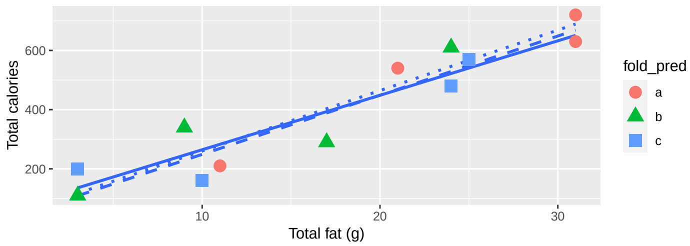
Math 158 - Jo Hardin

Thursday 3/3/2022

Name: ____

Names of people you worked with: ____

Consider the following dataset measuring fat content and calories for 12 fast food items.[1] Cross validated models have been fit for $v = 3$ folds.



The values of the observations in group `a` are as follows:

```
## # A tibble: 4 x 3
##   calories total_fat fold_pred
##      <dbl>     <dbl> <chr>
## 1      630        31 a
## 2      210        11 a
## 3      720        31 a
## 4      540        21 a
```

Calculate R^2 and RMSE for the observations in fold `a`. (That is, calculate exactly two numbers.)

a and b points

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     56.1     61.9      0.906 0.400
## 2 total_fat       20.4      2.99     6.84  0.000480
```

a and c points

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     49.1     57.7      0.851 0.427
## 2 total_fat       20.0      2.65     7.54  0.000282
```

b and c points

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     80.6     59.3       1.36 0.223
## 2 total_fat       18.4      3.52      5.22 0.00197
```

[1] the data actually come from a much larger and real dataset

Solution:

```
a_pts <- ff %>%
  filter(fold_pred == "a")

bc_mod <- ff %>% filter(fold_pred != "a") %>% lm(calories ~ total_fat, data = .)

bc_mod %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     80.6     59.3       1.36 0.223
## 2 total_fat       18.4      3.52      5.22 0.00197
```

```
bc_mod %>%
  predict(a_pts)
```

```
##        1        2        3        4
## 650.8049 282.9193 650.8049 466.8621
```

```
bc_mod %>%
  augment(newdata = a_pts)
```

```
## # A tibble: 4 x 5
##   calories total_fat fold_pred .fitted .resid
##      <dbl>     <dbl> <chr>       <dbl>  <dbl>
## 1      630        31 a            651.  -20.8
## 2      210        11 a            283.  -72.9
## 3      720        31 a            651.   69.2
## 4      540        21 a            467.   73.1
```

```
bc_mod %>%
  augment(newdata = a_pts) %>%
  summarize(R2 = 1 - sum(.resid^2) / sum((calories - mean(calories))^2),
            RMSE = sqrt(sum(.resid^2)/4))
```

```
## # A tibble: 1 x 2
##      R2  RMSE
##   <dbl> <dbl>
## 1 0.893  63.0
```