## PM592: Regression Analysis for Health Data Science
## Lab 5 – Regression and ANOVA
**Data Needed:** *CHS, Cereals*

**Outline**

- ANOVA – function and test
- Sums of Squares Types
- Extra Sums of Squares

### 1. The ANOVA Function

**1.1.** The ANOVA table breaks down the total variation in Y into variation that's explained by 1) our regression model and 2) error.

```
> anova(fev.m1)
Analysis of Variance Table

Response: fev
            Df    Sum Sq  Mean Sq F value    Pr(>F)
bmi          1  15488176 15488176  162.35 < 2.2e-16 ***
Residuals 1103 105224936    95399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**1.2.** Recall from Week 4 we created our own anova.full() function; this simply takes the output of the ANOVA table from above and adds a "total" column so we can see the total sums of squares that exist.

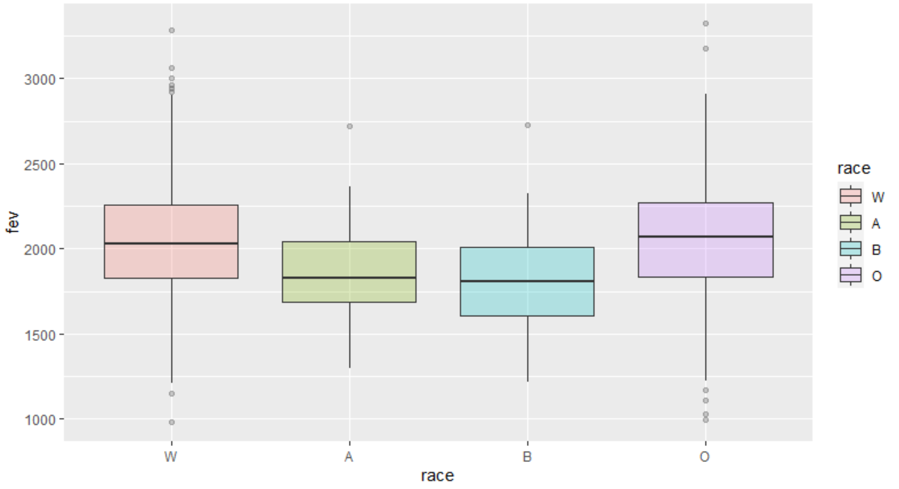
```
> anova.full(fev.m1)
# A tibble: 3 x 6
  rowname      Df     Sum.Sq   Mean.Sq F.value    Pr..F.
  <chr>     <int>      <dbl>     <dbl>   <dbl>     <dbl>
1 bmi           1  15488176. 15488176.    162.  8.60e-35
2 Residuals  1103 105224936.    95399.      NA  NA
3 Total      1104 120713111.   109342.      NA  NA
```

**1.3.** Running a lm() on a categorical independent variable is equivalent to performing an ANOVA. In the lm() framework, the ANOVA table tells us that categorical race is a significant predictor of FEV. On further investigation of the slope parameter estimates, Asian and Black participants had a lower average FEV compared to white participants.

```
> fev_race.m <-
+   lm(fev ~ factor(race), data = chs)
> anova(fev_race.m)
Analysis of Variance Table

Response: fev
               Df    Sum Sq Mean Sq F value    Pr(>F)
factor(race)    3   4151716 1383905  13.072 2.161e-08 ***
Residuals    1101 116561395  105869
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> chs %>%
+   ggplot(aes(x = race, y = fev, fill = race)) +
+   geom_boxplot(alpha = .25)
```



```
> summary(fev_race.m)

Call:
lm(formula = fev ~ factor(race), data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1061.37  -217.26    -6.61   208.25  1267.24 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2046.22      13.18 155.195  < 2e-16 ***
factor(race)A  -180.30      47.87  -3.767 0.000174 ***
factor(race)B  -236.12      48.32  -4.887 1.18e-06 ***
factor(race)O    10.23      20.99   0.487 0.626040    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 1101 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.03439, Adjusted R-squared:  0.03176 
F-statistic: 13.07 on 3 and 1101 DF,  p-value: 2.161e-08
```

**1.4.** We could have instead performed a traditional ANOVA and would have arrived at the same conclusion.

```
> aov(fev ~ race, data = chs) %>% summary()
              Df    Sum Sq Mean Sq F value   Pr(>F)    
race           3   4151716 1383905   13.07 2.16e-08 ***
Residuals   1101 116561395  105869                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   95 bservations deleted due to missingness
```

**1.5.** The aov() function gives us the ability to correct for multiple comparisons.

**1.5.1.** Recall that when we make multiple comparisons among groups we must control the Type I error rate (See OpenIntro 7.5.6).

**1.5.2.** When we perform several t-tests in succession, each with probability $\alpha$ of committing Type I error, the probability of finding a significant result due to chance increases.

**1.5.3.** One of the most common ways to control the Type I error rate when making multiple comparisons is the Tukey HSD (honestly significant difference) test.

**1.5.4.** Here, we see significant differences in FEV between Asian-white, Black-white, other-Asian, and other-Black groups.

```
> aov(fev ~ race, data = chs) %>% TukeyHSD()
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = fev ~ race, data = chs)

$race
          diff        lwr        upr     p adj
A-W -180.29847 -303.45733  -57.13962 0.0009993
B-W -236.11910 -360.43391 -111.80430 0.0000070
O-W   10.23067  -43.77155   64.23290 0.9618976
B-A  -55.82063 -224.10785  112.46659 0.8287165
O-A  190.52915   64.90010  316.15820 0.0005848
O-B  246.34978  119.58730  373.11226 0.0000040
```

**1.5.5.** The ability to perform multiple comparisons is a benefit of using the aov() function; performing the test in a linear regression will only provide comparisons to the baseline/reference group.

## 2. Sums of Squares Types

**2.1.** The types of sums of squares we compute from the ANOVA table are divided into three types

### 2.2. Type I SS

**2.2.1.** This is the sums of squares of each variable added *sequentially*

**2.2.2.** The sums of squares for each term in the model takes into account all variables that had been entered into the model before it.

**2.2.3.** Here we see that wheeze is not significant in the model. Adjusting for wheeze, male is significant.

```
> fev.m4 <-
+   lm(fev ~ wheeze + male, data = chs)
> anova(fev.m4)
Analysis of Variance Table

Response: fev
            Df    Sum Sq Mean Sq F value    Pr(>F)
wheeze       1    268054  268054  2.6105    0.1065
male         1   6319859 6319859 61.5470 1.069e-14 ***
Residuals 1039 106688063  102683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**2.2.4.** With regard to the p-values presented, the order of entry matters with Type I SS. In this next analysis we see that male is significant in the model. Adjusting for male, wheeze is significant as well.

```
> fev.m5 <-
+   lm(fev ~ male + wheeze, data = chs)
> anova(fev.m5)
Analysis of Variance Table

Response: fev
            Df    Sum Sq Mean Sq F value    Pr(>F)
male         1   5970735 5970735 58.1470 5.476e-14 ***
wheeze       1    617178  617178  6.0105   0.01438 *
Residuals 1039 106688063  102683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

### 2.3. Type II SS

**2.3.1.** This is the sums of squares for the main effects in the model.

**2.3.2.** So far we have only examined main effects, so we will not discuss this type of SS further

### 2.4. Type III SS

**2.4.1.** This is the sums of squares for each variable considering (adjusting for) all other terms in the model.

**2.4.2.** The order of variable entry does not matter for Type III SS.

**2.4.3.** For any given variable, the Type III SS for that variable is equivalent to the Type I SS had that variable been the last one entered in the model.

**2.4.4.** Type II and III SS can be obtained using the Anova() function in the CAR package.

**2.4.5.** Here we see that, adjusting for wheeze, BMI is significant in the model. Furthermore, adjusting for BMI, wheeze is significant.

```
> Anova(fev.m4, type = 3)
Anova Table (Type III tests)

Response: fev
                Sum Sq   Df    F value    Pr(>F)
(Intercept) 1710233309    1 16655.4004 < 2.2e-16 ***
wheeze          617178    1     6.0105   0.01438 *
male           6319859    1    61.5470 1.069e-14 ***
Residuals    106688063 1039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## 3. Extra Sums of Squares

**3.1.** Suppose we want to compare two models to see which fits better

**3.1.1.** Example 1: Does BMI do a good job at predicting FEV?

$$\text{Model 1: } \hat{Y}_{FEV} = \beta_0 + \beta_{BMI}X_{BMI}$$
$$\text{Model 0: } \hat{Y}_{FEV} = \beta_0$$

How would we figure this out? There are two pieces of information in the output that tell us about the effect of BMI compared to the null (intercept-only) model. When we compare a model with only one independent variable to the null model, the test of $H_0: \beta_1 = 0$ is equivalent to the null hypothesis for the F-test ($H_0$: the fit of the current model is equal to the fit of the intercept-only model).

Conclusion: A model with BMI fits significantly better than the intercept-only model (p<.001). The slope for BMI is significantly nonzero (p<.001).

```
> fev.m1 %>% summary()

Call:
lm(formula = fev ~ bmi, data = chs)

Residuals:
   Min     1Q Median     3Q    Max 
-991.5 -207.9   -5.4  200.9 1304.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1450.760     46.497   31.20   <2e-16 ***
bmi           31.363      2.461   12.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308.9 on 1103 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.1283,	Adjusted R-squared:  0.1275 
F-statistic: 162.4 on 1 and 1103 DF,  p-value: < 2.2e-16
```

**3.1.2.** Example 2: Does “having pets” do a good job at predicting FEV when added to a model with BMI?

$$\text{Model 1: } \hat{Y}_{FEV} = \beta_0 + \beta_{BMI}X_{BMI} + \beta_{PETS}X_{PETS}$$
$$\text{Model 0: } \hat{Y}_{FEV} = \beta_0 + \beta_{BMI}X_{BMI}$$

Now the F-test in the lm() output shows us the significance of the *entire model* vs. the intercept-only model. To see if having pets affects FEV in addition to BMI, we can look at the p-value of the slope coefficient for pets.

Conclusion: After adjusting for BMI, having pets is positively associated with FEV (p=.049).

```
> fev.m2 <-
+   lm(fev ~ bmi + pets, data = chs)
> fev.m2 %>% summary()

Call:
lm(formula = fev ~ bmi + pets, data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1002.91  -211.28    -7.13   196.77  1295.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1413.41      50.15  28.186   <2e-16 ***
bmi            31.55       2.46  12.824   <2e-16 ***
```

```
pets          43.86      22.23   1.973   0.0487 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308.5 on 1102 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.1314,  Adjusted R-squared:  0.1298
F-statistic: 83.34 on 2 and 1102 DF,  p-value: < 2.2e-16
```

**3.1.3.** Example 3: Does Race do a good job at predicting FEV when added to a model with BMI?

Model 1: $\hat{Y}_{FEV} = \beta_0 + \beta_{BMI}X_{BMI} + \beta_{Asian}X_{Asian} + \beta_{Black}X_{Black} + \beta_{Other}X_{Other}$

Model 0: $\hat{Y}_{FEV} = \beta_0 + \beta_{BMI}X_{BMI}$

This is a bit more complicated as there is no single slope coefficient for us to look at. We must find a way to test the addition of *all* the dummy variables associated with race simultaneously.

Conclusion: None, yet. We know Asian and Black differ from White, but we do not have any information about their predictive ability as a set.

```
> fev.m3 %>% summary()

Call:
lm(formula = fev ~ bmi + race, data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1022.47  -205.90    -7.81   197.24  1278.09 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1470.455     45.828  32.087  < 2e-16 ***
bmi           31.782      2.437  13.040  < 2e-16 ***
raceA       -168.763     44.576  -3.786 0.000161 ***
raceB       -266.691     45.047  -5.920 4.29e-09 ***
raceO        -22.225     19.700  -1.128 0.259476    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 302.9 on 1100 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.1637,  Adjusted R-squared:  0.1606
F-statistic: 53.82 on 4 and 1100 DF,  p-value: < 2.2e-16
```

**3.2.** We can use an F-test based on the principle of **extra sums of squares** to test the effect of adding several variables to the model on the fit of the model.

**3.3.** Note that this approach can only be used on two models that are **nested**.

**3.4.** Notation for the extra sums of squares F-test (assume model 0 is nested within model 1)

**3.4.1.** $SSE_1$ = Error sum of squares from model 1

**3.4.2.** $SSE_0$ = Error sum of squares from model 0

**3.4.3.** $DFE_1$ = Error degrees of freedom from model 1

**3.4.4.** $DFE_0$ = Error degrees of freedom from model 0

**3.5.** The F-statistic is given as:

$$\boldsymbol{F} = \frac{\frac{SSE_0 - SSE_1}{DFE_0 - DFE_1}}{\frac{SSE_1}{DFE_1}}$$

And has an F-distribution with ($DFE_0$ – $DFE_1$) and $DFE_1$ degrees of freedom under the null hypothesis.

**3.6.** Warning: Models 1 and 0 must be based on exactly the same sample size, which can be a problem if some observations are omitted because of missing values on the additional variables.

**3.7.** Example 3 by Hand

**3.7.1.** The full model has SSE = 100954938 on 1100 df.

```
> anova(fev.m3)
Analysis of Variance Table

Response: fev
            Df    Sum Sq  Mean Sq F value    Pr(>F)    
bmi          1  15488176 15488176 168.758 < 2.2e-16 ***
race         3   4269997  1423332  15.509 6.935e-10 ***
Residuals 1100 100954938    91777                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**3.7.2.** The reduced model has SSE = 105224936 on 1103 df.

```
> anova(fev.m1)
Analysis of Variance Table

Response: fev
            Df    Sum Sq  Mean Sq F value    Pr(>F)    
bmi          1  15488176 15488176  162.35 < 2.2e-16 ***
Residuals 1103 105224936    95399                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**3.7.3.** $F = \frac{\frac{105224936-100954938}{1103-1100}}{\frac{100954938}{1100}} = 15.51$ on 3, 1100 df

```
> 1-pf(15.51, 3, 1100)
[1] 6.920453e-10
```

Conclusion: Adding the set of race dummy variables significantly improves the fit when added to a model with BMI (p<.001).

**3.8.** Example 3 by Code

**3.8.1.** We can use the anova() function to perform the extra sums of squares F-test to compare the fit of two nested models.

**3.8.2.** Our conclusion is the same as if we performed the test by hand in (1.7).

```
> anova(fev.m1, fev.m3)
Analysis of Variance Table
```

```
Model 1: fev ~ bmi
Model 2: fev ~ bmi + race
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)    
1   1103 105224936                                  
2   1100 100954938  3   4269997 15.509 6.935e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Lab 5 Exercises

| | |
|---|---|
| Objective(s): | Analyze univariable and multivariable regressions, use the ANOVA table to determine statistical significance, interpret regression coefficients, evaluate influential values |
| Datasets Required: | `cereals` |

The “rating” variable in the cereals data set contains a rating based on Consumer Reports. In this lab we will test the hypothesis that there is a difference in consumer rating based on 1) manufacturer, 2) calories, and 3) sugar. You will need the GGally and olsrr packages installed.

1) While *calories* and *sugars* are numeric variables, *mfr* is nominal. Create a factor variable for *mfr*.

   a) What is the default reference category for mfr?

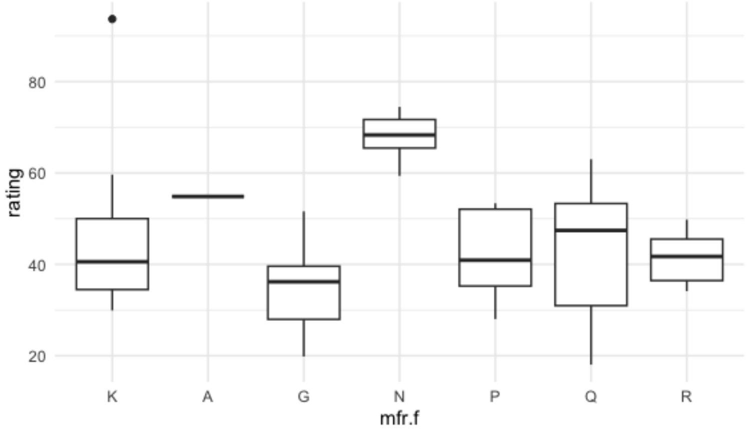
      “A”

   b) Use the count() function to determine the number of cereals per manufacturer. Based on this information, change the reference category to a more appropriate category. Which category did you choose and why?

      ```
      > cereals$mfr.f <- relevel(cereals$mfr.f, ref = "K")
      ```

      “K” is now the reference category because it has the highest number of cereals

   c) Create a boxplot of *rating* by *mfr*. Based on this boxplot, do you think an ANOVA would be significant?



      Yes, I believe that an ANOVA would be significant, as the ratings differ a lot based on manufacturer.

2) Perform 3 separate univariable analyses for the effect of *calories, sugars,* and *mfr* on *rating*.

   a) Is the calorie content of cereal related to rating? Explain the relationship and provide evidence (parameter estimate and p-value) to support your answer.

A one-unit increase in calories is expected to decrease the rating by 0.497 units. The p-value of 4.14e-12 indicates that this relationship is statistically significant. The R^2 value of 0.475 indicates that calories explains 47.5% of the variance in ratings.

b) Is the sugar content of cereal related to rating? Explain the relationship and provide evidence (parameter estimate and p-value) to support your answer.

A one-unit increase in sugars is expected to decrease the rating by 2.40 units. The p-value of 1.15e-15 indicates that the relationship between sugars and rating is statistically significant. The R^2 value of 0.577 indicates that calories explains 57.7% of the variance in ratings.

c) Is the manufacturer of cereal related to rating? Explain the relationship and provide evidence (parameter estimate and p-value) to support your answer.

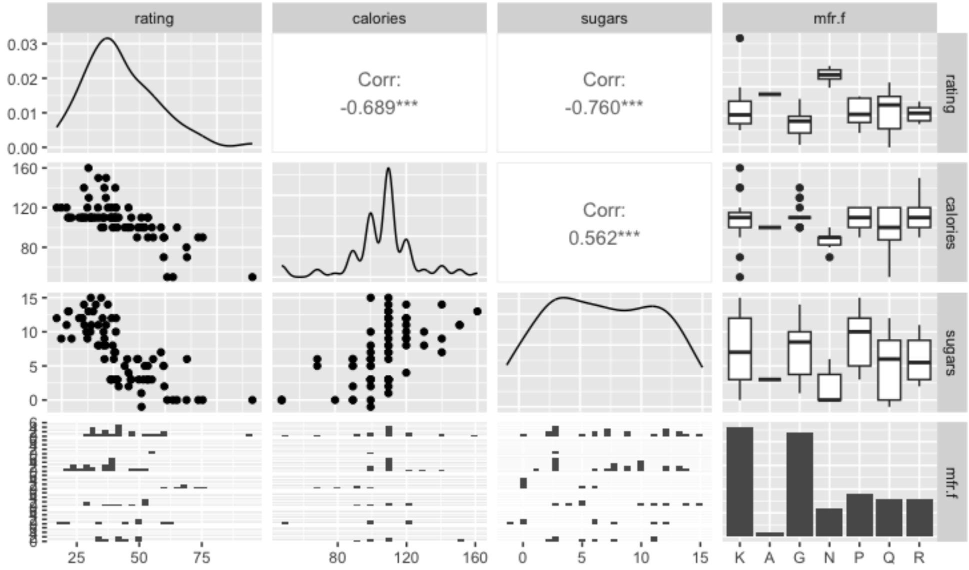
Since manufacturer is a nomial variable, the F-test tests whether manufacturer is related to rating. The F-statistic is 6.804 on 6 and 70 degrees of freedom, and the p-value is 1.032e-05 (p << 0.001), which indicates that the relationship between manufacturer and cereal rating is statistically significant. The R^2 value of 0.368 indicates that manufacturer explains about 36.8% of the variance in ratings.

3) When we use factor variables in the lm() function, dummy variables are created for each level of the factor. Explain the meaning of each of the dummy variables that were created in (2c).

The dummy variables created for each level of the factor indicates which manufacturer each individual observation belongs to. The reference, “K”, is the group for which all other manufacturers are compared to. For example, the coefficient for “A” is the difference in mean rating between manufacturer “A” vs “K”, and coefficient for “G” is the difference in mean rating between manufacturer “G” vs “K”, etc. If the dummy variables were in the “wide” format there would be n – 1 number of features with n being the number of categories. And each observation would have a `1` for the category it belongs to and a `0` for all others. The reference group is identified by having a `0` in all categories.

4) Use the ggpairs() function to provide a figure of all possible correlations among variables (Note: select only the variables included in the model!)

a) Provide the figure.



b) Based on this figure, do you have any concerns about linearity?

From the figure, it looks like rating vs sugars and calories vs sugars may follow a nonlinear relationship.

c) Based on this figure, do you have any concerns about collinearity?

Not particularly, sugars and calories have a correlation coefficient of 0.562, which isn’t concerningly high.

5) Fit a multiple regression model that includes *calories, sugars,* and *mfr.*

a) Explain if there is an effect of calories: provide a p-value for the effect and interpret the beta coefficient associated with calories.

Holding manufacturer and sugars constant, a one-unit increase in calories is predicted to decrease rating by 0.24956. The p-value 2.61e-06 indicates that this association is significant.

b) Explain if there is an effect of sugars: provide a p-value for the effect and interpret the beta coefficient associated with calories.

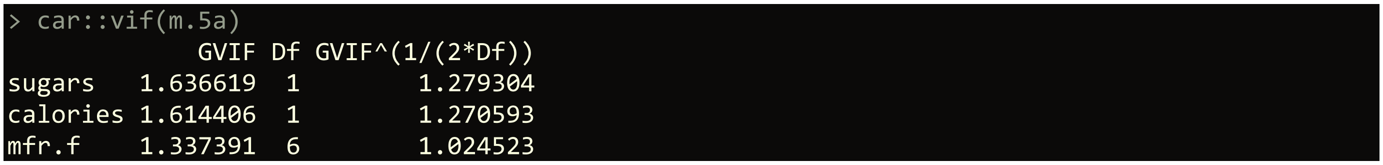
Holding manufacturer and calories constant, a one-unit increase in calories is predicted to decrease rating by 1.54203. The p-value 6.68e-10 indicates that this association is significant.

c) Explain if there is an overall effect of manufacturer. This will require computing the extra sums of squares F-test, comparing a model with “calories + sugars” to a model with “calories + sugars + mfr”.

The extra sums of squares F-test resulted in a p-value of 2.42e-06, which indicates that adding the set of manufacturer dummy variables significantly improves the fit when added to a model with calories and sugar (p<.001).

6) Determine if there is evidence of collinearity.

a) Use the vif() function in the car package to find the variance inflation factor of each variable.

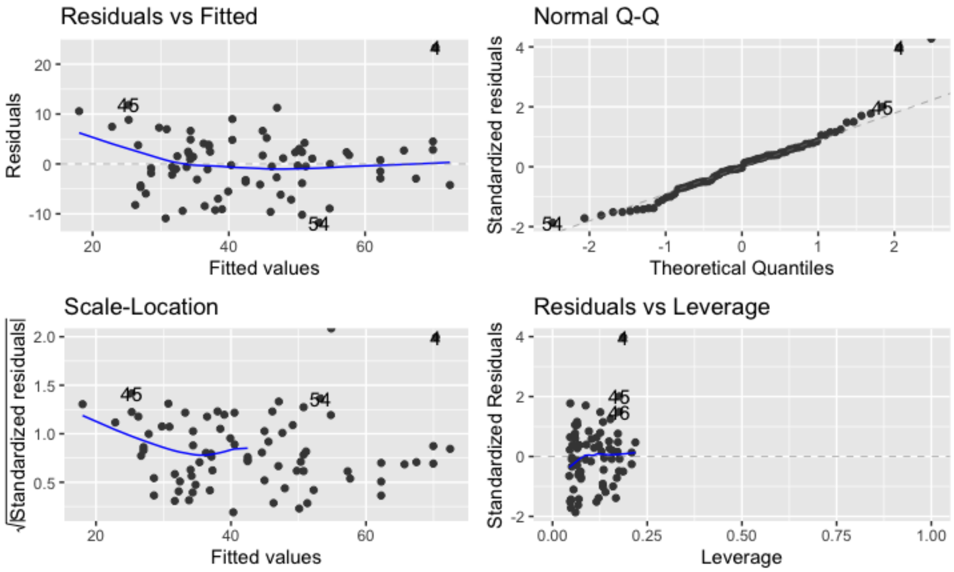


```
> car::vif(m.5a)
             GVIF Df GVIF^(1/(2*Df))
sugars   1.636619  1        1.279304
calories 1.614406  1        1.270593
mfr.f    1.337391  6        1.024523
```

b) Based on the VIF, do you suspect there is collinearity?

Based on the VIF, I do not suspect presence of collinearity, none of the VIFs are above 10.

7) Determine if model assumptions are met by using the autoplot() function.



a) Does the Residuals vs. Fitted plot show evidence of a violation of linearity?

No, the median line of points seems to be overall consistent at 0 for the fitted values.

b) Does the Normal Q-Q plot show evidence of a violation of normality?

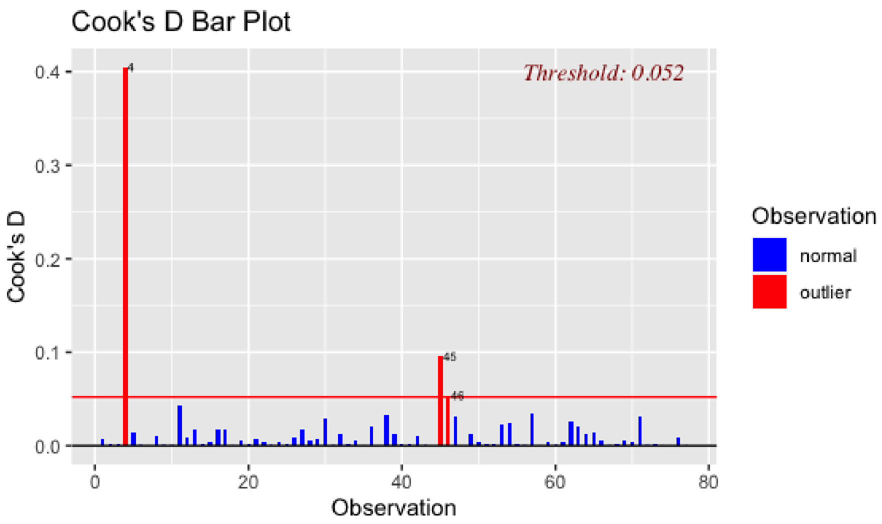
Normality assumption seems okay, the Q-Q plot of the residuals follows a straight line for the most part.

c) Does the Scale-Location plot show evidence of a violation of homoscedasticity?

No, the scale-location plot generally looks okay so no flagrant violation of homoscedasticity.

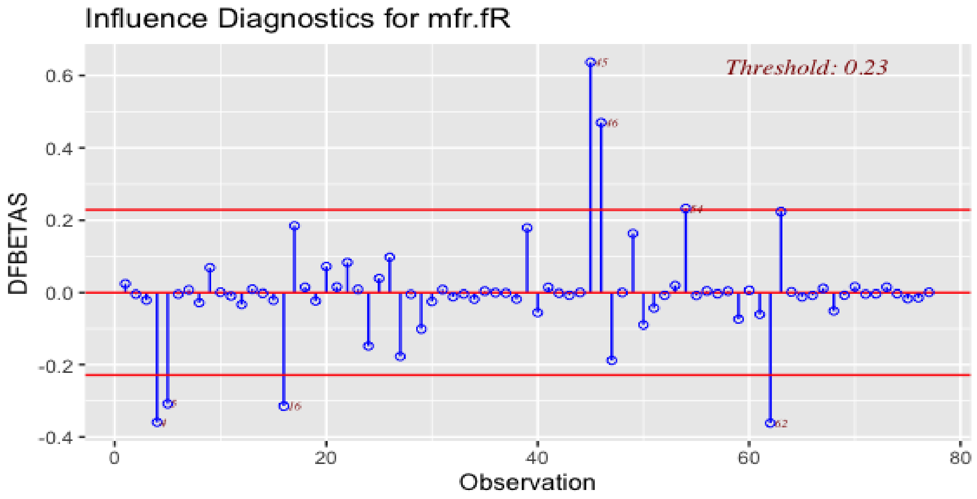
8) Determine if there are any influential values.

**a)** Use the **ols_plot_cooksd_bar()** function to **determine if any observations influence the fitted values.**



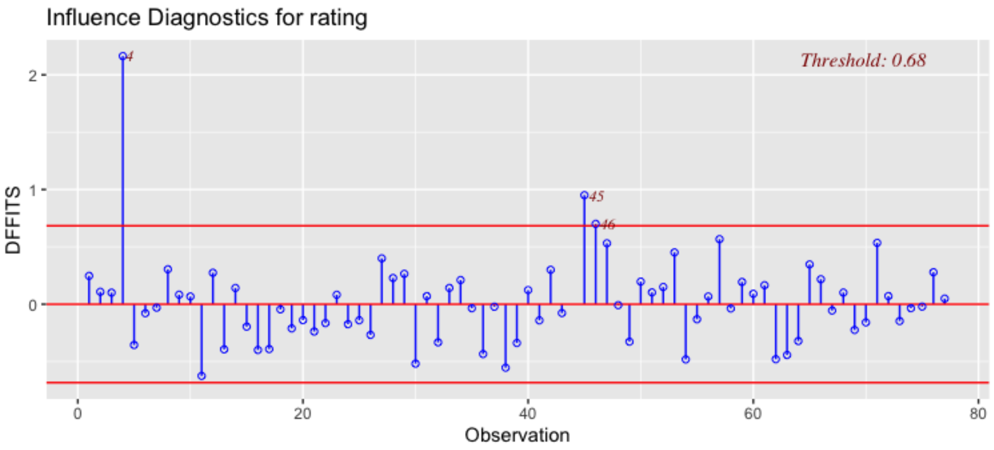
It appears that observation 4 may influence the fitted values, although since its value is not above 1 it is not extremely influential in an absolute sense.

b) Use the **ols_plot_dfbetas()** function with the option "print_plot = F" to **observe which observations influence the value of the intercept and each slope parameter**. Which observation is influencing the value of the intercept the most?



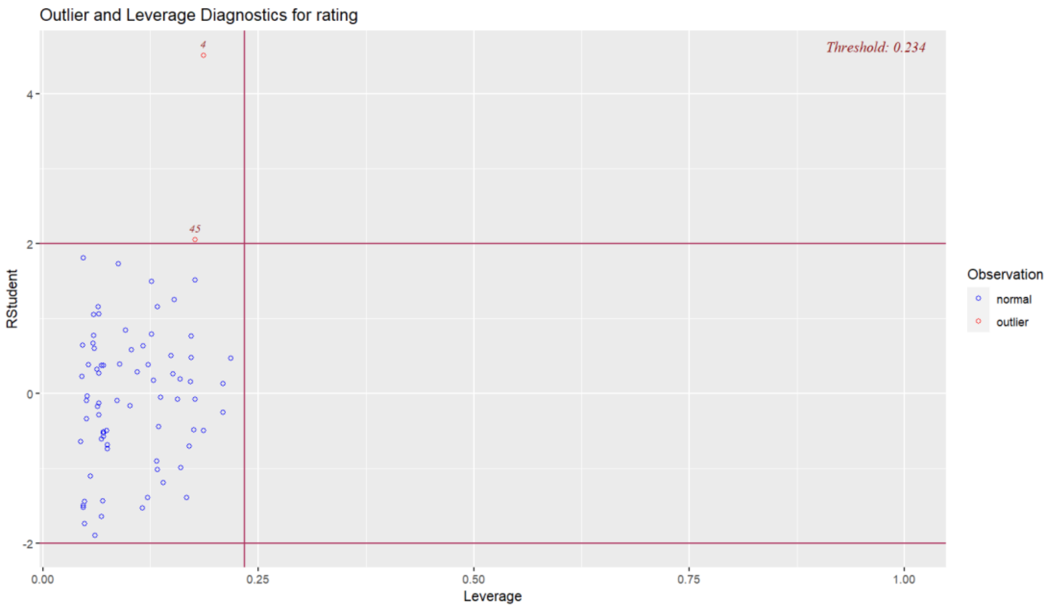
From the DFBETAS plot it looks like observation 45 has the most influence on the model parameters (intercept and slopes).

c) Use the **ols_plot_dffits()** function to determine if **any observations influence the fit**.



Observation 4 is seems to influence the fit (i.e. the coefficients change with the exclusion of observation 4 compared to the exclusion of the others).

d) Based on the ols_plot_resid_lev() function, are there any data points that are classified as having high leverage and a high studentized residual?



Although observation 4 is not fit well (has a high residual), it does not have high leverage. No points in the data set have both high studentized residual and high leverage.

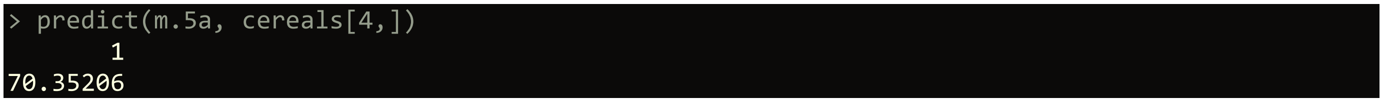
9) Based on your answer to (8), find one potentially influential observation you wish to investigate.

   Observation 4

   a) What rating did this observation have?

      Observation 4 had a rating of 93.7.

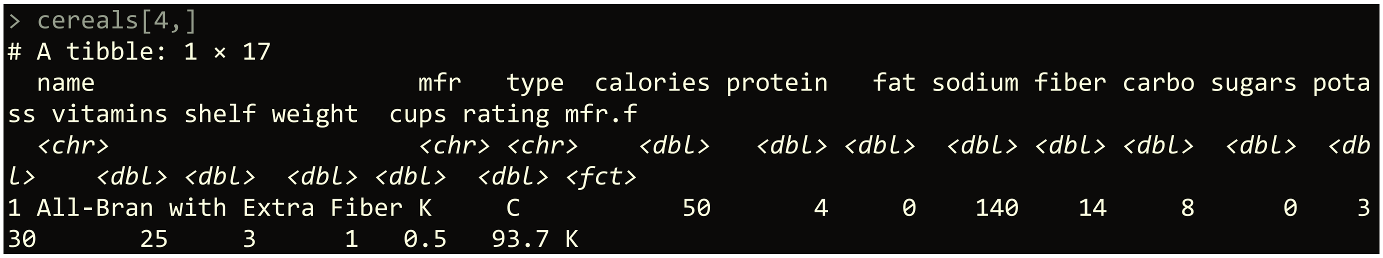
   b) What rating did your model predict this observation to have? (hint: use the predict() function)



```
> predict(m.5a, cereals[4,])
       1
70.35206
```

This observation was predicted to have a rating of 70.35

   c) How would you proceed with your model with respect to this particular observation?



```
> cereals[4,]
# A tibble: 1 × 17
  name                       mfr   type  calories protein   fat sodium fiber carbo sugars pota
ss vitamins shelf weight  cups rating mfr.f
  <chr>                      <chr> <chr>    <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>   <db
l>    <dbl> <dbl>  <dbl> <dbl>  <dbl> <fct>
1 All-Bran with Extra Fiber K     C           50       4     0    140    14     8      0    3
30       25     3      1   0.5   93.7 K
```

There is no obvious reason to exclude the data from the model, as the data looks okay, I would keep the data in the model.