Predicting rating on breakfast cereals with linear regression - Applied Statistics report

Authors: Lúcia Moreira and Nirbhaya Shaji
Date: January, 6th 2020

Summary

This report addresses the linear regression problem of predicting the rating of breakfast cereals based on 11 variables, both numerical and categorical. Best linear model for predicting rating was chosen based on analysing multicollinearity between the predictors and using variable selection techniques (based on the AIC criteria). Suitability of the chosen model was evaluated using analysis of variance tests, adjusted determination coefficient and the 95 % coefficients confidence interval. The best linear model comprises 5 numerical and 1 categorical statistically relevant variables for predicting the cereal *rating* with an adjusted determination factor of 0.965.

Dataset

The “cereals” dataset from the June, 2019 ‘lgrdata’ R package [1] summarizes 77 different brands of breakfast cereals, including calories, proteins, fats, and so on, and gives a 'rating' that indicates the overall nutritional value of the cereal. The dataset has the following 13 variables: Cereal name (character), Cereals Manufacturer (factor -letter code), Cold.or.Hot (factor 'C' or 'H'), calories (integer), protein (integer), fat (integer), sodium (integer), fiber (double), carbo (double), sugars (integer), potass (integer), vitamins (integer), Health rating of the cereal (double). The goal in this project is to predict the rating of a cereal type (a regression problem) using the 9 numerical and the 2 categorical (Manufacturer: 7 factors and Cold.or.Hot: 2 factors) variables, accounting then for 11 predictors, 1 response variable and 77 instances.

Question (a) - Data description

The data has missing values only in 4 numerical cells in total, those instances were removed.

There are two stated categorical variables, *Manufacturer* and the type of consumption of cereals named *Cold.or.Hot.* Both are of factors with 7 and 2 levels, respectively as shown in Table 1.

Table 1 : Categorical variables, levels and values

Manufacturer	Factor w/ 7 levels	“A”,”G”,”K”,”N”,”P”,”Q”,”R”
Cold.or.Hot	Factor w/ 2 levels	“C”,”H”
Vitamins	Factor w/3 levels	“0”, “25”, “100”

Cold.or.Hot categorical variable comes out as highly imbalanced with just 1 instance that is Hot “*H*” and the rest being Cold “*C*”. This way, categorical *Cold.or.Hot* was removed from our analysis. Regarding *Manufacturer*, there are 7 different manufacturers of cereals with, *K* and *Q* having the majority of instances and A present only in a single instance, so this single level was also discharged. Variable *vitamins* was also considered as an additional categorical, with 3 levels: 0, 25 and 100 (Table 1). Fig. 1 shows the number of instances in each categorical variable after removal of the non-relevant levels/categoricals. The number of total complete instances to be used in this regression problem is then 73.

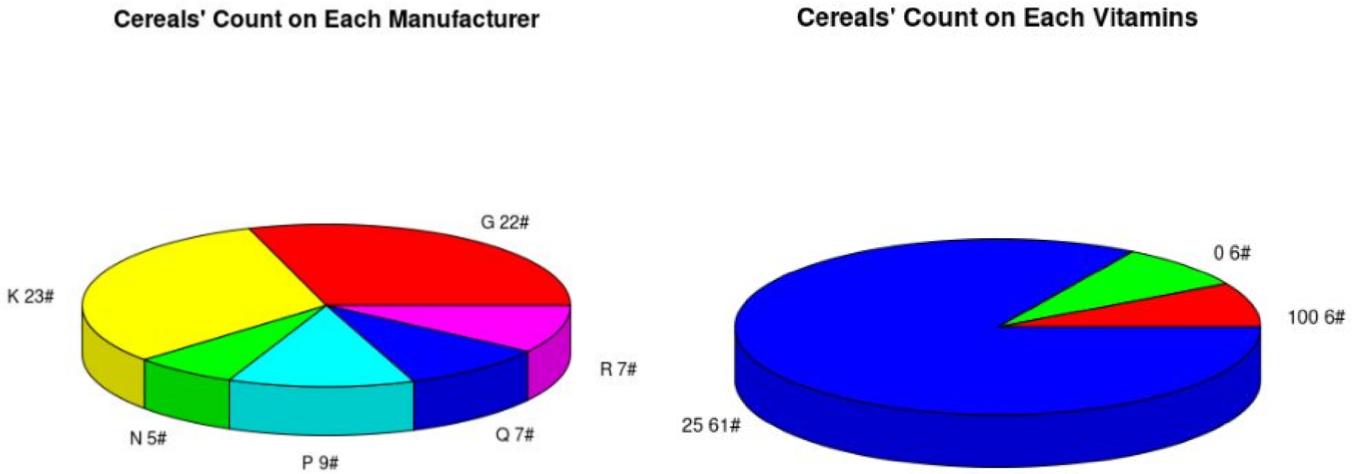


Fig. 1 : Pie diagram representing the number of instances for each level in the two categorical variables

Table 2 shows the statistics summary (min, max, median, mean, etc) for the numerical variables in the dataset. We can see that response variable *rating* is rated in a 0% -100% range. There is no information on the dimension of the variables on the dataset.

Table 2 : Numerical variables summary

Numerical Variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Calories	50	100	110	106	110	160
Protein	1	2	3	2.545	3	6
Fat	0	0	1	1.013	2	5
Sodium	0	130	180	159.7	210	320
Fiber	0	1	2	2.152	3	14
Carbo	5	12	14.5	14.8	17	23
Sugars	0	3	7	7.026	11	15
Potass	15	42	90	98	120	330
Rating	18.04	33.17	40.40	42.67	50.83	93.70

To have an initially idea towards which nutrients are essential for a nutritious breakfast per rating, plots for rating versus each nutrient was studied as shown in Fig. 2. Higher protein, fiber and potassium seems to have a positive effect on rating while calories, sugar and fat cause a negative effect on the rating of a particular cereal. It seems like the rating is more based on health concerns rather than taste.

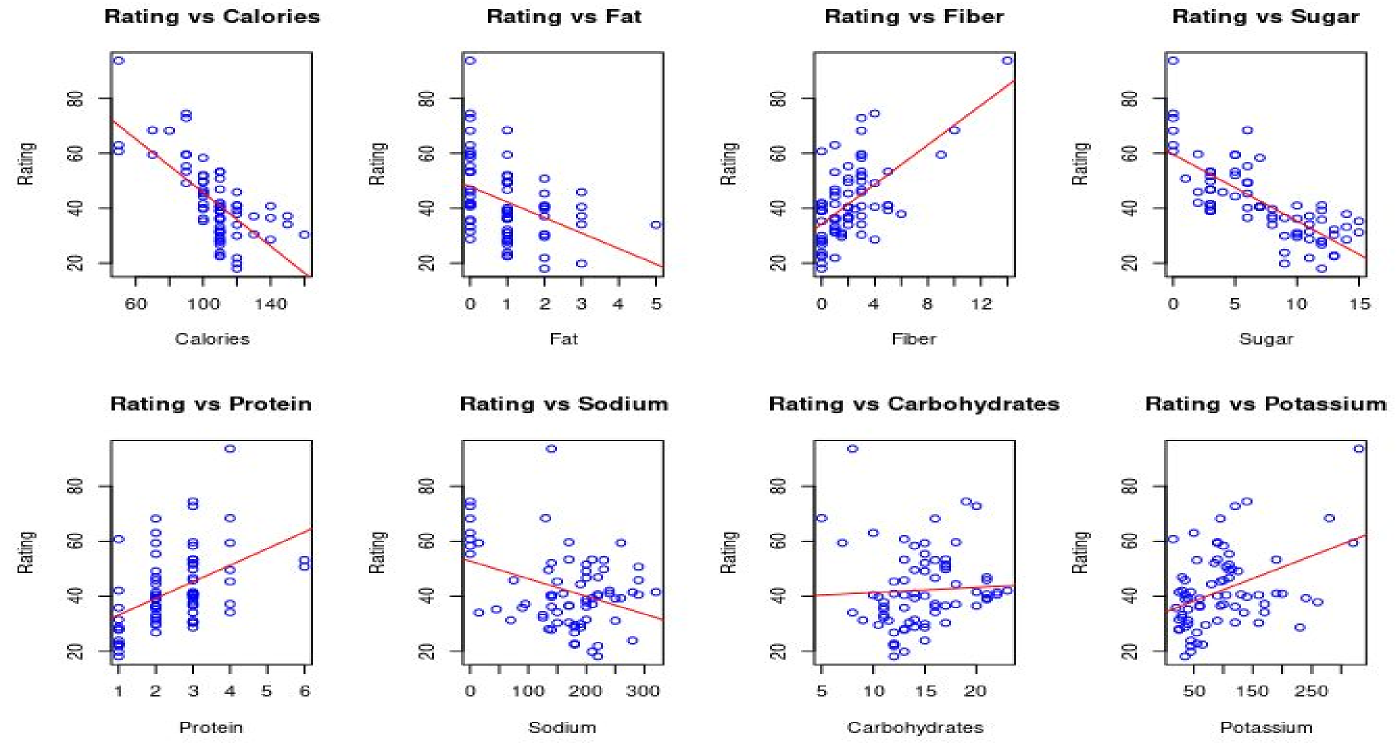


Fig. 2 : Rating vs numerical predictors.

Fig. 3 shows how the numerical predictors are distributed, a normal curve plotted over their respective histogram was made.

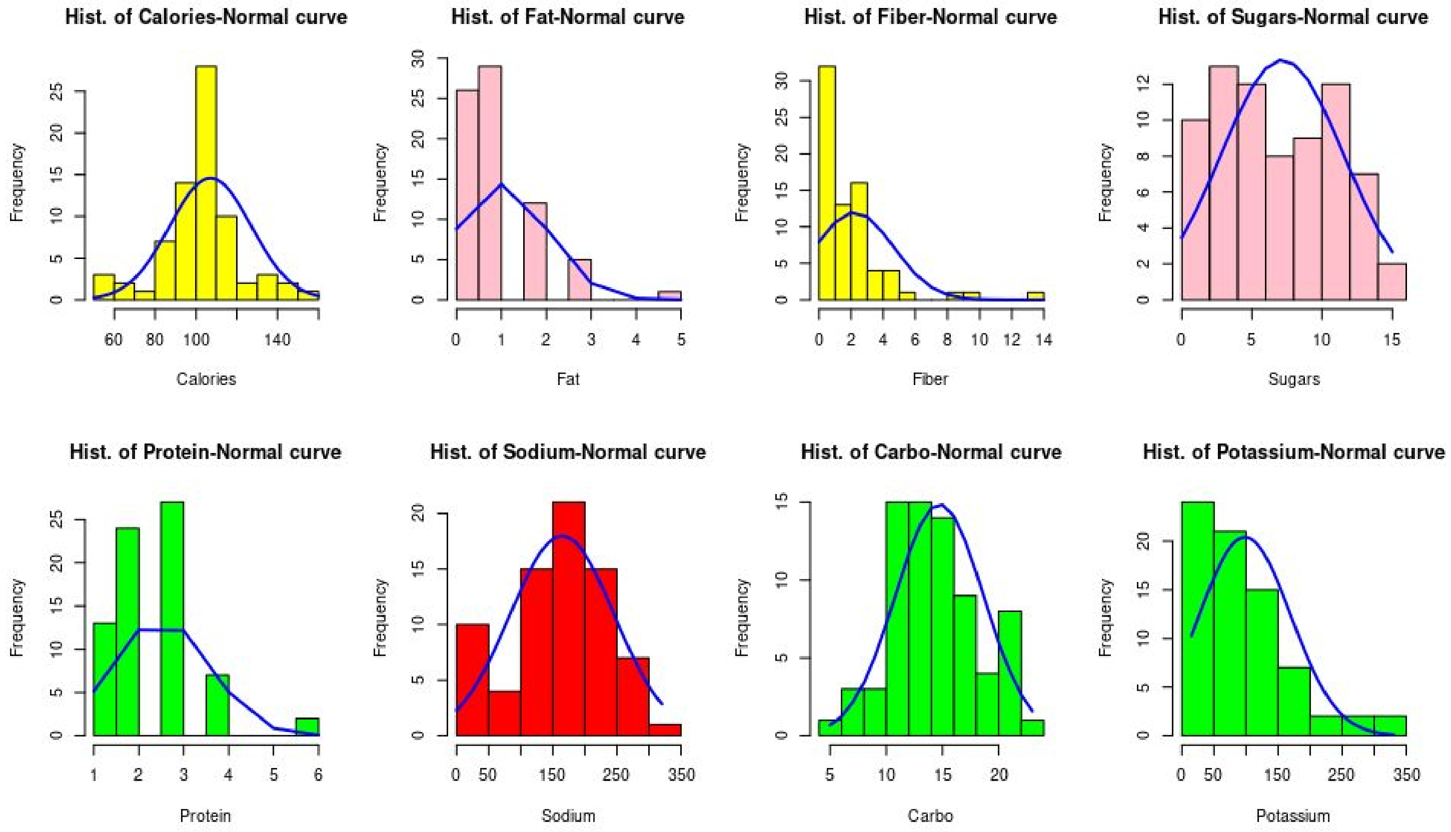


Fig. 3 : Histogram and normal curve for each numerical predictor.

Figure 4 shows the Pearson correlation coefficient between *rating* and all the other predictors. Sugars and calories variables shows a negative linear correlation whereas fiber and protein predictors show some weak positive correlation, as we saw above. Both the categorical predictors are shown as having no linear relation to rating. Strongest predictors correlation is on *fiber* and *potassium* variables with a correlation of 0.91.

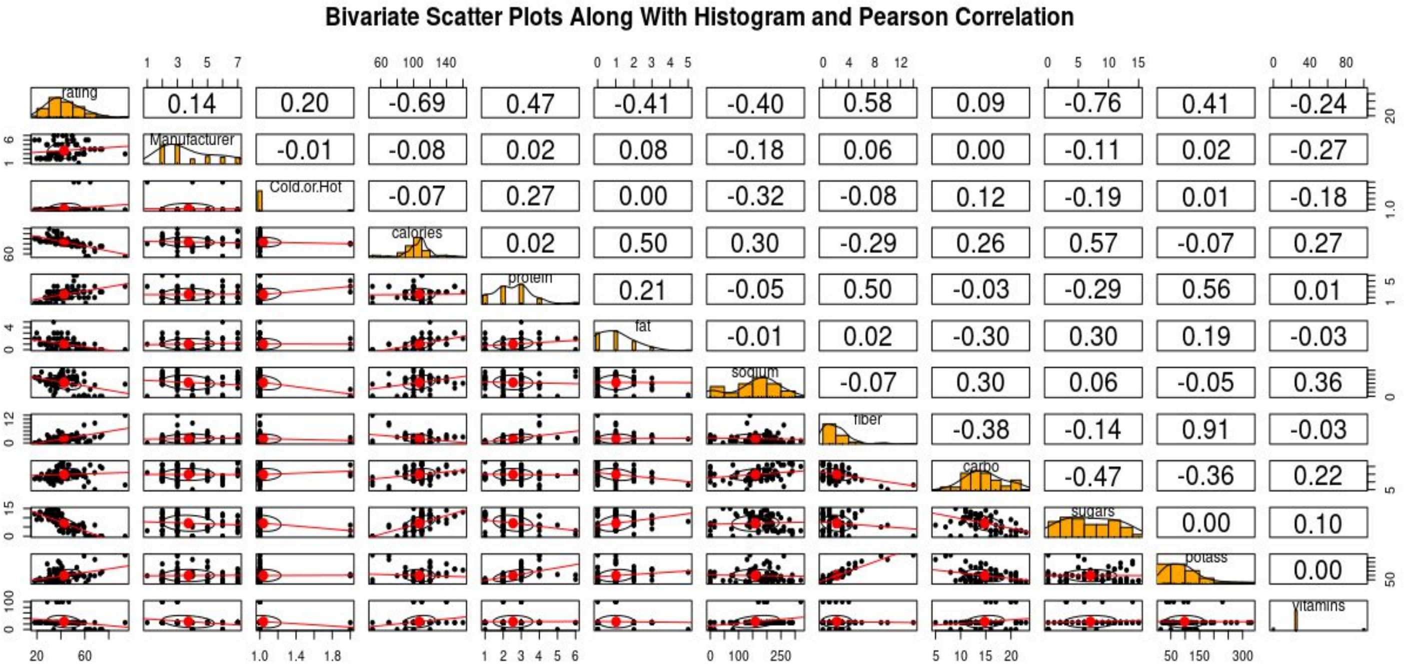


Fig. 4 : Bivariate Scatter Plots along with Histogram and Pearson Correlation

Response variable distribution

Figure 5 shows the histogram of the *rating* variable, a quasi normal distribution could be considered, despite the Shapiro-Wilk test indeed rejected the null hypothesis for a normal distribution (p-value= 0.00345). For the purpose of the present work based on a linear regression setting, *rating* will be considered having a normal distribution (considered a current practice in this type of analysis).

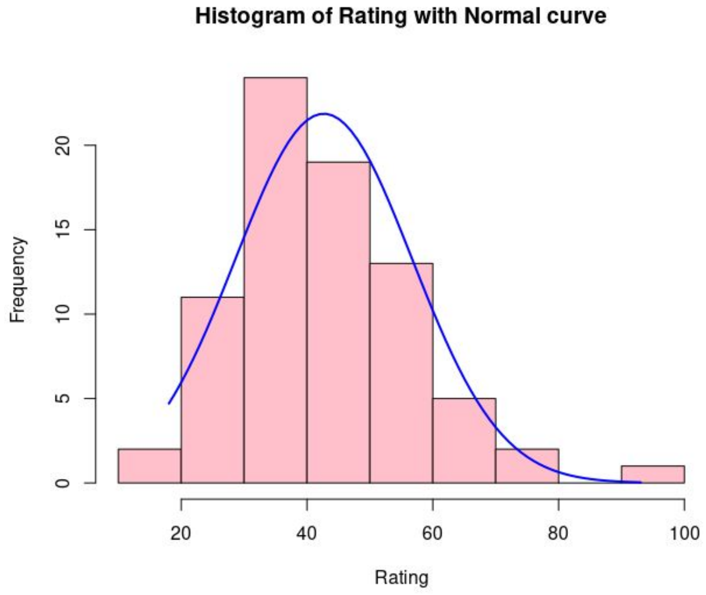


Fig. 5 :Distribution of the *rating* variable (response)

Question (b) - Linear regression model choice

A full linear model with predictors all the 8 numerical variables and the 2 categorical variables, was initially considered:

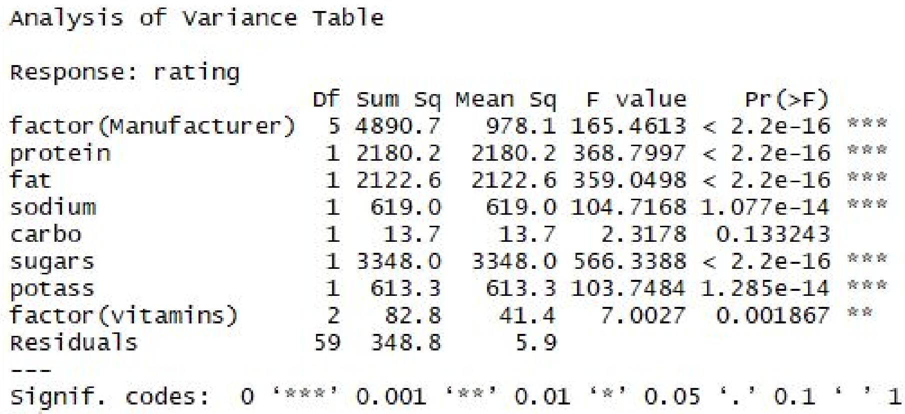
```
mod_a=lm(rating~ factor(Manufacturer)+ factor(vitamins) +calories +protein +fat +sodium +fiber +carbo +sugars
+potass, data=mydata)
```

Initial analysis considered the effect of multicollinearity. The Variance Inflation Factor (VIF) criterium was applied to mod_a. Table 3 shows the VIF values obtained from this model. One can see that using as threshold 10.3 points, the variables: *calories*, *carbo, fiber* and *sugars* show significant multicollinearity and should not be considered simultaneously in the same linear model. We performed a step-wise removal by removing in each step the variable with the highest VIF value. In the first step we have removed variable *calories*, reaching mod_b that still shows collinearity between *potass* and *fiber*, and as expected from their high correlation coefficient (see above). Finally after removal of the highest VIF value on mod_b (*fiber*), we obtain mod_c where all the colinearities were now removed (Table 3).

Table 3 : VIF values for each of the numerical variables in several models

Model	*Manufacturer*	*calories*	*protein*	*fat*	*sodium*	*fiber*	*carbo*	*sugars*	*potass*	*vitamins*
mod_a	7.16	18.99	2.93	6.32	2.38	17.38	15.92	17.83	14.41	3.93
mod_b	6.46	-	2.10	1.69	2.37	17.32	3.74	4.01	14.32	3.93
mod_c	4.75	-	2.01	1.65	2.37	-	2.64	2.24	2.30	3.37

Anova analysis of mod_c is shown in Fig. 6. We can see that the coefficient of variable *carbo* is not statistically relevant in this model.

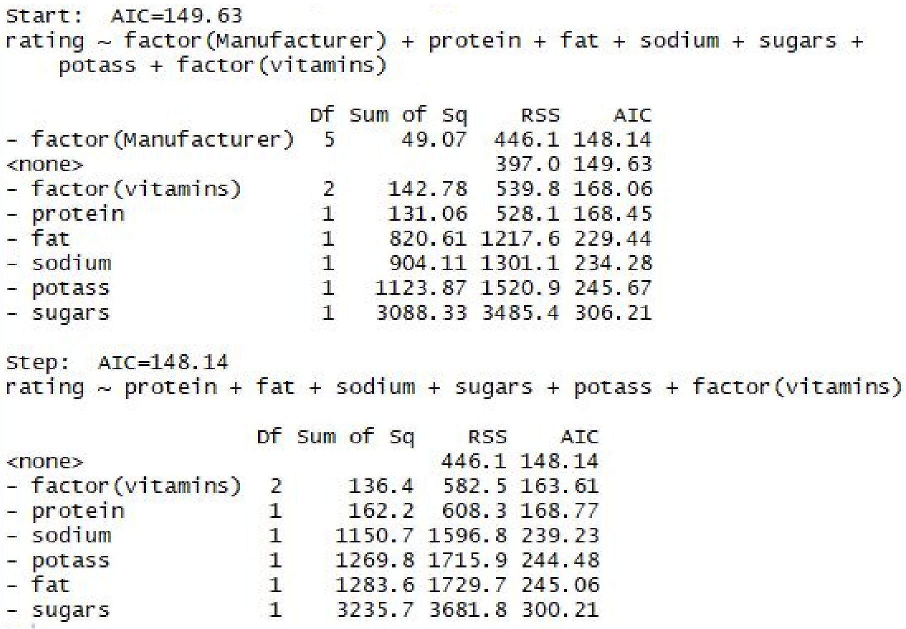


```
Analysis of Variance Table

Response: rating
                     Df Sum Sq Mean Sq  F value    Pr(>F)    
factor(Manufacturer)  5 4890.7   978.1 165.4613 < 2.2e-16 ***
protein               1 2180.2  2180.2 368.7997 < 2.2e-16 ***
fat                   1 2122.6  2122.6 359.0498 < 2.2e-16 ***
sodium                1  619.0   619.0 104.7168 1.077e-14 ***
carbo                 1   13.7    13.7   2.3178  0.133243    
sugars                1 3348.0  3348.0 566.3388 < 2.2e-16 ***
potass                1  613.3   613.3 103.7484 1.285e-14 ***
factor(vitamins)      2   82.8    41.4   7.0027  0.001867 ** 
Residuals            59  348.8     5.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 6: Anova analysis on mod_c

Next step we have considered a step-wise approach for the selection of the most relevant predictors for this problem based on the AIC criteria. The *step()* function was considered using as the starting point the model (mod_d) without the variable *carbo*, as the search algorithm *both* directions and with a scope from a null model (0 predictors) to the model (mod_c above) without collinearity effects.

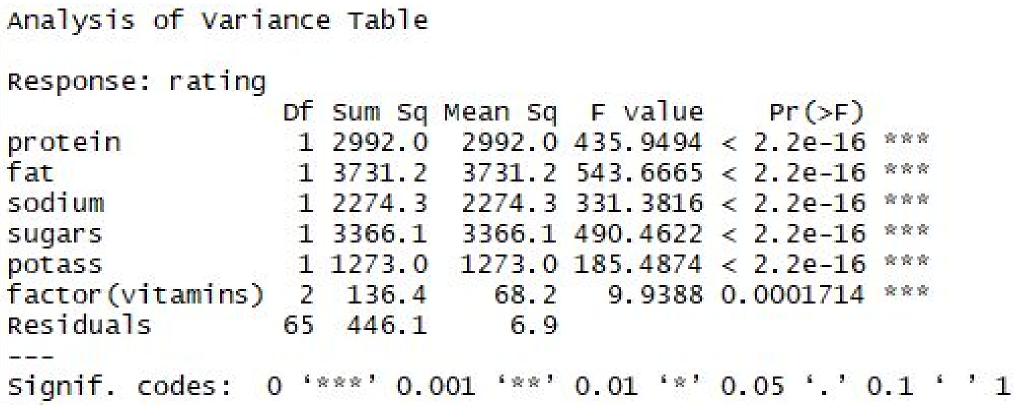
Figure 7 shows the output from the R code to perform such task. The model selected by the AIC criterium removed the categorical variable Manufacturer (mod_e).



```
Start:  AIC=149.63
rating ~ factor(Manufacturer) + protein + fat + sodium + sugars +
    potass + factor(vitamins)

                       Df Sum of Sq    RSS    AIC
- factor(Manufacturer)  5     49.07  446.1 148.14
<none>                               397.0 149.63
- factor(vitamins)      2    142.78  539.8 168.06
- protein               1    131.06  528.1 168.45
- fat                   1    820.61 1217.6 229.44
- sodium                1    904.11 1301.1 234.28
- potass                1   1123.87 1520.9 245.67
- sugars                1   3088.33 3485.4 306.21

Step:  AIC=148.14
rating ~ protein + fat + sodium + sugars + potass + factor(vitamins)

                   Df Sum of Sq    RSS    AIC
<none>                           446.1 148.14
- factor(vitamins)  2     136.4  582.5 163.61
- protein           1     162.2  608.3 168.77
- sodium            1    1150.7 1596.8 239.23
- potass            1    1269.8 1715.9 244.48
- fat               1    1283.6 1729.7 245.06
- sugars            1    3235.7 3681.8 300.21
```

Fig. 7 : Variable selection approach

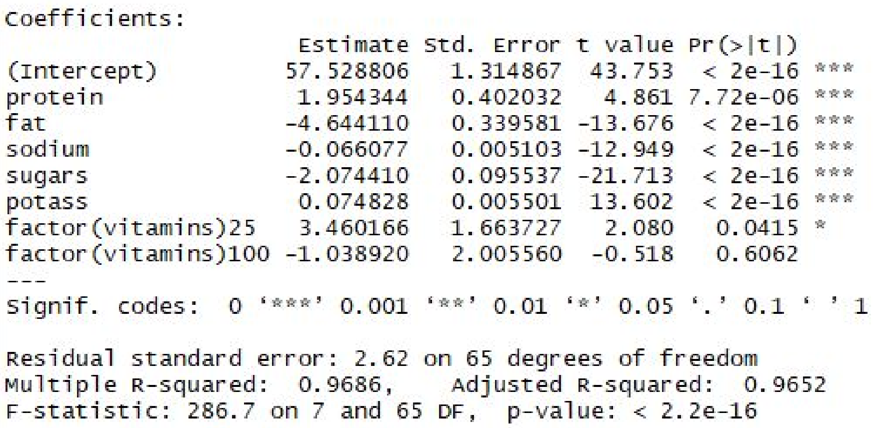
Finally, an Analysis of Variance approach was considered on output model from variable selection task (mod_e). Figure 8 shows the outcome from such analysis. One can see that all the coefficients are statistically significant from the anova analysis, being our final chosen model.



```
Analysis of Variance Table

Response: rating
                 Df Sum Sq Mean Sq  F value    Pr(>F)    
protein           1 2992.0  2992.0 435.9494 < 2.2e-16 ***
fat               1 3731.2  3731.2 543.6665 < 2.2e-16 ***
sodium            1 2274.3  2274.3 331.3816 < 2.2e-16 ***
sugars            1 3366.1  3366.1 490.4622 < 2.2e-16 ***
potass            1 1273.0  1273.0 185.4874 < 2.2e-16 ***
factor(vitamins)  2  136.4    68.2   9.9388 0.0001714 ***
Residuals        65  446.1     6.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 8 : Anova summary output from mod_e

Fig. 9 shows a summary of the best model chosen. One can see that the adjusted determination coefficient from this fitting is 0.9686 indicating that 96.9% of the variability of the *rating* response is described by the linear model, so this model explains almost all the variance in the response variable.



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         57.528806   1.314867  43.753  < 2e-16 ***
protein              1.954344   0.402032   4.861 7.72e-06 ***
fat                 -4.644110   0.339581 -13.676  < 2e-16 ***
sodium              -0.066077   0.005103 -12.949  < 2e-16 ***
sugars              -2.074410   0.095537 -21.713  < 2e-16 ***
potass               0.074828   0.005501  13.602  < 2e-16 ***
factor(vitamins)25   3.460166   1.663727   2.080   0.0415 *  
factor(vitamins)100 -1.038920   2.005560  -0.518   0.6062    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.62 on 65 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.9652 
F-statistic: 286.7 on 7 and 65 DF,  p-value: < 2.2e-16
```

Fig. 9 : Summary output from the selected model.

Fig. 10 shows the analysis of the residuals for the chosen model (model_d). Results show that residuals statistically follow a normal distribution, pointing to the adequacy of the model.

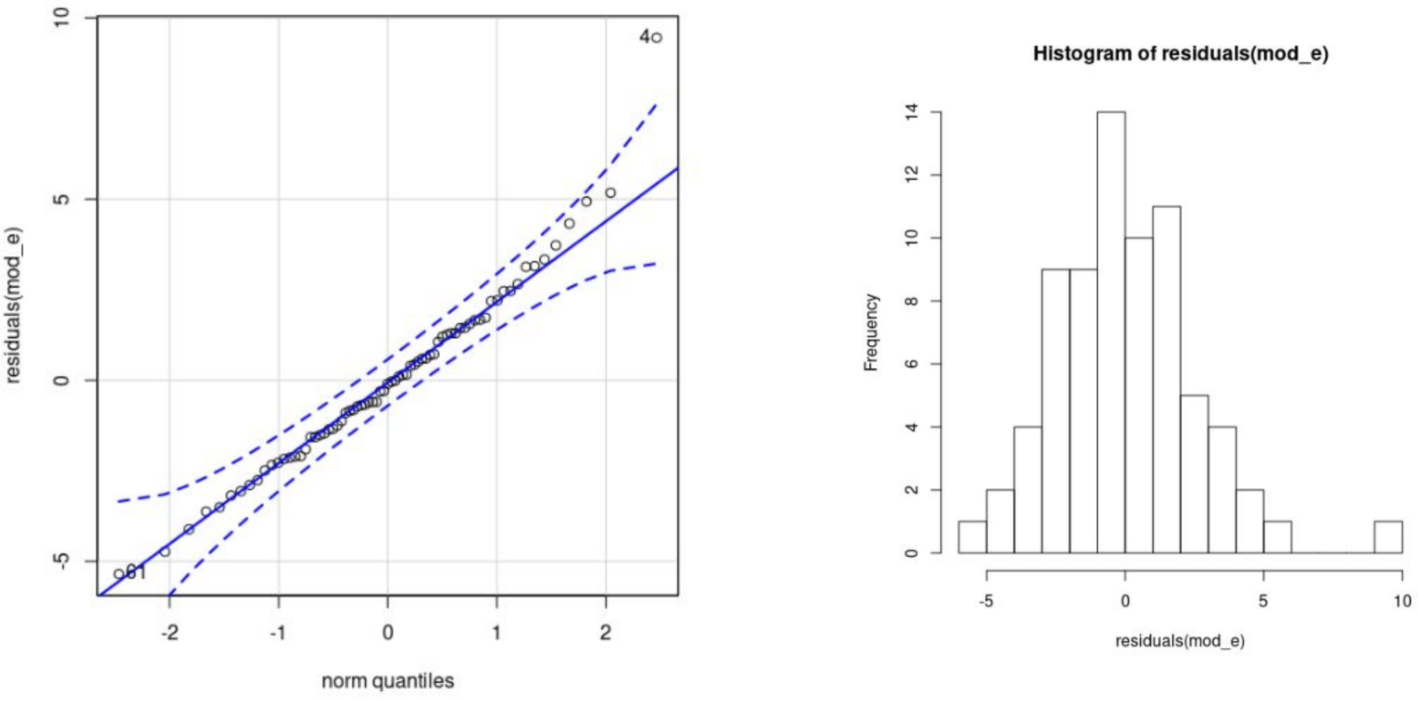
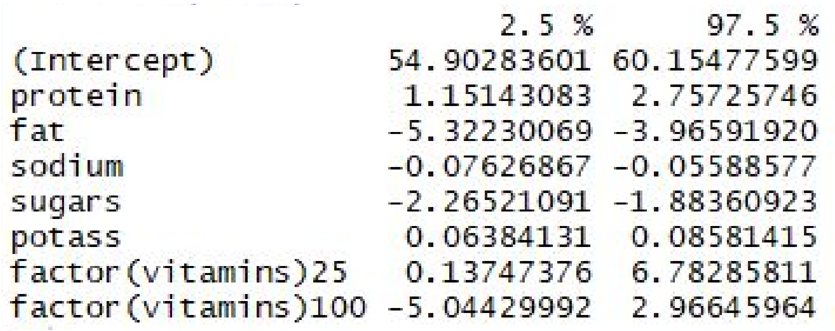


Fig. 10 : Residuals analysis: qqplot(left), histogram(right)

An analysis of the confidence intervals for the variable coefficients in the chosen model and for a 95 % confidence level is shown on Fig. 11. Results shows that indeed none of the regression coefficients are zero for the numerical variables.



```
                          2.5 %      97.5 %
(Intercept)         54.90283601 60.15477599
protein              1.15143083  2.75725746
fat                 -5.32230069 -3.96591920
sodium              -0.07626867 -0.05588577
sugars              -2.26521091 -1.88360923
potass               0.06384131  0.08581415
factor(vitamins)25   0.13747376  6.78285811
factor(vitamins)100 -5.04429992  2.96645964
```

Fig. 11 : 95 % confidence intervals for the coefficients of the chosen model (mod_e).

Question (c) - i, ii, iii, iv, v

i) Choosing X_1 as protein and X_2 as Vitamins, for studying the gross and adjusted effects, first, two simple linear regression models (model1 and model2) with each of these predictors were considered for the gross effects.

The gross effect of protein in model1 is +6.031, which means for each unit of increase in protein variable there is an increase in rating of 6.031. The gross effect of Vitamin "25", with respect to the reference factor Vitamin 0, on rating in model2 is -21.573. And for Vitamin 100 its -24.111. Using our final model, model_e, the adjusted effect of protein dropped to +1.95 while for Vitamin 25 and 100 it increases to +3.46 and -1.03, respectively.

Ii) The 95% confidence (upper: blue, lower: red) and prediction (green) intervals for the *rating* variable as a function of the numerical *protein* variable are shown in Fig. 12 while using the mean value for the remaining ones and mode for the categorical. Fig. 13 shows the correspondent values.

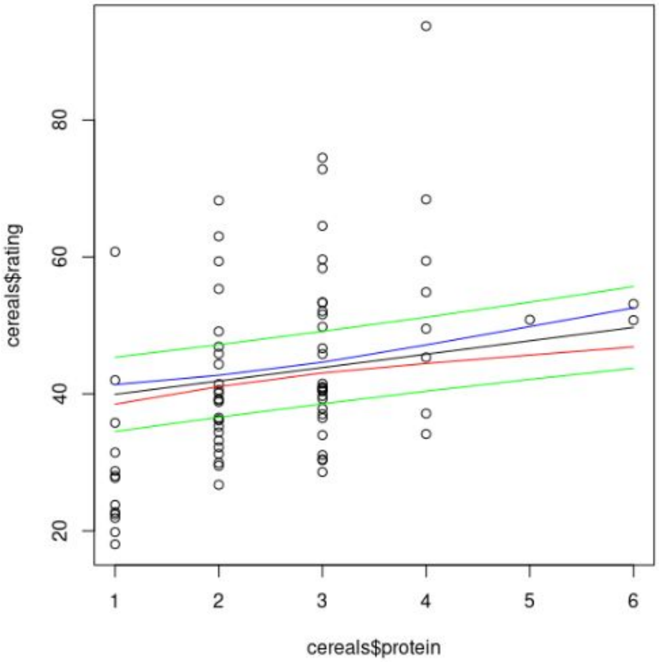
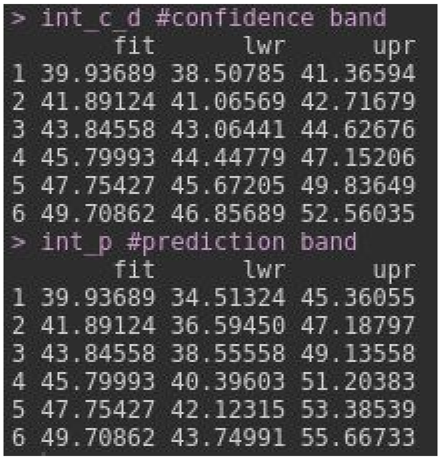


Fig. 12 : Prediction (green) and confidence (upper: blue, lower: red) intervals for rating as a function of protein content (mean values were considered for the other numerical variables and the mode for the categorical one).



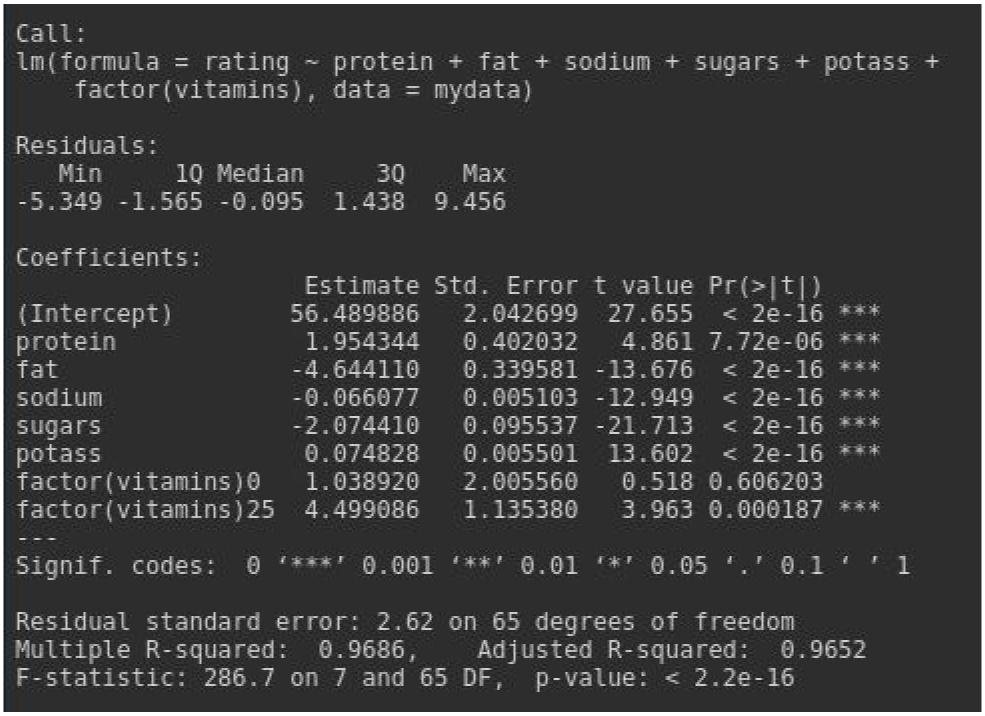
```
> int_c_d #confidence band
       fit      lwr      upr
1 39.93689 38.50785 41.36594
2 41.89124 41.06569 42.71679
3 43.84558 43.06441 44.62676
4 45.79993 44.44779 47.15206
5 47.75427 45.67205 49.83649
6 49.70862 46.85689 52.56035
> int_p #prediction band
       fit      lwr      upr
1 39.93689 34.51324 45.36055
2 41.89124 36.59450 47.18797
3 43.84558 38.55558 49.13558
4 45.79993 40.39603 51.20383
5 47.75427 42.12315 53.38539
6 49.70862 43.74991 55.66733
```

Fig. 13 : Prediction and confidence bands for rating as a function of protein content.

iii) For categorical variables the rate is relative to the given base line. For our data the base level is Vitamin 0. From summary of our selected final model (Fig. 9), we can see that vitamin 25 has an estimate of 3.46, which means the rating is 3.46 higher among vitamin 25 compared to vitamin 0. Also for vitamin 100 the rating is -1.038 higher compared to vitamin 0.

To see the effect on the response from changing from out category variable's level third (100) to level second (25), we can add the symmetrical or negative value of current third's effect to current second's effect, ie -(−1.03) + 3.460 = 4.498.

To confirm the above value of 4.498, we can relevel the factors to a baseline of vitamin 100 and generate the model once again to see the estimates of vitamin 25.



```
Call:
lm(formula = rating ~ protein + fat + sodium + sugars + potass +
    factor(vitamins), data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-5.349 -1.565 -0.095  1.438  9.456

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        56.489886   2.042699  27.655  < 2e-16 ***
protein             1.954344   0.402032   4.861 7.72e-06 ***
fat                -4.644110   0.339581 -13.676  < 2e-16 ***
sodium             -0.066077   0.005103 -12.949  < 2e-16 ***
sugars             -2.074410   0.095537 -21.713  < 2e-16 ***
potass              0.074828   0.005501  13.602  < 2e-16 ***
factor(vitamins)0   1.038920   2.005560   0.518 0.606203
factor(vitamins)25  4.499086   1.135380   3.963 0.000187 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.62 on 65 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.9652
F-statistic: 286.7 on 7 and 65 DF,  p-value: < 2.2e-16
```

Fig 14: Final model summary after releveling the category variables factor.

The confidence band values for 90% and 95% does not change significantly when changing from Vitamin 25 to Vitamin 100 (Table 4).

Table 4 : Confidence and prediction bands for 90% and 95%

Confidence Interval	90%	95%	90%	95%	90%	95%
	fit		Upper limit		Lower limit	
Vitamin 25	42.855	42.855	42.158	42.158	43.551	43.552
Vitamin 100	38.356	38.356	36.172	36.172	40.540	40.540

iv) Table 5 shows that the standardized coefficients on our final mod_e. The standardized coefficient for protein shows that for a change in 1 standard deviation of our X_1 protein, there is a change of 0.149 to the response rating, which is lower than the other numerical predictors.

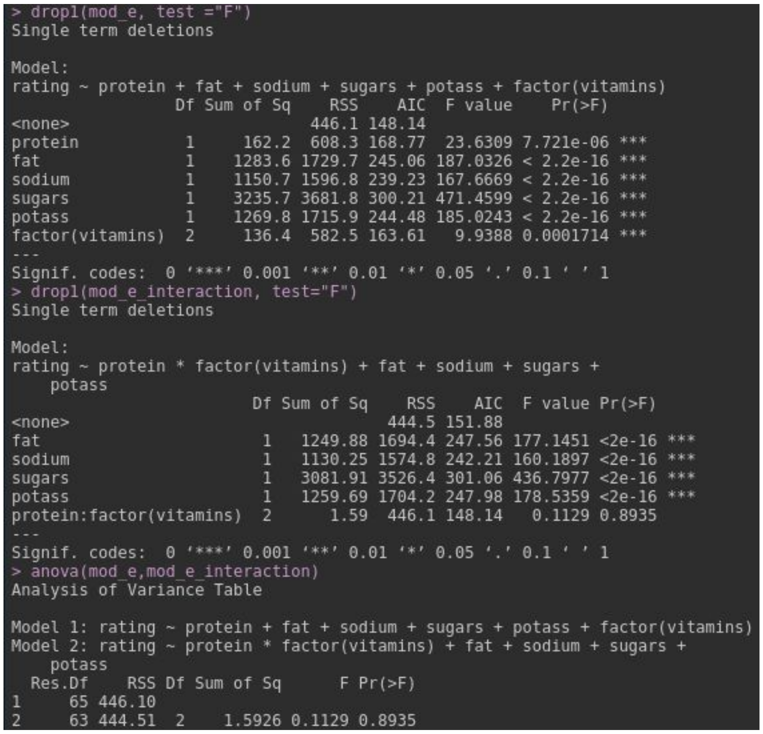
Table 5: Standardized coefficients from mod_e (final model)

protein	fat	sodium	sugars	potass	factor(vitamins)25	factor(vitamins)100
0.149	-0.335	-0.381	-0.644	0.380	0.092	-0.020

v) From the anova table analysis of our final model (mod_e):

```
mod_e_interaction=lm(rating~ calories + protein*factor(Vitamins) + fat + sodium +  sugars + potass + vitamins,
data=mydata)
```

we can see in the below Fig. 15 that, from the results of drop1 and anova() that considering the interaction is not providing any significant improvement to the model.



```
> drop1(mod_e, test ="F")
Single term deletions

Model:
rating ~ protein + fat + sodium + sugars + potass + factor(vitamins)
                 Df Sum of Sq    RSS    AIC  F value    Pr(>F)    
<none>                         446.1 148.14                       
protein           1     162.2  608.3 168.77  23.6309 7.721e-06 ***
fat               1    1283.6 1729.7 245.06 187.0326 < 2.2e-16 ***
sodium            1    1150.7 1596.8 239.23 167.6669 < 2.2e-16 ***
sugars            1    3235.7 3681.8 300.21 471.4599 < 2.2e-16 ***
potass            1    1269.8 1715.9 244.48 185.0243 < 2.2e-16 ***
factor(vitamins)  2     136.4  582.5 163.61   9.9388 0.0001714 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(mod_e_interaction, test="F")
Single term deletions

Model:
rating ~ protein * factor(vitamins) + fat + sodium + sugars + 
    potass
                         Df Sum of Sq    RSS    AIC  F value Pr(>F)    
<none>                                 444.5 151.88                    
fat                       1   1249.88 1694.4 247.56 177.1451 <2e-16 ***
sodium                    1   1130.25 1574.8 242.21 160.1897 <2e-16 ***
sugars                    1   3081.91 3526.4 301.06 436.7977 <2e-16 ***
potass                    1   1259.69 1704.2 247.98 178.5359 <2e-16 ***
protein:factor(vitamins)  2      1.59  446.1 148.14   0.1129 0.8935    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(mod_e,mod_e_interaction)
Analysis of Variance Table

Model 1: rating ~ protein + fat + sodium + sugars + potass + factor(vitamins)
Model 2: rating ~ protein * factor(vitamins) + fat + sodium + sugars + 
    potass
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     65 446.10                           
2     63 444.51  2    1.5926 0.1129 0.8935
```

Fig. 15 : drop1 and ANOVA test responses for mod_e and mod_e_interaction

Reference

[1] -   “Example Datasets for a Learning Guide to R”, https://cran.r-project.org/web/packages/lgrdata/lgrdata.pdf, June 2019.