SDM_Assignment1_2

Sri Balaji Muruganandam

17/09/2021

Setting Working Directory

```
rm(list = ls())
setwd("G:\\SDM_Sem01\\Assignment1")
```

Loading the preprocessed data

```
load("cereal_clean_data.RData")
```

2) Perform a multiple regression on the dataset you pre-processed in question one. The response variable is rating. Use the lm() function in R.

Including all the predictors except name.

```
model_fit <- lm(rating~.-name, data = c_data)
names(model_fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"       "call"          "terms"
## [13] "model"
```

Computing Confidence Interval

```
confint(model_fit)
```

```
##                    2.5 %        97.5 %
## (Intercept)  0.5843316932  0.6246301838
## mfr         -0.0012710095  0.0003431790
## type         0.0025351237  0.0249819691
## calories    -0.0022635401 -0.0018438363
## protein      0.0306215517  0.0343225327
## fat         -0.0205679211 -0.0159130072
## sodium      -0.0005834055 -0.0005469454
## fiber        0.0279373436  0.0303700158
## carbo        0.0096321404  0.0111289711
## sugars      -0.0091044860 -0.0076213531
## potass      -0.0479186812 -0.0292800723
## vitamins    -0.0538802654 -0.0402138711
## shelf       -0.0038746038 -0.0002301983
## weight      -0.0262464681  0.0143064458
## cups        -0.0117798965  0.0017035100
```

Summary of Regression model

```
summary(model_fit)
```

```
##
## Call:
## lm(formula = rating ~ . - name, data = c_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.014520 -0.002401  0.000114  0.002989  0.011609
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.045e-01  1.007e-02  60.009  < 2e-16 ***
## mfr         -4.639e-04  4.035e-04  -1.150   0.2548
## type         1.376e-02  5.611e-03   2.452   0.0171 *
## calories    -2.054e-03  1.049e-04 -19.576  < 2e-16 ***
## protein      3.247e-02  9.251e-04  35.101  < 2e-16 ***
## fat         -1.824e-02  1.164e-03 -15.676  < 2e-16 ***
## sodium      -5.652e-04  9.114e-06 -62.014  < 2e-16 ***
## fiber        2.915e-02  6.081e-04  47.944  < 2e-16 ***
## carbo        1.038e-02  3.742e-04  27.744  < 2e-16 ***
## sugars      -8.363e-03  3.707e-04 -22.558  < 2e-16 ***
## potass      -3.860e-02  4.659e-03  -8.285  1.6e-11 ***
## vitamins    -4.705e-02  3.416e-03 -13.772  < 2e-16 ***
## shelf       -2.052e-03  9.110e-04  -2.253   0.0279 *
## weight      -5.970e-03  1.014e-02  -0.589   0.5581
## cups        -5.038e-03  3.370e-03  -1.495   0.1402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005211 on 60 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic:  3796 on 14 and 60 DF,  p-value: < 2.2e-16
```

Predicting for a new set of values

```
model_predict <- predict(model_fit, newdata = data.frame("name"=c("Kix","Life"),"mfr" = c(2,3
),"type" = c(1,1),"calories"= c(115,130),"protein"= c(2,4),"fat" = c(2,0),"sodium" = c(170,20
0),"fiber" = c(2.0,5.0),"carbo" = c(12.0,5.0),"sugars" = c(8,2),"potass" = c(1.5,2.0),"vitami
ns" = c(0.25,1.0),"shelf" = c(3,1),"weight" = c(1.00,1.00),"cups" = c(0.75,0.67)))
```

Summary of prediction

```
model_predict
```

```
##         1         2
## 0.3439248 0.4120229
```

```
summary(model_predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3439  0.3609  0.3780  0.3780  0.3950  0.4120
```

We can see that for prediction 1, the rating is 0.343 and for prediction 2 the rating is 0.412

a) Which predictors appear to have a significant relationship to the response.

From the summary, we can find that the predictors **protein, fiber, carbo** have a significant relationship to the response.

b) What does the coefficient variable for “sugar” suggest?

```
cor(c_data[,2:16])
```

```
##                  mfr        type    calories     protein         fat
## mfr       1.000000000  0.01754318 -0.08157807  0.02943179  0.05747391
## type      0.017543182  1.00000000  0.05856933 -0.29419035 -0.08056080
## calories -0.081578067  0.05856933  1.00000000  0.02212511  0.49887923
## protein   0.029431794 -0.29419035  0.02212511  1.00000000  0.22389823
## fat       0.057473907 -0.08056080  0.49887923  0.22389823  1.00000000
## sodium   -0.191709757  0.31659458  0.29773324 -0.04681618 -0.02470907
## fiber     0.071510035  0.05762794 -0.29590305  0.50276606  0.01682321
## carbo    -0.065735925  0.27390469  0.26220057 -0.14247863 -0.30496318
## sugars   -0.140959856  0.22571070  0.56448646 -0.32536180  0.25456808
## potass    0.007118336 -0.06439729  0.03497074  0.64777852  0.30575879
## vitamins -0.277742226  0.12028742  0.26285797  0.01345222 -0.04715353
## shelf    -0.020743053  0.13914915  0.09479824  0.14278818  0.25302498
## weight   -0.240634574  0.03323600  0.69663325  0.21665127  0.21751316
## cups     -0.062941939 -0.01093062  0.09194001 -0.25250594 -0.16465689
## rating    0.158779897 -0.12349288 -0.69437729  0.46819070 -0.39337642
##                sodium       fiber       carbo      sugars       potass
## mfr       -0.19170976  0.07151003 -0.06573592 -0.14095986  0.007118336
## type       0.31659458  0.05762794  0.27390469  0.22571070 -0.064397292
## calories   0.29773324 -0.29590305  0.26220057  0.56448646  0.034970737
## protein   -0.04681618  0.50276606 -0.14247863 -0.32536180  0.647778523
## fat       -0.02470907  0.01682321 -0.30496318  0.25456808  0.305758791
## sodium     1.00000000 -0.07453352  0.38435115  0.08231043 -0.093679370
## fiber     -0.07453352  1.00000000 -0.35367820 -0.15280645  0.791016449
## carbo      0.38435115 -0.35367820  1.00000000 -0.31001634 -0.261815781
## sugars     0.08231043 -0.15280645 -0.31001634  1.00000000 -0.052788691
## potass    -0.09367937  0.79101645 -0.26181578 -0.05278869  1.000000000
## vitamins   0.35293551 -0.04203515  0.29062165  0.10216296  0.001883807
## shelf     -0.07991557  0.30504207 -0.09738592  0.09462118  0.357543454
## weight     0.30981705  0.24556027  0.14094851  0.45495357  0.450101236
## cups       0.13284731 -0.51398000  0.35507984 -0.01585318 -0.507586139
## rating    -0.38805034  0.60350256  0.02078492 -0.75184121  0.386140074
##                vitamins       shelf     weight        cups      rating
## mfr       -0.277742226 -0.02074305 -0.2406346 -0.06294194  0.15877990
## type       0.120287417  0.13914915  0.0332360 -0.01093062 -0.12349288
## calories   0.262857969  0.09479824  0.6966333  0.09194001 -0.69437729
## protein    0.013452216  0.14278818  0.2166513 -0.25250594  0.46819070
## fat       -0.047153533  0.25302498  0.2175132 -0.16465689 -0.39337642
## sodium     0.352935508 -0.07991557  0.3098171  0.13284731 -0.38805034
## fiber     -0.042035149  0.30504207  0.2455603 -0.51398000  0.60350256
## carbo      0.290621652 -0.09738592  0.1409485  0.35507984  0.02078492
## sugars     0.102162961  0.09462118  0.4549536 -0.01585318 -0.75184121
## potass     0.001883807  0.35754345  0.4501012 -0.50758614  0.38614007
## vitamins   1.000000000  0.30231739  0.3202295  0.14292499 -0.22201105
## shelf      0.302317394  1.00000000  0.1939763 -0.33261530  0.03848749
## weight     0.320229471  0.19397627  1.0000000 -0.19935746 -0.30127603
## cups       0.142924990 -0.33261530 -0.1993575  1.00000000 -0.22655393
## rating    -0.222011054  0.03848749 -0.3012760 -0.22655393  1.00000000
```

The coefficeient variable for sugar is -0.75184121

c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

Using : to fit data. It is used to represent the set of predictors

```
model_fit2 <- lm(rating~mfr:cups, data = c_data)
summary(model_fit2)
```

```
##
## Call:
## lm(formula = rating ~ mfr:cups, data = c_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24421 -0.09794 -0.02112  0.08316  0.51200
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4252588  0.0331616  12.824   <2e-16 ***
## mfr:cups    -0.0001382  0.0096651  -0.014    0.989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1407 on 73 degrees of freedom
## Multiple R-squared:  2.8e-06,    Adjusted R-squared:  -0.0137
## F-statistic: 0.0002044 on 1 and 73 DF,  p-value: 0.9886
```

Using * to fit data. If we are creating a new feature my multipying two predictors we can use *

As protein and fiber have high correlation between rating, creating a new feature by multiplying both.

```
model_fit3 <- lm(rating~.-name + protein*fiber, data = c_data)
summary(model_fit3)
```

```
##
## Call:
## lm(formula = rating ~ . - name + protein * fiber, data = c_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0130604 -0.0021736  0.0002117  0.0025579  0.0098650
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.160e-01  1.018e-02  60.520  < 2e-16 ***
## mfr           -5.626e-04  3.797e-04  -1.482  0.14373
## type           1.045e-02  5.372e-03   1.946  0.05648 .
## calories      -2.059e-03  9.838e-05 -20.932  < 2e-16 ***
## protein        3.401e-02  1.003e-03  33.890  < 2e-16 ***
## fat           -1.769e-02  1.106e-03 -16.005  < 2e-16 ***
## sodium        -5.561e-04  9.055e-06 -61.410  < 2e-16 ***
## fiber          3.409e-02  1.719e-03  19.832  < 2e-16 ***
## carbo          1.036e-02  3.509e-04  29.519  < 2e-16 ***
## sugars        -8.278e-03  3.487e-04 -23.739  < 2e-16 ***
## potass        -4.527e-02  4.887e-03  -9.263 4.18e-13 ***
## vitamins      -4.763e-02  3.209e-03 -14.845  < 2e-16 ***
## shelf         -1.761e-03  8.594e-04  -2.049  0.04493 *
## weight        -1.006e-02  9.598e-03  -1.048  0.29881
## cups          -4.628e-03  3.163e-03  -1.463  0.14868
## protein:fiber -1.211e-03  3.979e-04  -3.043  0.00349 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004886 on 59 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.9988
## F-statistic:  4031 on 15 and 59 DF,  p-value: < 2.2e-16
```

```
model_fit4 <- lm(rating~.-name + type*protein, data = c_data)
summary(model_fit4)
```

```
##
## Call:
## lm(formula = rating ~ . - name + type * protein, data = c_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.014498 -0.002215  0.000000  0.002932  0.011625
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.163e-01  5.730e-02  10.757 1.56e-15 ***
## mfr          -4.399e-04  4.224e-04  -1.041   0.3019
## type          1.898e-03  5.666e-02   0.033   0.9734
## calories     -2.037e-03  1.332e-04 -15.295  < 2e-16 ***
## protein       2.953e-02  1.400e-02   2.109   0.0392 *
## fat          -1.840e-02  1.406e-03 -13.094  < 2e-16 ***
## sodium       -5.650e-04  9.222e-06 -61.270  < 2e-16 ***
## fiber         2.912e-02  6.359e-04  45.792  < 2e-16 ***
## carbo         1.027e-02  6.437e-04  15.956  < 2e-16 ***
## sugars       -8.460e-03  5.952e-04 -14.215  < 2e-16 ***
## potass       -3.859e-02  4.697e-03  -8.215 2.37e-11 ***
## vitamins     -4.698e-02  3.458e-03 -13.588  < 2e-16 ***
## shelf        -2.075e-03  9.243e-04  -2.244   0.0286 *
## weight       -5.274e-03  1.074e-02  -0.491   0.6253
## cups         -4.926e-03  3.439e-03  -1.432   0.1573
## type:protein  2.862e-03  1.360e-02   0.210   0.8341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005253 on 59 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic:  3486 on 15 and 59 DF,  p-value: < 2.2e-16
```

type X protein and type X fiber interactions are significant when used along with the model