实验9、多元线性回归分析

年级： 15级　　专业：生信　　学号：1513401013　　姓名：郑磊

编号	一	二	三	四	总分	评阅人
得分						

软硬件平台：

1. 硬件平台：（硬件配置）i5，2.9HZ处理器，16G内存，64位操作系统

2. 系统平台：（操作系统及其版本号）Windows10 企业版

3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）R3.4.1　，Rstudio

一、目的要求：

1、加深对多元线性回归分析的理解；

2、加深对哑变量（Proxy and dummy variables）的理解；

3、熟悉并掌握多元线性回归分析所涉及的R语言函数和脚本。

二、实验内容：

1、以“Healthy_Breakfast”的数据为例，进行多元回归分析（不考虑哑变量）：

以rating作为因变量，其他数值型数据列作为自变量，进行多元回归分析，分析过程的R语言代码参考理论授课环节的PPT。具体过程如下：

1.1、 读取数据表；

```
dir="D:/RFile/实验九"

setwd(dir) # 设定工作目录

file="Data_Healthy_Breakfast.txt"

data<-read.table(file,head=TRUE,sep="\t")

#查看数据信息
```

```
head(data)

ncol(data)

nrow(data)
```

1.2、两两组合绘制散点图和拟合曲线，从总体上看看不同变量之间的关联：

```
png("lec10_Healthy_Breakfast_pairs.png")

pairs(data[,4:16],panel=panel.smooth)

dev.off()
```

1.3、以“rating”数据列为因变量(y),对其他所有数据列进行多元回归分析，并查看分析结果；

```
lm0<-lm(rating~.,data=data[,4:16])

summary(lm0)
```

1.4、以【向后】逐步回归法计算最终多元回归模型（记录逐步回归的结果），并查看分析结果（summary）,并根据分析结果，写出相应的多元线性方程；

```
lm.step<-step(lm0,direction="backward")

summary(lm.step)
```

1.5、查看回归结果的统计图谱，通过这四张统计图来讨论回归结果的可靠性；

```
png(file = "lec10_Healthy_Breakfast_lm_data.png")

par(mfrow=c(2,2)) #同一个图形文件中绘制2*2=4个图像

plot(lm.step)

dev.off()
```

1.6、多重共线性分析：

理想中的线性模型各个自变量应该是线性无关的，若自变量间存在共线

性，则会降低回归系数的准确性。一般用方差膨胀因子VIF(Variance Inflation Factor)来衡量共线性，《统计学习》中认为VIF超过5或10就存在共线性，《R语言实战》中认为VIF大于4则存在共线性。理想中的线性模型VIF=1，表完全不存在共线性。

```
library(car)
vif(lm.step)
```

1.7、检查离群点、高杠杆点、强影响点，保存屏幕反馈结果（如果有的话）和统计图：纵坐标超过+2或小于-2的点可被认为是离群点，水平轴超过0.2或0.3的就是高杠杆值（通常为预测值的组合）。圆圈大小与影响成比例，圆圈很大的点可能是对模型参数的估计造成的不成比例影响的强影响点。

```
#car包里influencePlot()函数能一次性同时检查离群点、高杠杆点、强影响点
png("lec10_Healthy_Breakfast_influencePlot.png")
influencePlot(lm.step,id.method = "identity", main="Influence Plot",sub="Circle size is proportional to Cook's distance")
dev.off()
```

2、以“Healthy_Breakfast”的数据为例，进行多元回归分析（考虑哑变量）：

2.1、将该数据表的第2(mfr)和3(type)两列数据，转换为哑变量（Proxy and dummy variables），与4：16列重新组合成一个新的数据表；

```
table(data[,2]) #查看第2列数据种类
table(data[,3]) #查看第3列数据种类
data2<-data.frame(matrix(NA,77,20))
```

```
for(i in 1:nrow(data))
{
 #第二列mfr分类
if(data[i,2]=="A"){data2[i,1:6]<-c(0,0,0,0,0,0)}
if(data[i,2]=="G"){data2[i,1:6]<-c(1,0,0,0,0,0)}
if(data[i,2]=="K"){data2[i,1:6]<-c(0,1,0,0,0,0)}
if(data[i,2]=="N"){data2[i,1:6]<-c(0,0,1,0,0,0)}
if(data[i,2]=="P"){data2[i,1:6]<-c(0,0,0,1,0,0)}
if(data[i,2]=="Q"){data2[i,1:6]<-c(0,0,0,0,1,0)}
if(data[i,2]=="R"){data2[i,1:6]<-c(0,0,0,0,0,1)}
#第三列type分类
if(data[i,3]=="C"){data2[i,7]<-0}
if(data[i,3]=="H"){data2[i,7]<-1}
}
data2[,8:20]<-data[,4:16]
```

2.2、将原数据表的第一列数据作为新数据表的行标题，列标题作为新数据表相应列的列标题：

```
rownames(data2)<- data[,1]
colnames(data2)<-
c(paste("mfr_",c("G","K","N","P","Q","R"),sep=""),"type_CH",colnames(data)[4:16])
head(data2)
```

2.3、将所有数据列两两组合绘制散点图和拟合曲线，看看不同数据列之间的关

联：

```
png("lec10_Healthy_Breakfast_pairs2.png")
pairs(data2,panel=panel.smooth)
dev.off()
```

2.4、以“rating”数据列为因变量(y),对其他所有数据列进行多元回归分析，并查看分析结果；

```
lm0<-lm(rating~.,data=data2)
summary(lm0)
```

2.5、以【向后】逐步回归法计算最终多元回归模型（记录逐步回归的结果），并查看分析结果（summary）,并根据分析结果，写出相应的多元线性方程；

```
lm.step<-step(lm0,direction="backward")
summary(lm.step)
```

2.6、查看回归结果的统计图谱：

```
png(file = "lec10_Healthy_Breakfast_lm_data2.png")
par(mfrow=c(2,2)) #同一个图形文件中绘制2*2=4个图像
plot(lm.step)
dev.off()
```

2.7、多重共线性分析：

```
vif(lm.step)
```

2.8、检查离群点、高杠杆点、强影响点，保存屏幕反馈结果（如果有的话）和统计图：

```
png("lec10_Healthy_Breakfast_influencePlot2.png")
```

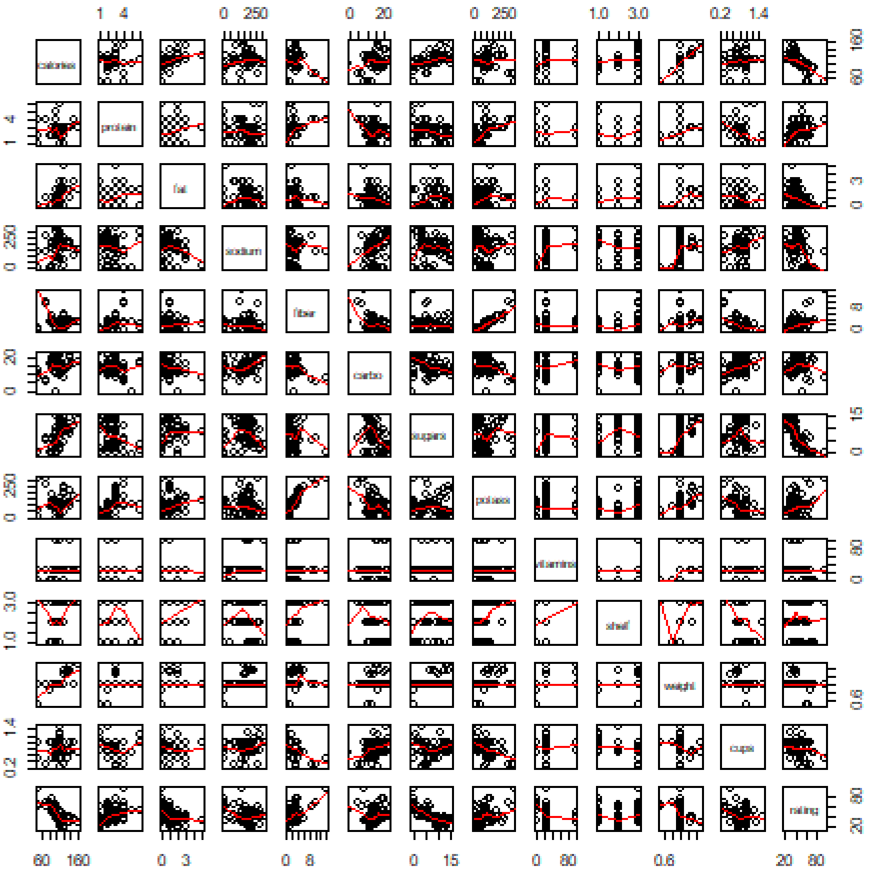
```
    influencePlot(lm.step,id.method = "identity", main="Influence
Plot",sub="Circle size is proportional to Cook's distance")
    dev.off()
```

三、实验结果：

1.1

```
> head(data)
                       name mfr type calories protein fat sodium fibe
r carbo sugars potass vitamins shelf weight cups   rating
1                 100%_Bran   N    C       70       4   1    130  10.
0   5.0      6    280       25     3      1 0.33 68.40297
2         100%_Natural_Bran   Q    C      120       3   5     15   2.
0   8.0      8    135        0     3      1 1.00 33.98368
3                  All-Bran   K    C       70       4   1    260   9.
0   7.0      5    320       25     3      1 0.33 59.42551
4 All-Bran_with_Extra_Fiber   K    C       50       4   0    140  14.
0   8.0      0    330       25     3      1 0.50 93.70491
5            Almond_Delight   R    C      110       2   2    200   1.
0  14.0      8     -1       25     3      1 0.75 34.38484
6   Apple_Cinnamon_Cheerios   G    C      110       2   2    180   1.
5  10.5     10     70       25     1      1 0.75 29.50954
> ncol(data)
[1] 16
> nrow(data)
[1] 77
```

1.2



1.3

```
> summary(lm0)

Call:
lm(formula = rating ~ ., data = data[, 4:16])

Residuals:
       Min         1Q     Median         3Q        Max
-5.243e-07 -2.577e-07  4.643e-08  2.264e-07  5.657e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  3.630e-07  1.513e+08   <2e-16 ***
calories    -2.227e-01  5.663e-09 -3.933e+07   <2e-16 ***
protein      3.273e+00  5.092e-08  6.428e+07   <2e-16 ***
fat         -1.691e+00  6.226e-08 -2.717e+07   <2e-16 ***
```

```
sodium      -5.449e-02  4.962e-10 -1.098e+08   <2e-16 ***
fiber        3.443e+00  4.309e-08  7.992e+07   <2e-16 ***
carbo        1.092e+00  1.743e-08  6.268e+07   <2e-16 ***
sugars      -7.249e-01  1.819e-08 -3.986e+07   <2e-16 ***
potass      -3.399e-02  1.473e-09 -2.307e+07   <2e-16 ***
vitamins    -5.121e-02  1.928e-09 -2.657e+07   <2e-16 ***
shelf       -3.721e-08  5.285e-08 -7.040e-01    0.484
weight      -4.298e-07  5.206e-07 -8.260e-01    0.412
cups         1.379e-07  1.924e-07  7.170e-01    0.476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1

Residual standard error: 3.044e-07 on 64 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 1.349e+16 on 12 and 64 DF,  p-value: < 2.2e-16
```

1.4

```
> lm.step<-step(lm0,direction="backward")
Start:  AIC=-2298.99
rating ~ calories + protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins + shelf + weight + cups

           Df Sum of Sq     RSS      AIC
- shelf     1      0.00    0.00 -2300.40
- cups      1      0.00    0.00 -2300.38
- weight    1      0.00    0.00 -2300.18
<none>                     0.00 -2298.99
- potass    1     49.32   49.32   -10.30
- vitamins  1     65.40   65.40    11.43
- fat       1     68.39   68.39    14.87
- calories  1    143.32  143.32    71.84
- sugars    1    147.24  147.24    73.91
- carbo     1    364.10  364.10   143.63
- protein   1    382.86  382.86   147.50
- fiber     1    591.87  591.87   181.04
- sodium    1   1117.66 1117.66   229.99
```

```
Step:  AIC=-2300.4
rating ~ calories + protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins + weight + cups

           Df Sum of Sq     RSS      AIC
- weight    1      0.00    0.00 -2301.71
- cups      1      0.00    0.00 -2301.35
<none>                     0.00 -2300.40
- potass    1     50.06   50.06   -11.15
- fat       1     71.85   71.85    16.67
- vitamins  1     78.33   78.33    23.32
- calories  1    143.54  143.54    69.95
- sugars    1    148.98  148.98    72.82
- carbo     1    376.08  376.08   144.12
- protein   1    383.76  383.76   145.68
- fiber     1    592.42  592.42   179.11
- sodium    1   1176.19 1176.19   231.92

Step:  AIC=-2301.71
rating ~ calories + protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins + cups

           Df Sum of Sq     RSS      AIC
- cups      1      0.00    0.00 -2302.35
<none>                     0.00 -2301.71
- potass    1     53.48   53.48    -8.06
- vitamins  1     79.60   79.60    22.55
- fat       1     84.96   84.96    27.57
- sugars    1    149.24  149.24    70.95
- calories  1    232.81  232.81   105.20
- carbo     1    376.15  376.15   142.14
- protein   1    383.99  383.99   143.73
- fiber     1    603.95  603.95   178.60
- sodium    1   1185.17 1185.17   230.51

Step:  AIC=-2302.35
```

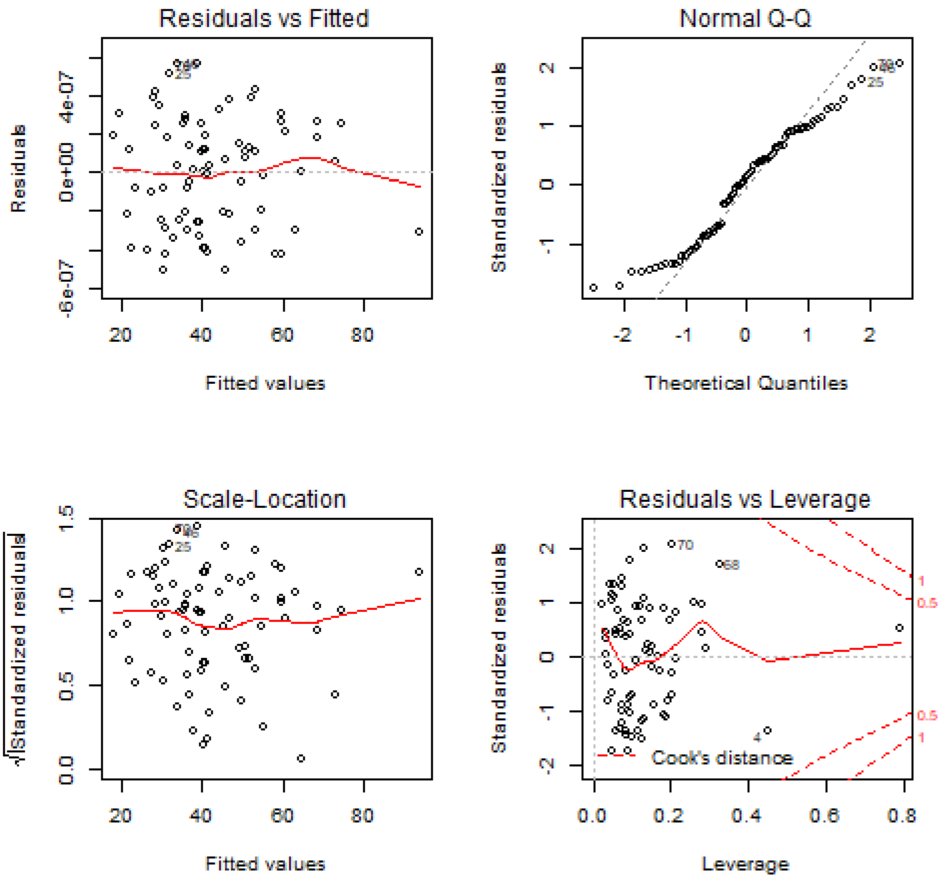
```
rating ~ calories + protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins

           Df Sum of Sq     RSS      AIC
<none>                     0.00 -2302.35
- potass    1     53.55   53.55    -9.97
- vitamins  1     80.18   80.18    21.12
- fat       1     85.00   85.00    25.61
- sugars    1    150.00  150.00    69.35
- calories  1    235.09  235.09   103.94
- carbo     1    384.29  384.29   141.79
- protein   1    386.61  386.61   142.25
- fiber     1    625.16  625.16   179.25
- sodium    1   1185.72 1185.72   228.54
Warning messages:
1: 对几乎是完全拟合再进行模型选择是没有用的
2: 对几乎是完全拟合再进行模型选择是没有用的
3: 对几乎是完全拟合再进行模型选择是没有用的
4: 对几乎是完全拟合再进行模型选择是没有用的
> summary(lm.step)

Call:
lm(formula = rating ~ calories + protein + fat + sodium + fiber +
    carbo + sugars + potass + vitamins, data = data[, 4:16])

Residuals:
       Min         1Q     Median         3Q        Max
-5.111e-07 -2.547e-07  3.200e-08  2.440e-07  5.676e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.550e-07  215364561   <2e-16 ***
calories    -2.227e-01  4.396e-09  -50659136   <2e-16 ***
protein      3.273e+00  5.038e-08   64965144   <2e-16 ***
fat         -1.691e+00  5.553e-08  -30461632   <2e-16 ***
sodium      -5.449e-02  4.790e-10 -113771893   <2e-16 ***
fiber        3.443e+00  4.168e-08   82611577   <2e-16 ***
```

```
carbo         1.092e+00  1.687e-08   64770285   <2e-16 ***
sugars       -7.249e-01  1.791e-08  -40465741   <2e-16 ***
potass       -3.399e-02  1.406e-09  -24177294   <2e-16 ***
vitamins     -5.121e-02  1.731e-09  -29585288   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1

Residual standard error: 3.027e-07 on 67 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 1.819e+16 on 9 and 67 DF,  p-value: < 2.2e-16
```

回归方程：

rating = -2.227e-01*calories + 3.273e+00*protein-1.691e+00*fat-5.449e-02*sodium +3.443e+00*fiber + 1.092e+00*carbo-7.249e-01*sugars-3.399e-02*potass-5.121e-02*vitamins

1.5

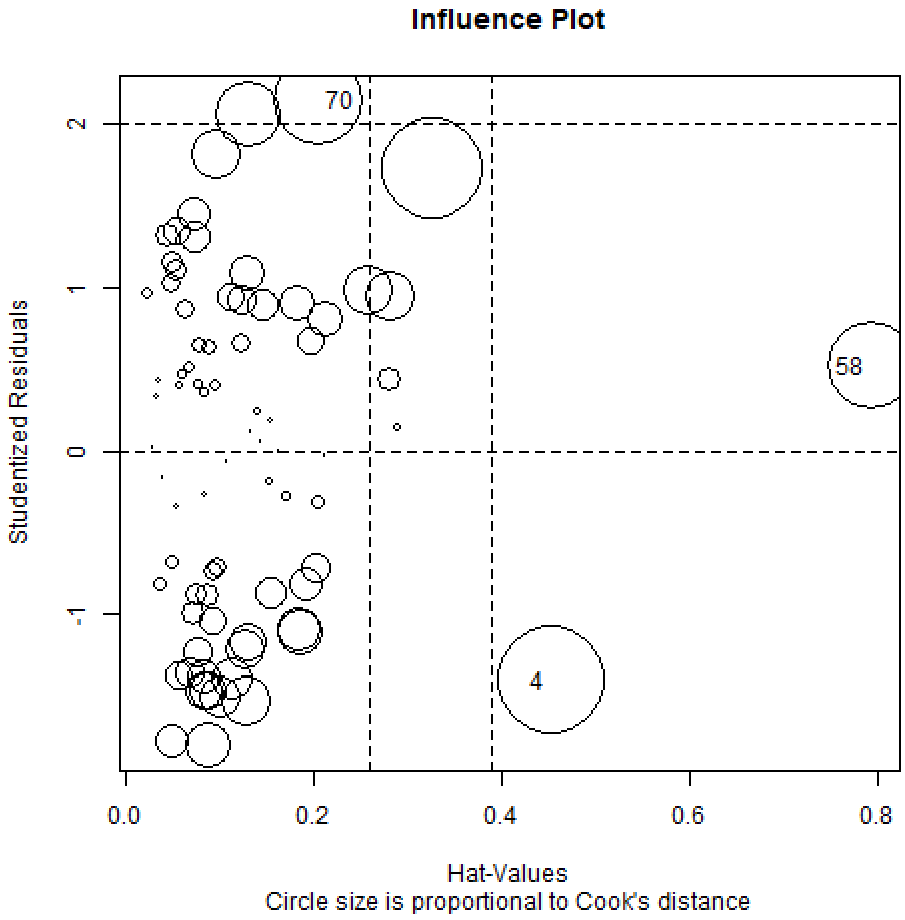


1.6

```
> vif(lm.step)
calories  protein      fat   sodium    fiber    carbo   sugars   potass vitamins
6.088104 2.524295 2.591169 1.337614 8.188331 4.321449 5.260148 8.334746 1.240958

>
```

1.7

```
> influencePlot(lm.step,id.method = "identity", main="Influence Plot
",sub="Circle size is proportional to Cook’s distance")
      StudRes       Hat     CookD
4  -1.3958455 0.4532860 0.1592878
58  0.5258852 0.7930972 0.1071660
70  2.1396958 0.2049968 0.1120691
```



2.1

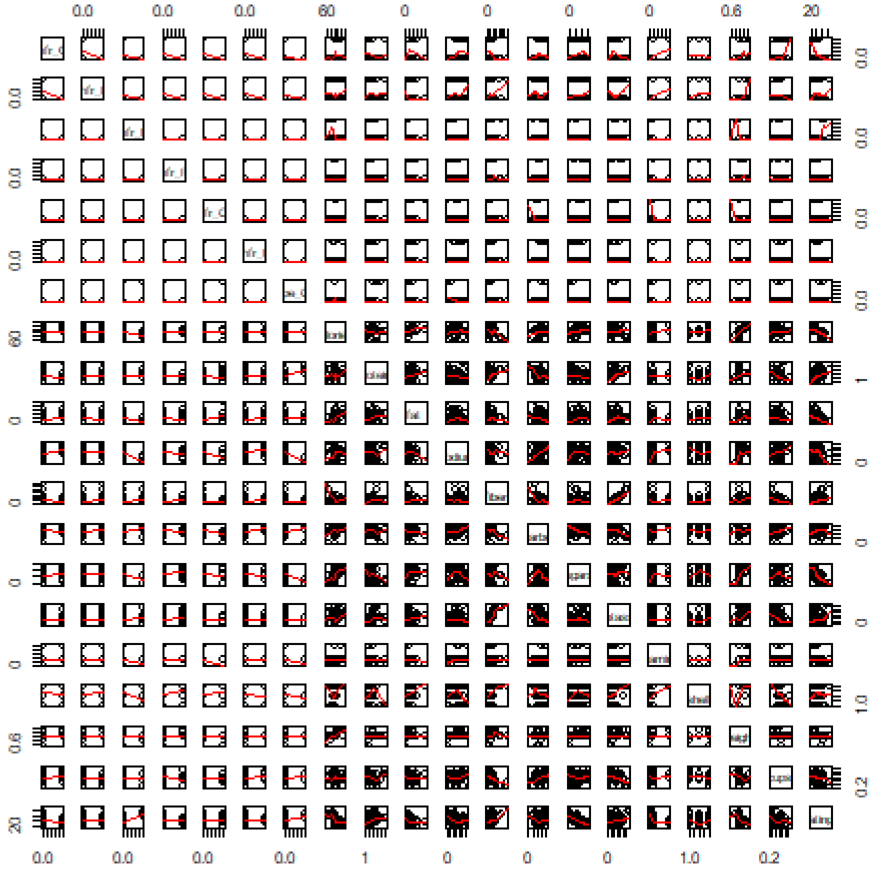
```
> table(data[,2]) #查看第2列数据种类

 A  G  K  N  P  Q  R 
 1 22 23  6  9  8  8 
> table(data[,3]) #查看第3列数据种类

 C  H 
74  3 
```

2.2

```
> head(data2)
                          mfr_G mfr_K mfr_N mfr_P mfr_Q mfr_R type_CH
 calories
100%_Bran                     0     0     1     0     0     0       0
       70
100%_Natural_Bran             0     0     0     0     1     0       0
      120
All-Bran                      0     1     0     0     0     0       0
       70
All-Bran_with_Extra_Fiber     0     1     0     0     0     0       0
       50
Almond_Delight                0     0     0     0     0     1       0
      110
Apple_Cinnamon_Cheerios       1     0     0     0     0     0       0
      110
                          protein fat sodium fiber carbo sugars potas
s vitamins
100%_Bran                       4   1    130  10.0   5.0      6    28
0       25
100%_Natural_Bran               3   5     15   2.0   8.0      8    13
5        0
All-Bran                        4   1    260   9.0   7.0      5    32
0       25
All-Bran_with_Extra_Fiber       4   0    140  14.0   8.0      0    33
0       25
Almond_Delight                  2   2    200   1.0  14.0      8     -
1       25
```

```
Apple_Cinnamon_Cheerios           2    2     180    1.5   10.5       10       7
0        25
                          shelf weight cups   rating
100%_Bran                     3        1 0.33 68.40297
100%_Natural_Bran             3        1 1.00 33.98368
All-Bran                      3        1 0.33 59.42551
All-Bran_with_Extra_Fiber     3        1 0.50 93.70491
Almond_Delight                3        1 0.75 34.38484
Apple_Cinnamon_Cheerios       1        1 0.75 29.50954
```

2.3



2.4

```
> lm0<-lm(rating~.,data=data2)
> summary(lm0)

Call:
```

```
lm(formula = rating ~ ., data = data2)

Residuals:
       Min         1Q     Median         3Q        Max
-4.941e-07 -2.017e-07 -8.050e-09  2.243e-07  5.396e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  5.693e-07  9.648e+07   <2e-16 ***
mfr_G        3.315e-07  4.786e-07  6.930e-01    0.491
mfr_K        3.446e-07  4.927e-07  6.990e-01    0.487
mfr_N        4.411e-07  4.664e-07  9.460e-01    0.348
mfr_P        5.655e-07  4.972e-07  1.137e+00    0.260
mfr_Q        2.405e-07  4.812e-07  5.000e-01    0.619
mfr_R        3.778e-07  4.872e-07  7.750e-01    0.441
type_CH      9.659e-08  3.382e-07  2.860e-01    0.776
calories    -2.227e-01  6.630e-09 -3.359e+07   <2e-16 ***
protein      3.273e+00  5.302e-08  6.174e+07   <2e-16 ***
fat         -1.691e+00  7.439e-08 -2.274e+07   <2e-16 ***
sodium      -5.449e-02  6.439e-10 -8.463e+07   <2e-16 ***
fiber        3.443e+00  5.634e-08  6.112e+07   <2e-16 ***
carbo        1.092e+00  2.513e-08  4.346e+07   <2e-16 ***
sugars      -7.249e-01  2.346e-08 -3.090e+07   <2e-16 ***
potass      -3.399e-02  1.689e-09 -2.013e+07   <2e-16 ***
vitamins    -5.121e-02  2.027e-09 -2.527e+07   <2e-16 ***
shelf       -3.249e-08  5.835e-08 -5.570e-01    0.580
weight      -3.330e-07  6.149e-07 -5.420e-01    0.590
cups         1.928e-07  2.026e-07  9.520e-01    0.345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1

Residual standard error: 3.091e-07 on 57 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 8.26e+15 on 19 and 57 DF,  p-value: < 2.2e-16
```

2.5

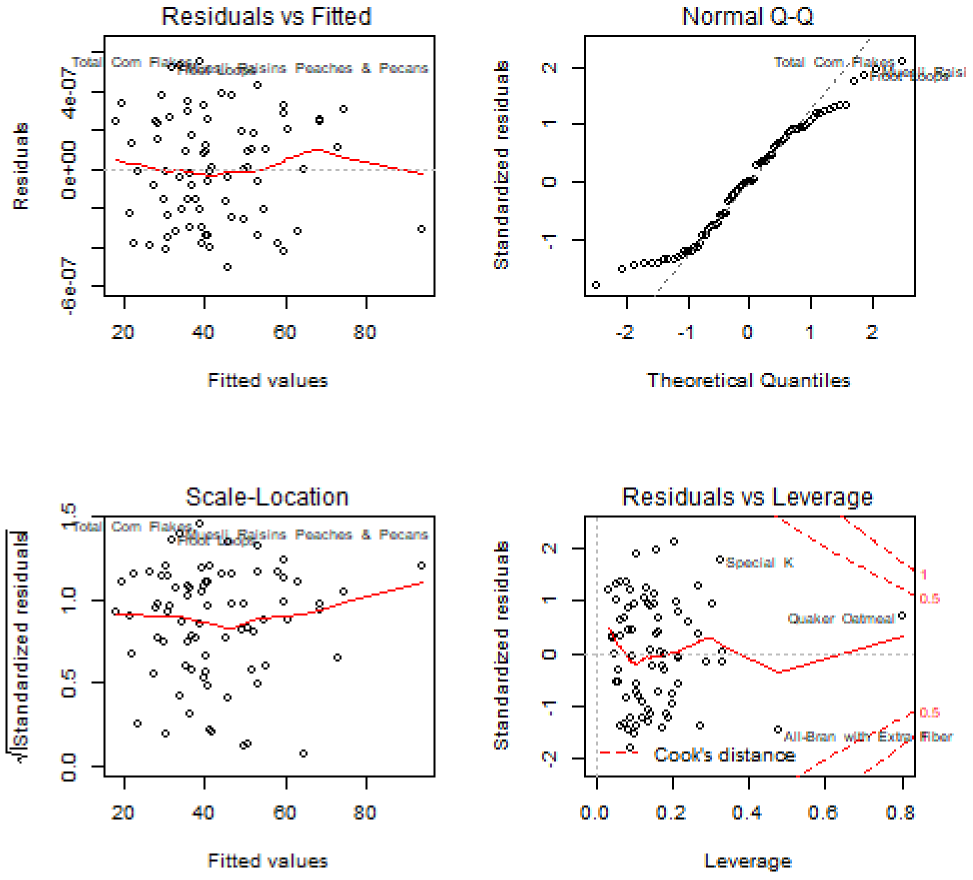
```
> lm.step<-step(lm0,direction="backward")
Start:  AIC=-2291.54
rating ~ mfr_G + mfr_K + mfr_N + mfr_P + mfr_Q + mfr_R + type_CH +
    calories + protein + fat + sodium + fiber + carbo + sugars +
    potass + vitamins + shelf + weight + cups

           Df Sum of Sq    RSS      AIC
- type_CH   1      0.00   0.00 -2293.43
- mfr_Q     1      0.00   0.00 -2293.20
- weight    1      0.00   0.00 -2293.15
- shelf     1      0.00   0.00 -2293.12
- mfr_G     1      0.00   0.00 -2292.89
- mfr_K     1      0.00   0.00 -2292.88
- mfr_R     1      0.00   0.00 -2292.73
- mfr_N     1      0.00   0.00 -2292.34
- cups      1      0.00   0.00 -2292.33
- mfr_P     1      0.00   0.00 -2291.81
<none>                    0.00 -2291.54
- potass    1     38.71  38.71   -14.95
- fat       1     49.41  49.41     3.83
- vitamins  1     61.02  61.02    20.09
- sugars    1     91.27  91.27    51.09
- calories  1    107.84 107.84    63.94
- carbo     1    180.54 180.54   103.61
- fiber     1    357.02 357.02   156.12
- protein   1    364.24 364.24   157.66
- sodium    1    684.43 684.43   206.23

Step:  AIC=-2293.43
rating ~ mfr_G + mfr_K + mfr_N + mfr_P + mfr_Q + mfr_R + calories +
    protein + fat + sodium + fiber + carbo + sugars + potass +
    vitamins + shelf + weight + cups

           Df Sum of Sq    RSS      AIC
- mfr_Q     1      0.00   0.00 -2295.20
- weight    1      0.00   0.00 -2295.13
- shelf     1      0.00   0.00 -2295.07
- mfr_G     1      0.00   0.00 -2294.83
- mfr_K     1      0.00   0.00 -2294.82
- mfr_R     1      0.00   0.00 -2294.67
- mfr_N     1      0.00   0.00 -2294.22
- cups      1      0.00   0.00 -2294.04
<none>                   0.00 -2293.43
- mfr_P     1      0.00   0.00 -2293.37
- potass    1     40.57  40.57   -13.34
- fat       1     54.79  54.79     9.80
- vitamins  1     62.28  62.28    19.66
- calories  1    118.07 118.07    68.92
- sugars    1    136.88 136.88    80.30
- carbo     1    286.40 286.40   137.15
- fiber     1    357.64 357.64   154.25
- protein   1    364.38 364.38   155.69
- sodium    1    684.58 684.58   204.25

Step:  AIC=-2295.2
rating ~ mfr_G + mfr_K + mfr_N + mfr_P + mfr_R + calories + protein +

    fat + sodium + fiber + carbo + sugars + potass + vitamins +
    shelf + weight + cups

           Df Sum of Sq    RSS      AIC
- shelf     1      0.00   0.00 -2296.90
- weight    1      0.00   0.00 -2296.81
- mfr_K     1      0.00   0.00 -2296.63
- mfr_G     1      0.00   0.00 -2296.55
- mfr_R     1      0.00   0.00 -2296.40
- cups      1      0.00   0.00 -2295.84
- mfr_N     1      0.00   0.00 -2295.81
<none>                   0.00 -2295.20
- mfr_P     1      0.00   0.00 -2292.90
- potass    1     41.63  41.63   -13.35
- fat       1     54.92  54.92     7.98
- vitamins  1     62.50  62.50    17.94
- calories  1    122.08 122.08    69.49
- sugars    1    137.95 137.95    78.90
- carbo     1    296.14 296.14   137.72
- fiber     1    370.75 370.75   155.02
- protein   1    375.50 375.50   156.00
- sodium    1    730.46 730.46   207.24

Step:  AIC=-2296.9
rating ~ mfr_G + mfr_K + mfr_N + mfr_P + mfr_R + calories + protein +

    fat + sodium + fiber + carbo + sugars + potass + vitamins +
    weight + cups

           Df Sum of Sq    RSS      AIC
- weight    1      0.00   0.00 -2298.58
- mfr_K     1      0.00   0.00 -2298.25
- mfr_G     1      0.00   0.00 -2298.11
- mfr_R     1      0.00   0.00 -2297.88
- cups      1      0.00   0.00 -2297.07
- mfr_N     1      0.00   0.00 -2297.01
<none>                    0.00 -2296.90
- mfr_P     1      0.00   0.00 -2294.46
- potass    1     41.94  41.94   -14.78
- fat       1     57.35  57.35     9.31
- vitamins  1     71.96  71.96    26.79
- calories  1    122.16 122.16    67.54
- sugars    1    140.11 140.11    78.09
- carbo     1    321.45 321.45   142.04
- fiber     1    375.35 375.35   153.97
- protein   1    377.58 377.58   154.43
- sodium    1    786.36 786.36   210.92

Step:  AIC=-2298.58
rating ~ mfr_G + mfr_K + mfr_N + mfr_P + mfr_R + calories + protein +

    fat + sodium + fiber + carbo + sugars + potass + vitamins +
    cups

           Df Sum of Sq    RSS      AIC
- mfr_K     1      0.00   0.00 -2299.80
- mfr_G     1      0.00   0.00 -2299.76
- mfr_R     1      0.00   0.00 -2299.05
- mfr_N     1      0.00   0.00 -2298.66
<none>                    0.00 -2298.58
- cups      1      0.00   0.00 -2298.39
- mfr_P     1      0.00   0.00 -2295.74
- potass    1     43.41  43.41   -14.14
- fat       1     68.43  68.43    20.91
- vitamins  1     72.41  72.41    25.27
- sugars    1    140.95 140.95    76.55
- calories  1    215.59 215.59   109.28
- carbo     1    323.81 323.81   140.60
- protein   1    377.85 377.85   152.48
- fiber     1    392.99 392.99   155.51
- sodium    1    796.29 796.29   209.88

Step:  AIC=-2299.8
rating ~ mfr_G + mfr_N + mfr_P + mfr_R + calories + protein + 
    fat + sodium + fiber + carbo + sugars + potass + vitamins + 
    cups

           Df Sum of Sq    RSS      AIC
- mfr_G     1      0.00   0.00 -2301.66
- mfr_R     1      0.00   0.00 -2301.05
- mfr_N     1      0.00   0.00 -2300.66
- cups      1      0.00   0.00 -2299.89
<none>                    0.00 -2299.80
- mfr_P     1      0.00   0.00 -2296.33
- potass    1     48.83  48.83    -7.08
- vitamins  1     72.48  72.48    23.34
- fat       1     76.31  76.31    27.31
- sugars    1    148.55 148.55    78.60
- calories  1    223.32 223.32   109.99
- carbo     1    345.83 345.83   143.67
- protein   1    378.71 378.71   150.66
- fiber     1    494.75 494.75   171.24
- sodium    1    797.48 797.48   208.00

Step:  AIC=-2301.66
rating ~ mfr_N + mfr_P + mfr_R + calories + protein + fat + sodium + 
    fiber + carbo + sugars + potass + vitamins + cups

           Df Sum of Sq    RSS     AIC
- mfr_R     1      0.00   0.00 -2303.04
- mfr_N     1      0.00   0.00 -2302.43
- cups      1      0.00   0.00 -2301.70
<none>                    0.00 -2301.66
- mfr_P     1      0.00   0.00 -2298.31
- potass    1     52.65  52.65    -3.27
- vitamins  1     73.40  73.40    22.31
- fat       1     81.63  81.63    30.49
- sugars    1    148.91 148.91    76.78
- calories  1    230.16 230.16   110.31
- carbo     1    345.86 345.86   141.67
- protein   1    380.92 380.92   149.11
- fiber     1    553.01 553.01   177.81
- sodium    1    915.80 915.80   216.65

Step:  AIC=-2303.04
rating ~ mfr_N + mfr_P + calories + protein + fat + sodium + 
    fiber + carbo + sugars + potass + vitamins + cups

           Df Sum of Sq    RSS     AIC
- mfr_N     1      0.00   0.00 -2304.06
<none>                    0.00 -2303.04
- cups      1      0.00   0.00 -2303.01
- mfr_P     1      0.00   0.00 -2300.07
- potass    1     52.83  52.83    -5.01
- vitamins  1     76.35  76.35    23.34
- fat       1     82.55  82.55    29.36
- sugars     1    148.91  148.91    74.78
- calories   1    230.30  230.30   108.36
- carbo      1    359.65  359.65   142.68
- protein    1    381.75  381.75   147.27
- fiber      1    562.39  562.39   177.11
- sodium     1    915.85  915.85   214.66

Step:  AIC=-2304.06
rating ~ mfr_P + calories + protein + fat + sodium + fiber + 
    carbo + sugars + potass + vitamins + cups

           Df Sum of Sq     RSS      AIC
<none>                     0.00 -2304.06
- cups      1      0.00    0.00 -2303.99
- mfr_P     1      0.00    0.00 -2301.71
- potass    1     53.48   53.48    -6.07
- vitamins  1     79.41   79.41    24.37
- fat       1     83.23   83.23    27.99
- sugars    1    149.20  149.20    72.93
- calories  1    230.56  230.56   106.45
- carbo     1    374.49  374.49   143.80
- protein   1    383.04  383.04   145.53
- fiber     1    603.05  603.05   180.48
- sodium    1   1182.89 1182.89   232.36
Warning messages:
1: 对几乎是完全拟合再进行模型选择是没有用的
2: 对几乎是完全拟合再进行模型选择是没有用的
3: 对几乎是完全拟合再进行模型选择是没有用的
4: 对几乎是完全拟合再进行模型选择是没有用的
5: 对几乎是完全拟合再进行模型选择是没有用的
6: 对几乎是完全拟合再进行模型选择是没有用的
7: 对几乎是完全拟合再进行模型选择是没有用的
8: 对几乎是完全拟合再进行模型选择是没有用的
9: 对几乎是完全拟合再进行模型选择是没有用的
> summary(lm.step)

Call:
```

```
lm(formula = rating ~ mfr_P + calories + protein + fat + sodium +
    fiber + carbo + sugars + potass + vitamins + cups, data = data2)

Residuals:
       Min         1Q     Median         3Q        Max
-5.118e-07 -2.412e-07 -4.000e-09  2.400e-07  5.571e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.934e-07  1.872e+08   <2e-16 ***
mfr_P        2.128e-07  1.095e-07  1.943e+00   0.0563 .
calories    -2.227e-01  4.343e-09 -5.128e+07   <2e-16 ***
protein        4.952e-08  6.610e+07   <2e-16 ***
fat         -1.691e+00  5.489e-08 -3.081e+07   <2e-16 ***
sodium        4.691e-10 -1.162e+08   <2e-16 ***
fiber        3.443e+00  4.152e-08  8.294e+07   <2e-16 ***
carbo        1.092e+00  1.672e-08  6.536e+07   <2e-16 ***
sugars      -7.249e-01  1.757e-08 -4.125e+07   <2e-16 ***
potass      -3.399e-02  1.376e-09 -2.470e+07   <2e-16 ***
vitamins    -5.121e-02  1.702e-09 -3.010e+07   <2e-16 ***
cups         2.383e-07  1.791e-07  1.330e+00   0.1880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1

Residual standard error: 2.961e-07 on 65 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 1.555e+16 on 11 and 65 DF,  p-value: < 2.2e-16
```

回归方程:

Rating = 2.128e-07*mfr_P -2.227e-01*calories + 3.273e+00*protein -1.691e+00 *fat +4.691e-10*sodium + 3.443e+00 *fiber + 1.092e+00*carbo-7.249e-01*sugars-3.399e-02*potass -5.121e-02*vitamins+2.383e-07*cups

2.6

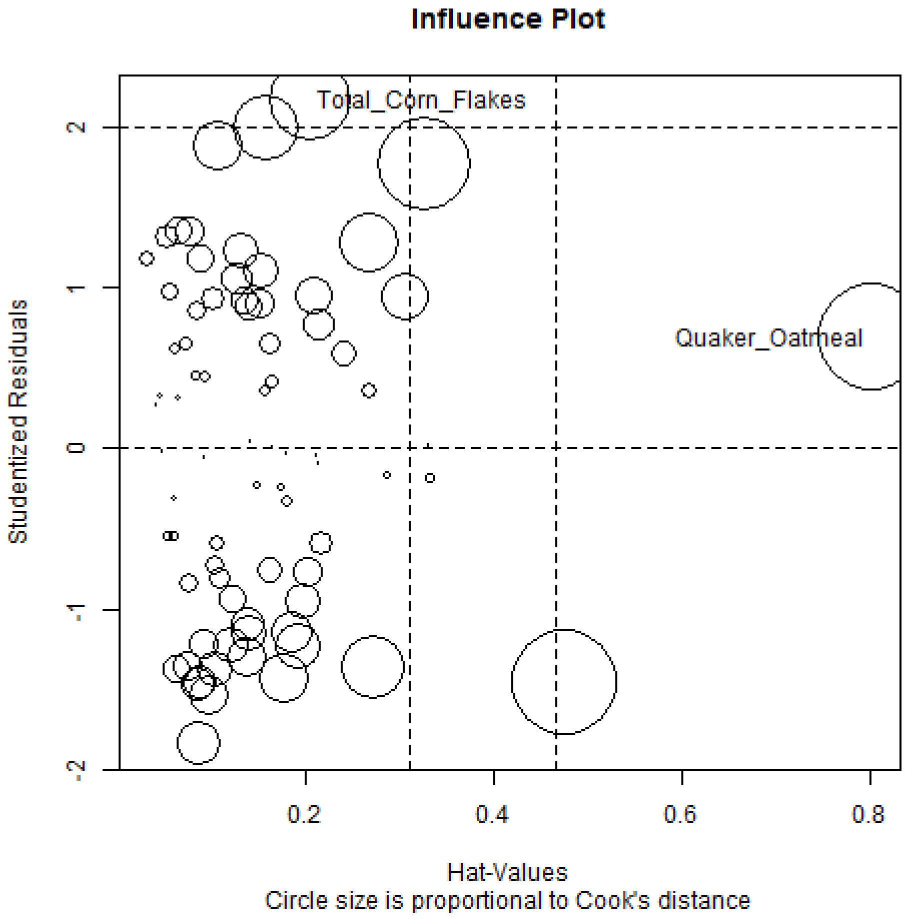


2.7

```
> vif(lm.step)
   mfr_P calories  protein      fat   sodium    fiber    carbo    suga
rs   potass
1.086704 6.207718 2.547822 2.646163 1.340810 8.488575 4.434566 5.2883
00 8.344903
vitamins     cups
1.252998 1.506106
```

2.8

```
> influencePlot(lm.step,id.method = "identity", main="Influence Plot
",sub="Circle size is proportional to Cook's distance")
                     StudRes       Hat      CookD
Quaker_Oatmeal     0.6908905 0.8024932 0.16293063
Total_Corn_Flakes 2.1704245 0.2057242 0.09618579
```



四 ．对哑变量设定前后的多元回归结果进行比较、分析和讨论。

在多元回归分析中，有些变量不能直接把相关数值放到模型中分析，实际上仅仅表示一种分类，此时设置哑变量会被当成数值型变量计算，提高线性回归的准确性。引入哑变量可使线形回归模型变得更复杂，但对问题描述更简明，一个方程能达到俩个方程的作用，而且接近现实。本例中的回归方程在增加哑变量后新增了两项，mfr_P和cups，更加接近现实。