Protein and fat have an impact on food calories

Wen Han 7879607

November 25, 2020

Question

As we all know that protein and fat play an important role in food calories. This project is to statistically reveal how these two ingredients impact the food calories.

With in this project, the explanatory variable Protein is the protein content(in grams) contains in one kind of food, the explanatory variable Fat is the fat content(in grams) contains in this kind of food, and the response variable Calories is the calorie content (in calories) in this kind of food.

Protein and fat have positive linear relationship with total food calories. In addition, the amount of these two ingredients can help to predict how much calories for specific kind of food contains-the more protein/fat it has, the more calories it provides.

Data Set

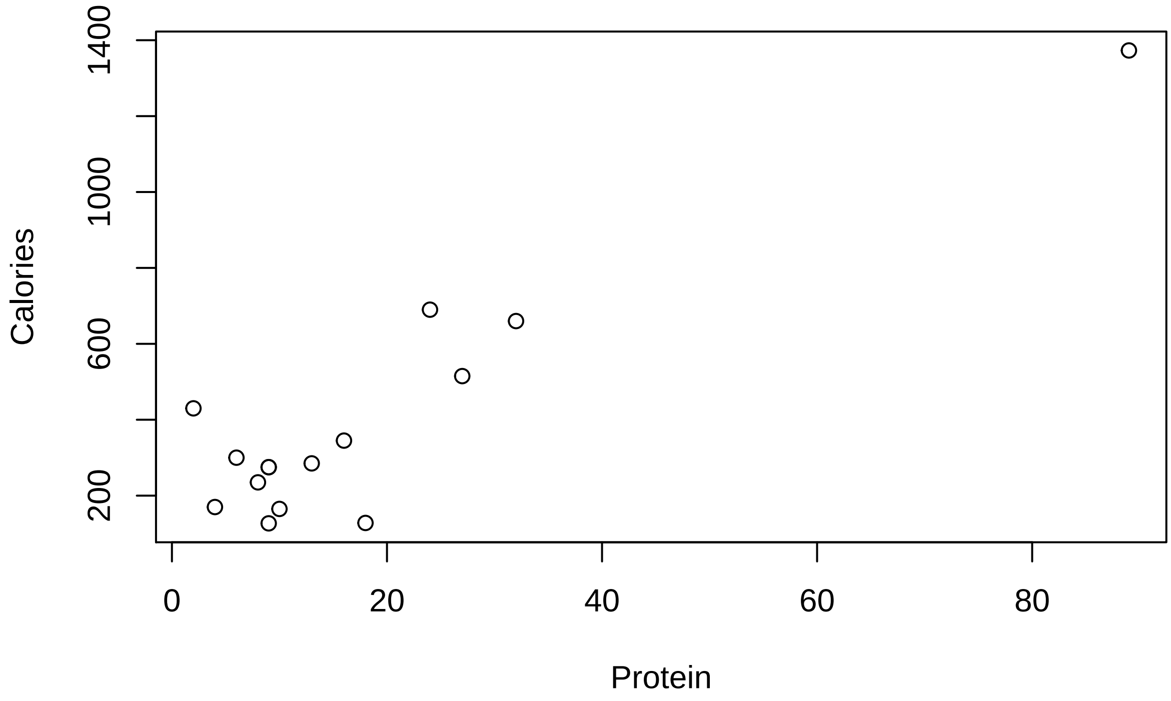
```
FoodCalories<-read.csv("nutrients_csvfile.csv")
Calories<-c(660,127,345,1373,515,165,690,235,128,275,285,300,275,170,430)
Protein<-c(32,9,16,89,27,10,24,8,18,9,13,6,9,4,2)
Fat<-c(40,5,20,42,28,8,24,11,4,10,14,18,10,15,44)
Calorie<-data.frame(Calories, Protein, Fat)
knitr::kable(Calorie, "pipe", col.name=c("Calorie", "Protein", "Fat"), align=c("l", "c","c"))
```

Calorie	Protein	Fat
660	32	40
127	9	5
345	16	20
1373	89	42
515	27	28
165	10	8
690	24	24
235	8	11
128	18	4
275	9	10
285	13	14
300	6	18
275	9	10
170	4	15
430	2	44

Pandit, N.(July 25, 2020). Nutritional Facts for most common foods, know the nutrients in your food: Fat, Carbs, Proteins etc. https://www.kaggle.com/niharika41298/nutrition-details-for-most-common-foods

Following is a scatterplot of Protein respect Calories and calculated r^2.

```
plot(y=Calories, x=Protein, xlab="Protein", ylab="Calories")
```



```
model<-lm(Calories~Protein)
summary(model)
```

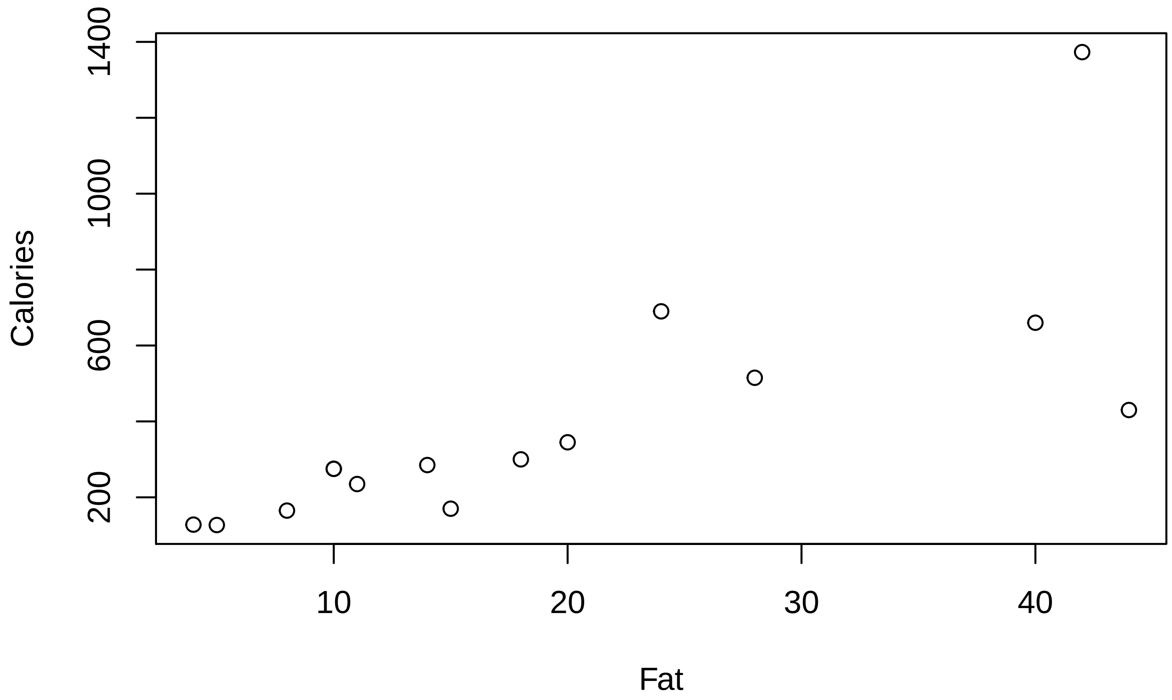
```
##
## Call:
## lm(formula = Calories ~ Protein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -264.663  -33.650   -2.545   40.193  258.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  143.481     46.071   3.114  0.00822 **
## Protein       13.843      1.667   8.306 1.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 133.2 on 13 degrees of freedom
## Multiple R-squared:  0.8414, Adjusted R-squared:  0.8293
## F-statistic: 68.99 on 1 and 13 DF,  p-value: 1.479e-06
```

-Here, the r^2 is 0.8414.

Following is a scatterplot of Fat respect Calories and calculated r^2.

```
plot(y=Calories, x=Fat, xlab="Fat", ylab="Calories")
```



```
model<-lm(Calories~Fat)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -423.69  -65.77  -10.19   36.63  556.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.556     99.209   0.348 0.733178
## Fat           18.617      4.235   4.396 0.000723 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 212.1 on 13 degrees of freedom
## Multiple R-squared:  0.5978, Adjusted R-squared:  0.5669
## F-statistic: 19.32 on 1 and 13 DF,  p-value: 0.0007234
```

-Here, the r^2 is 0.5978.

Preliminary Model

Following is the model for Protein~Calories, also the regression line.

```
model<-lm(Calories~Protein)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Protein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -264.663  -33.650   -2.545   40.193  258.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  143.481     46.071   3.114  0.00822 **
## Protein       13.843      1.667   8.306 1.48e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 133.2 on 13 degrees of freedom
## Multiple R-squared:  0.8414, Adjusted R-squared:  0.8293
## F-statistic: 68.99 on 1 and 13 DF,  p-value: 1.479e-06
```

$\hat{y}=143.481+13.843X_1$

Following is the model for Fat~Calories, also the regression line.

```
model<-lm(Calories~Fat)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -423.69  -65.77  -10.19   36.63  556.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.556     99.209   0.348 0.733178
## Fat           18.617      4.235   4.396 0.000723 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 212.1 on 13 degrees of freedom
## Multiple R-squared:  0.5978, Adjusted R-squared:  0.5669
## F-statistic: 19.32 on 1 and 13 DF,  p-value: 0.0007234
```

$\hat{y}$=34.556+18.617X_2

Following is the model for Protein and Fat~Calories, also the regression line.

```
model<-lm(Calories~Protein+Fat)
summary(model)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.64  -35.53  -22.46   37.75  190.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.593     37.349   0.525 0.609414
## Protein       10.646      1.191   8.939 1.19e-06 ***
## Fat            9.354      1.900   4.923 0.000352 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 79.76 on 12 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9387
## F-statistic: 108.3 on 2 and 12 DF,  p-value: 2.097e-08
```

$\hat{y}$=19.593+10.646X_1+9.354X_2.

The adjusted r^2 is increased from 0.5669 to 0.9387

Following is the full second-order model, also the regression line.

full second-order model

```
protein2<-Protein^2
fat2<-Fat^2
Model.full<-lm(Calories~Protein+Fat+protein2+fat2+Protein*Fat)
Model.full
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat + protein2 + fat2 + Protein *
##     Fat)
##
## Coefficients:
## (Intercept)      Protein          Fat     protein2         fat2  Protein:Fat
##   -80.00330      8.77592     23.98250      0.04907     -0.29423     -0.05502
```

$\hat{y}$=-80.00330 + 8.77592X_1 + 23.98250X_1 + 0.0497X_1^2 -0.29423X_2^2 -0.05502X_1X_2

Here is the complete ANOVA test, it needs to be identified if at least one of the model terms is significant.

ANOVA test

```
anova(Model.full)
```

```
## Analysis of Variance Table
##
## Response: Calories
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Protein      1 1223353 1223353 200.6276 1.854e-07 ***
## Fat          1  154169  154169  25.2834  0.000711 ***
## protein2     1     780     780   0.1279  0.728876
## fat2         1   20358   20358   3.3387  0.100943
## Protein:Fat  1     329     329   0.0540  0.821415
## Residuals    9   54879    6098
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(Model.full)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat + protein2 + fat2 + Protein *
##     Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.961 -39.287  -4.832  16.128 156.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80.00330  106.32416  -0.752   0.4710
## Protein       8.77592    6.76201   1.298   0.2266
## Fat          23.98250    8.70372   2.755   0.0223 *
## protein2      0.04907    0.06092   0.806   0.4412
## fat2         -0.29423    0.16019  -1.837   0.0994 .
## Protein:Fat  -0.05502    0.23675  -0.232   0.8214
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 78.09 on 9 degrees of freedom
## Multiple R-squared:  0.9623, Adjusted R-squared:  0.9413
## F-statistic: 45.89 on 5 and 9 DF,  p-value: 3.894e-06
```

(1)Level of significance: α =0.05

(2)Hypothesis: H_0:β_1=β_2=β_3=β_4=β_5=0 vs. H_A: At least one $\beta_i \neq 0$ (i=1,2,3,4,5)

(3)Decision Rule: Reject H_0 if p-value $\leq \alpha$

(4)Test statistic: F=45.89

(5)P-value: ≈ 0

(6)Conclusion: As p-value $\approx 0 < 0.5= \alpha$, reject H_0. Conclude that there is sufficient evidence that at least one of the model terms does a sufficient job at explaining the protein and fat have an impact with food calories.

Model Refinement

Here the summary() function is used on the full model to get the output for the t-tests on the individual co-efficients.

```
output<-lm(formula = Calories~Protein+Fat+protein2+fat2+Protein*Fat)
summary(output)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat + protein2 + fat2 + Protein *
##     Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.961 -39.287  -4.832  16.128 156.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80.00330  106.32416  -0.752   0.4710
## Protein       8.77592    6.76201   1.298   0.2266
## Fat          23.98250    8.70372   2.755   0.0223 *
## protein2      0.04907    0.06092   0.806   0.4412
## fat2         -0.29423    0.16019  -1.837   0.0994 .
## Protein:Fat  -0.05502    0.23675  -0.232   0.8214
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 78.09 on 9 degrees of freedom
## Multiple R-squared:  0.9623, Adjusted R-squared:  0.9413
## F-statistic: 45.89 on 5 and 9 DF,  p-value: 3.894e-06
```

β_1=8.77592 β_2=23.98250 β_3=0.04907 β_4=-0.29423 β_5=-0.05502

The Fat term is 0.0223 which is less than 0.05 seems significant. But there are other terms are not significant. VIF function is used to check which other term (or terms) is(or are) significant.

VIF

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(Model.full)
```

```
##     Protein         Fat    protein2        fat2 Protein:Fat 
##    47.86924    31.15515    34.16092    26.27430   118.45618
```

All the VIF numbers is bigger than 5. So, it has to reduce the largest VIF, which is Protein*Fat, and redo the summary() without that, propose a new model based on those results.

```r
reduced.model<-lm(formula = Calories~Protein+Fat+protein2+fat2)
summary(reduced.model)
```

```
## 
## Call:
## lm(formula = Calories ~ Protein + Fat + protein2 + fat2)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -79.95 -34.19  -6.44  16.66 162.03 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -61.8267    68.5388  -0.902   0.3882  
## Protein       7.4803     3.6416   2.054   0.0670 .
## Fat          23.0504     7.3502   3.136   0.0106 *
## protein2      0.0386     0.0390   0.990   0.3457  
## fat2         -0.2868     0.1493  -1.920   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 74.3 on 10 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.9468 
## F-statistic: 63.34 on 4 and 10 DF,  p-value: 4.587e-07
```

From this model, there is no other significant term since p-value all bigger than α. So, reducing Protein^2 term, redo the summary() without that, and propose another new model based on those results.

```r
reduced.model2<-lm(formula = Calories~Protein+Fat+fat2)
summary(reduced.model2)
```

```
## 
## Call:
## lm(formula = Calories ~ Protein + Fat + fat2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -96.502 -57.386   7.921  30.932 144.634
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -76.7709    66.7923  -1.149   0.2748    
## Protein      10.9095     1.1193   9.747 9.55e-07 ***
## Fat          20.7224     6.9572   2.979   0.0126 *  
## fat2         -0.2379     0.1408  -1.690   0.1192    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 74.23 on 11 degrees of freedom
## Multiple R-squared:  0.9583, Adjusted R-squared:  0.9469 
## F-statistic: 84.28 on 3 and 11 DF,  p-value: 7.13e-08
```

From this model, there is still other non significant terms. So, reducing Fatˆ2 term, redo the summary() without that, and propose another new model based on those results.

```
reduced.model3<-lm(formula = Calories~Protein+Fat)
summary(reduced.model3)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -120.64  -35.53  -22.46   37.75  190.40 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   19.593     37.349   0.525 0.609414    
## Protein       10.646      1.191   8.939 1.19e-06 ***
## Fat            9.354      1.900   4.923 0.000352 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 79.76 on 12 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9387 
## F-statistic: 108.3 on 2 and 12 DF,  p-value: 2.097e-08
```

Now, a nested F-test was perform to test that the terms that eliminated were in fact zero comparing the full model to my new reduced model. And also following with the full test.

Anova test

```
anova(reduced.model3, Model.full)
```

```
## Analysis of Variance Table
##
## Model 1: Calories ~ Protein + Fat
## Model 2: Calories ~ Protein + Fat + protein2 + fat2 + Protein * Fat
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     12 76346
## 2      9 54879  3     21467 1.1735 0.3728
```

(1)Level of significance: α =0.05

(2)Hypothesis: H_0: β_1=β2=0 vs. H_A: At least one $\beta_i \neq 0$ (i=1,2)

(3)Decision Rule: Reject H_0 if p-value $\leq \alpha$

(4)Test statistic: F=1.1735

(5)P-value: p-value $\approx$ 0.3728

(6)Conclusion: As p-value $\approx$ 0.3728> 0.5= α, fail to reject H_0. Conclude that there is insufficient evidence that at least one of co-efficients for $(protein)^2$ and $(fat)^2$ is non-zero.

Final Model and Assessment

The ANOVA test on the reduced model was perform to show it adequately explains the relationship with Y.

Anova test

```
anova(reduced.model3)
```

```
## Analysis of Variance Table
##
## Response: Calories
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Protein    1 1223353 1223353 192.286  9.51e-09 ***
## Fat        1  154169  154169  24.232 0.0003522 ***
## Residuals 12   76346    6362
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(reduced.model3)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.64  -35.53  -22.46   37.75  190.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.593     37.349   0.525 0.609414
## Protein       10.646      1.191   8.939 1.19e-06 ***
## Fat            9.354      1.900   4.923 0.000352 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.76 on 12 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9387
## F-statistic: 108.3 on 2 and 12 DF,  p-value: 2.097e-08
```

(1)Level of significance: α =0.05

(2)Hypothesis: H_0: β_1=β_2=0, vs. H_A: At least one $\beta_i \neq 0$ (i=1,2)

(3)Decision Rule: Reject H_0 if p-value $\leq \alpha$

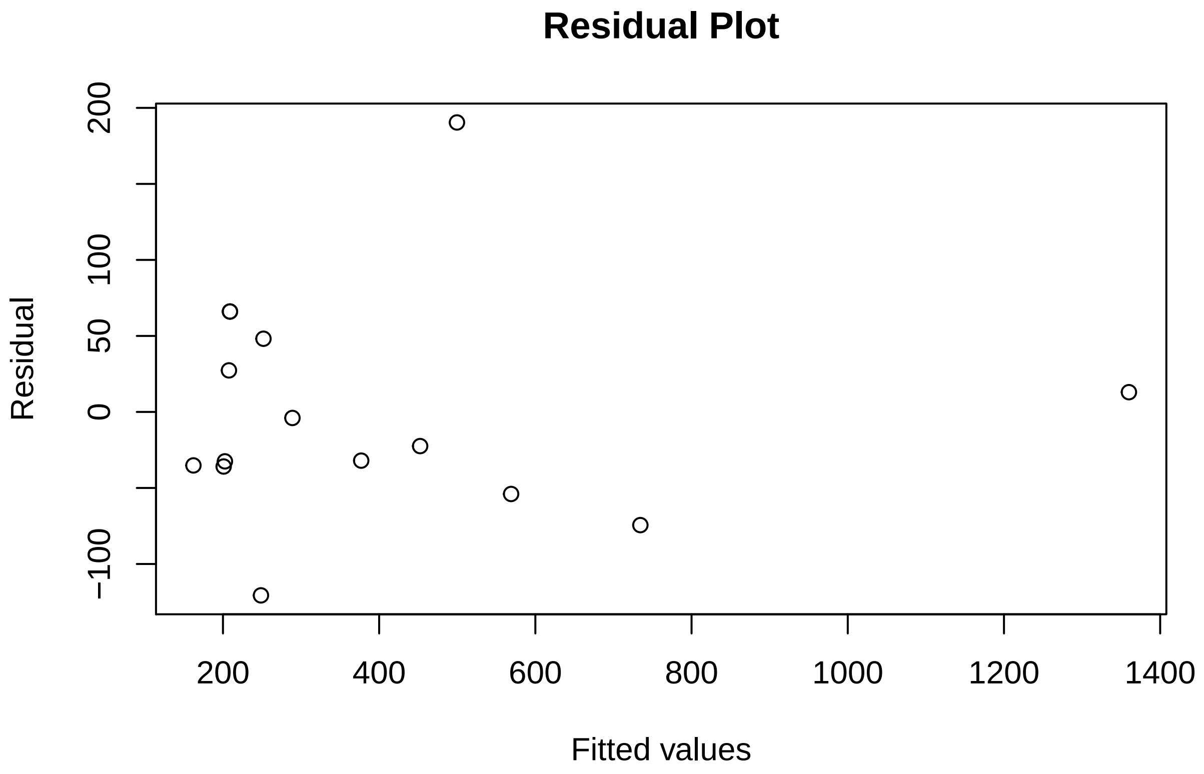
(4)Test statistic: F=108.3

(5)P-value: p-value $\approx$ 2.097e-08

(6)Conclusion: As p-value $\approx$ 2.097e-08 $<$ 0.5= α, reject H_0. Conclude that there is sufficient evidence that at least one of our model terms significantly explains the variation in Calories.

Residual Plot

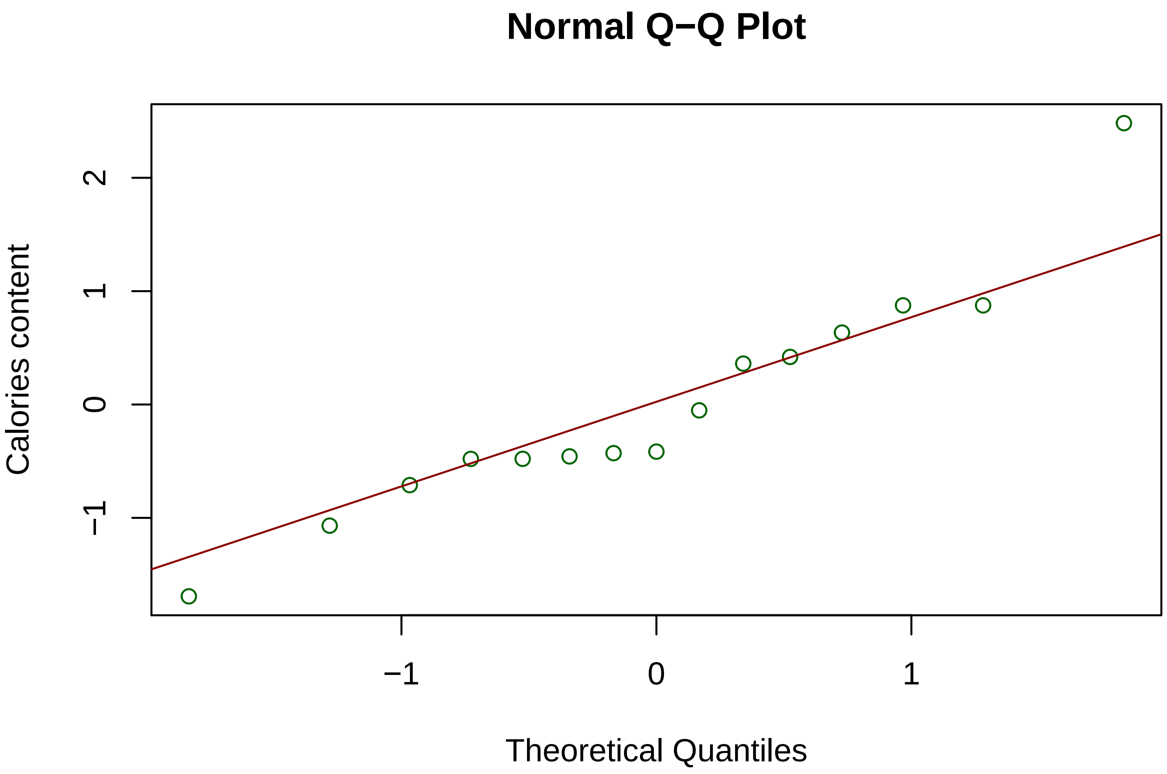
```
model.resid<-residuals(reduced.model3)
model.fitted<-fitted.values(reduced.model3)
plot(y=model.resid, x=model.fitted, xlab = "Fitted values", ylab = "Residual",
     main="Residual Plot")
```



The points in this plot look randomly distributed, but there is two outliers in the upper middle and upper right.There is a potential that there may be have extreme deviations from the assumptions if there are more points.

Now check the reduced.model3 assumption by using normal quantile plot.

```
model.stdres<-rstandard(reduced.model3)
qqnorm(model.stdres, ylab = "Calories content", col="dark green")
qqline(model.stdres, col="dark red")
```



Over half of the points are close or cross to the line. There are one outliers is away from the rest points. Overall, it requires us to use the model with caution.

Conclusion

Based on the ANOVA test and the reduced model we made, it is concluded that the explanatory variables Protein and Fat are able to predict the response variable Calories.

```
summary(reduced.model3)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Fat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.64  -35.53  -22.46   37.75  190.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.593     37.349   0.525 0.609414
## Protein       10.646      1.191   8.939 1.19e-06 ***
## Fat            9.354      1.900   4.923 0.000352 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 79.76 on 12 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9387
## F-statistic: 108.3 on 2 and 12 DF,  p-value: 2.097e-08
```

The final regression equation as the best estimate of the relationship between Y and X_1 and X_2 is: $\hat{y}$=19.593+10.646X_1+9.354X_2