Marcus Crowder
Statistics-3155
R-Homework Chapter 18

1. Regress Score on Calories, Type and Fat. Write down the fitted model. What is the interpretation of the coefficient of Type in this regression? Is that coefficient statistically significant? Explain.

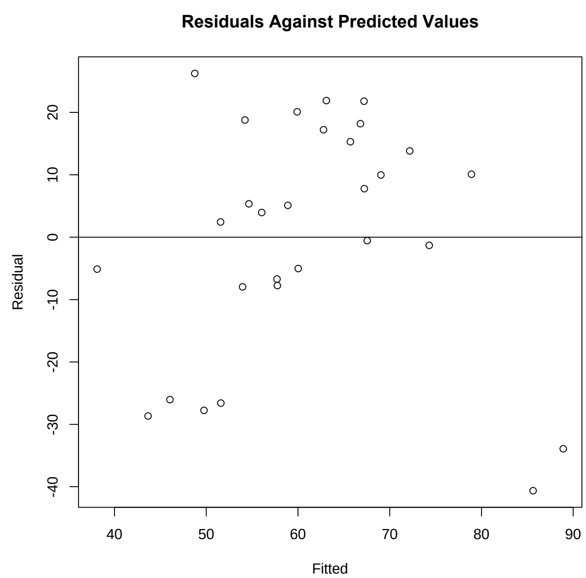
Fitted Model
Score = -148.8173 + 0.7430*Calories + 15.6344*Type – 3.8914*Fat

Interpretation of Type:
Given other predictors as constant, the mean Score of Pizza with cheese is greater than those with peperoni by 15.6344. The coefficient however is not Statistically significant in this model it would fail(be insignificant) under a 5% significance level because it is too high.

```
> imod1 <- lm(Score ~ Calories + Type + Fat, data = pizza)
> summary(imod1)

Call:
lm(formula = Score ~ Calories + Type + Fat, data = pizza)

Residuals:
   Min     1Q Median     3Q    Max
-40.63  -7.75   3.95  15.29  26.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -148.8173    77.9854  -1.908   0.0679 .
Calories       0.7430     0.3066   2.424   0.0229 *
Type          15.6344     8.1033   1.929   0.0651 .
Fat           -3.8914     2.1381  -1.820   0.0807 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 25 degrees of freedom
Multiple R-squared:  0.2873,	Adjusted R-squared:  0.2018
F-statistic:  3.36 on 3 and 25 DF,  p-value: 0.03464
```

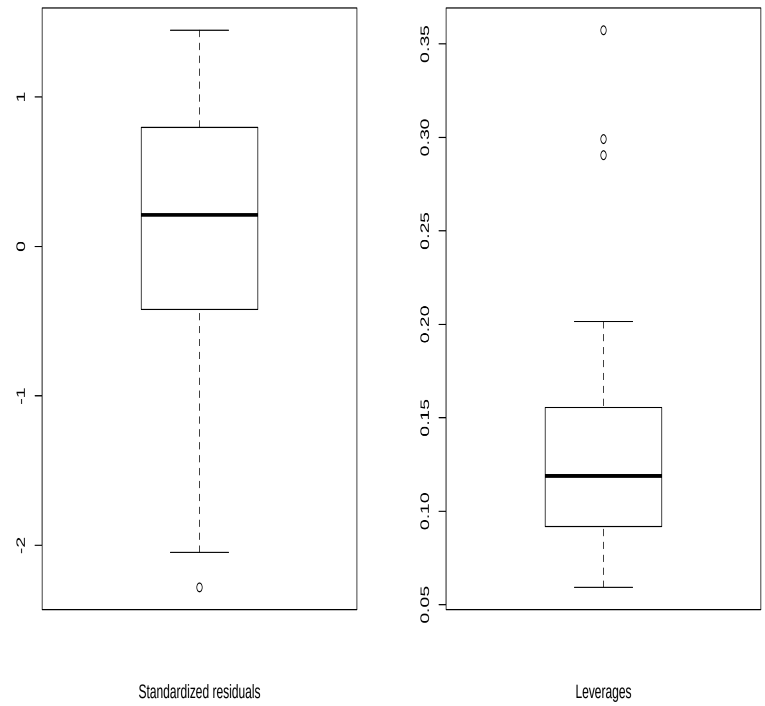
2. Based on the model obtained above plot its residuals against predicted values. Do you see any unusual observations in the plot? If yes, please identify them and find their standardized residuals, leverages, Cook's distances and DFFITS measures. Are the unusual points reflected in the residual plot influential? Is there any other influential case you can find? If yes, please identify the corresponding pizza.

Code:
> plot(imod1$fitted.values, imod1$residuals, xlab="Fitted", ylab="Residual", main = "Residuals Against Predicted Values")
> abline(0,0)



Their appear to be unusual observations in the lower right corner of this plot

Code:
> std.res <- rstandard(imod1)
> lev <- hatvalues(imod1)
> par(mfrow = c(1,2))
> boxplot(std.res, xlab='Standardized residuals')
> boxplot(lev, xlab = 'Leverages')



There seems to be only one point outside of the standardized residuals but up to 3 points with unusual leverages.
The standardized residual point appear to be beyond -2.05 (I initially thought it was beyond -2 but two points appeared in the in the R file so I switched it to 2.05 and one appeared so I left it out because the lines of the picture are slightly beyond -2 and only show 1 unusual standardized residual).

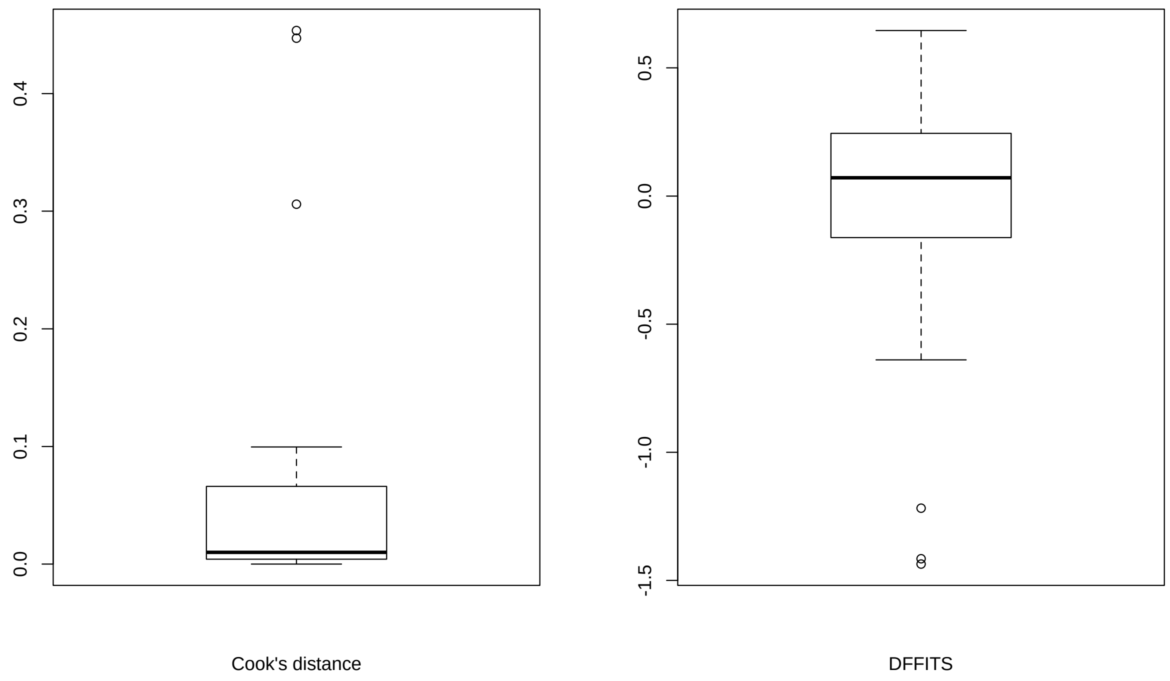
```
> pizza[which(abs(std.res) > 2.05), c(1,2,4,5,6)]
       Brand Score Calories Fat Type
16 Michelina    45      394  19    1
```

The pizza company that appeared is Michelina

On to the high leverage points. The common rule is to flag any case whose leverage is more than 3 times larger than the mean leverages but there seems to be no case that is larger than that which goes with our box plot graph. The leverages appear to be close to each other on the boxplot

```
> lev.cut <- 3*length(imod1$coefficients)/dim(pizza)[1]
> pizza[which(lev > lev.cut), c(1, 2, 4, 5, 6)]
[1] Brand    Score    Calories Fat      Type
<0 rows> (or 0-length row.names)
```

This also goes to show that the unusual residual point Michelina does not appear to be flagged here. Further investigation is needed to find out what makes it special.



Code:

\> cookd <- cooks.distance(imod1)
\> dffit <- dffits(imod1)
\> par(mfrow = c(1,2))
\> boxplot(cookd, xlab = "Cook's distance")
\> boxplot(dffit, xlab= 'DFFITS')

There seems to be only 3 points worth mentioning but with 2 points closer to each other than the 3rd which is lower so maybe remove both points.
But when using dffit

```
> lev.cut <- 3*length(imod1$coefficients)/dim(pizza)[1]
> pizza[which(lev > lev.cut), c(1, 2, 4, 5, 6)]
[1] Brand     Score     Calories Fat      Type
<0 rows> (or 0-length row.names)
> cookd <- cooks.distance(imod1)
> dffit <- dffits(imod1)
> par(mfrow = c(1,2))
> boxplot(cookd, xlab = "Cook's distance")
> boxplot(dffit, xlab= 'DFFITS')
> pizza[which.max(cookd), c(1,2,4,5,6)]
                         Brand Score Calories Fat Type
29 Healthy_Choice_pepperoni    15      280   4    0
```

The maximum cook point appears to be Healthy choice peperoni which did not show up in our standardized residuals.

```
> dffit.cut <- 2*sqrt(length(imod1$coefficients)/(dim(pizza)[1] - length(imod1$coefficients)))
> which(abs(diffit) > dffit.cut)
Error in which(abs(diffit) > dffit.cut) : object 'diffit' not found
> which(abs(dffit) > dffit.cut)
12 16 29
12 16 29
```

However when using the cutoff for DFFITS we see 29 again for healthy choice peperoni. For the final test we needto check if the point appears again for the max of dffit

```
> which(abs(dffit) > dffit.cut)
12 16 29
12 16 29
> pizza[which.max(dffit), c(1,2,4,5,6)]
   Brand Score Calories Fat Type
7 Kroger    75      292   9    1
```

However the point does the point does not appear again and we should not consider removing that point or any other point.

```
> dffit
          1           2           3           4           5           6           7           8           9          10          11          12 
 0.197396793  0.244613462  0.249692955  0.198452819  0.335190507  0.142840257  0.645914469  0.127472625 -0.025391631 -0.007488184  0.071128244 -1.436218663 
         13          14          15          16          17          18          19          20          21          22          23          24 
-0.106305785 -0.162161363 -0.150463774 -1.218180016 -0.639566152  0.518863836  0.480909415  0.336039544  0.316515216  0.100298596  0.149244888 -0.116788614 
         25          26          27          28          29 
 0.043643895 -0.192509345 -0.531328106 -0.572005439 -1.415416294 
```

the cooks distance for these points are negative but are influential if used with absolute value(points: 12,16,29).

With 29 being Healthy_Choice_Peperoni and the others being -

```
> pizza[12,]
   Brand Score Cost Calories Fat Type
12 Reggio    55 1.02      367  13    1
> pizza[16,]
      Brand Score Cost Calories Fat Type
16 Michelina    45 1.28      394  19    1
```

Michelina also appears in the standardized residual plot as an unusual point as well.

3. Let's remove all the influential cases from the dataset and refit the multiple regression model in Problem 1. Compare the new model to the old one based on their summaries. Check the assumptions for the new model.

To remove the outliers I used a neat R trick to concatenate all the influential points 12,16,29 Or Reggio, Michelina and Healthy_Choice_Peperoni.

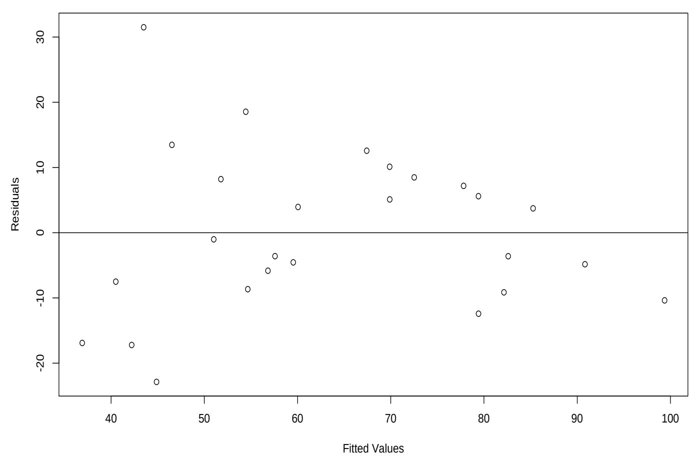
```
> pizza2 = pizza[-c(12,16,29),]
> pizza2
                              Brand Score Cost Calories Fat Type
1                   Freshetta4Cheese    89 0.98      364  15    1
2          Freschetta_stuffed_crust    86 1.23      334  11    1
3                           DiGiorno    85 0.94      332  12    1
4                        Amy_organic    81 1.92      341  14    1
5                            Safeway    80 0.84      307   9    1
6                               Tony    79 0.96      335  12    1
7                             Kroger    75 0.80      292   9    1
8           Tombstone_stuffed_crust    75 0.96      364  18    1
9                          Red_Baron    73 0.91      384  20    1
10                             Bobli    67 0.89      333  12    1
11          Tombstone_extra_cheese    60 0.94      328  14    1
13                              Jack    51 0.92      325  13    1
14                           Celeste    50 1.17      346  17    1
15                      McCain_Ellio    46 0.54      299   9    1
17                            Totino    25 0.67      322  14    1
18              Freschetta_pepperoni    89 0.96      385  18    0
19                DiGiorno_pepperoni    85 0.88      369  16    0
20 Tombstone_stuffed_crust_pepperoni    80 0.90      400  22    0
21               Tombstone_pepperoni    73 0.88      378  20    0
22               Red_Baron_pepperoni    64 0.89      400  23    0
23                    Tony_pepperoni    60 0.87      410  26    0
24      Red_Baron_deep_dish_pepperoni    55 1.28      412  25    0
25                Stouffer_pepperoni    54 1.26      343  14    0
26          Weight_Watchers_pepperoni    33 1.51      283   6    0
27                    Jeno_pepperoni    22 0.74      372  20    0
28                  Totino_pepperoni    20 0.64      367  20    0
```

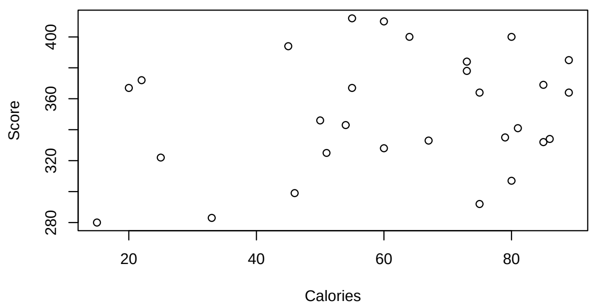
The result of removing those outliers produced something quite incredible. In comparison to the original multiple regression this one produced more favorable results. Type became Statistically significant under a 5% significance level and the P values of Calories and fat are significantly lowered making the model more useful for calculations. In addition the intercept changed from -148.8713 to -351.9436. Each of the coefficient estimates also changed along with lower std.Error values and higher test Stat. Lastly the F-Statistic went up! It's a Brave New World!

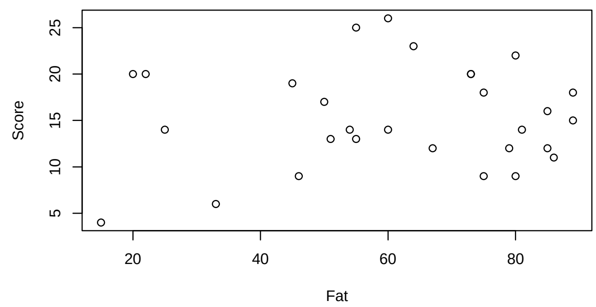
```
> imod2 <- lm(Score ~ Calories + Type + Fat, data = pizza2)
> summary(imod1)

Call:
lm(formula = Score ~ Calories + Type + Fat, data = pizza)

Residuals:
   Min     1Q Median     3Q    Max 
-40.63  -7.75   3.95  15.29  26.24 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -148.8173    77.9854  -1.908   0.0679 .
Calories       0.7430     0.3066   2.424   0.0229 *
Type          15.6344     8.1033   1.929   0.0651 .
Fat           -3.8914     2.1381  -1.820   0.0807 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 25 degrees of freedom
Multiple R-squared:  0.2873,	Adjusted R-squared:  0.2018 
F-statistic:  3.36 on 3 and 25 DF,  p-value: 0.03464

> summary(imod2)

Call:
lm(formula = Score ~ Calories + Type + Fat, data = pizza2)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.878  -8.372  -2.298   7.960  31.503 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -351.9436    65.4809  -5.375 2.14e-05 ***
Calories       1.5951     0.2559   6.234 2.84e-06 ***
Type          18.1209     6.3100   2.872  0.00886 ** 
Fat           -9.8278     1.7557  -5.598 1.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.04 on 22 degrees of freedom
Multiple R-squared:  0.6639,	Adjusted R-squared:  0.6181 
F-statistic: 14.48 on 3 and 22 DF,  p-value: 1.99e-05
```

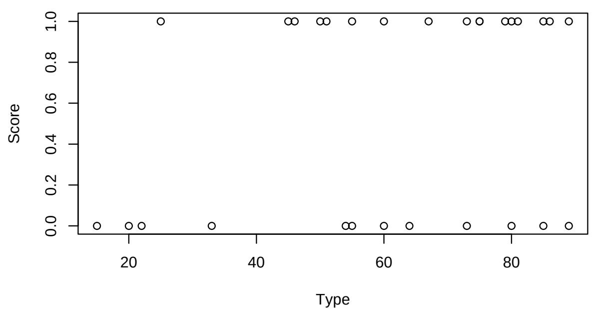
> plot(imod2$fitted.values, imod2$residuals, xlab= 'Fitted Values', ylab = 'Residuals')
> abline(0,0)
The equal variance assumption appears to be somewhat met with equal spread of residuals

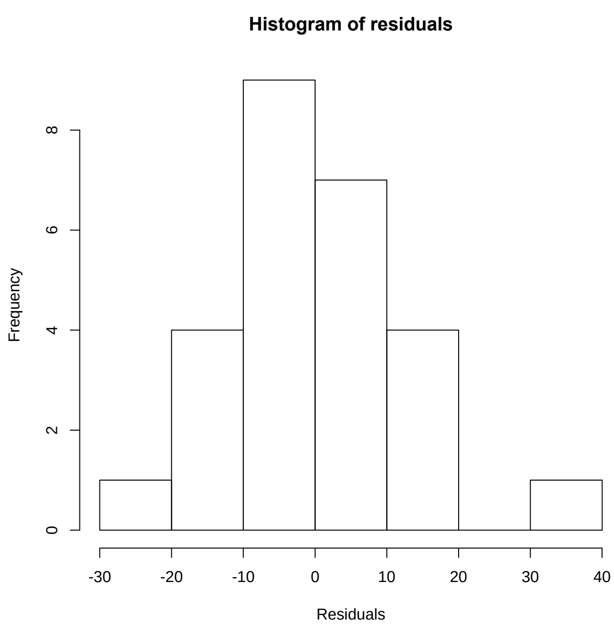
There also appear to be no patterns in the imod model so the linearity condition can be said to be met for the multiple regression model. When plotting score against each individual value their also appears to be no obvious patterns so linearity condition is met.

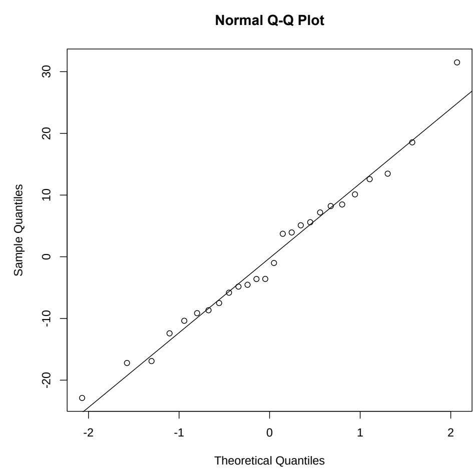












The model of the historgram appear to be unimodel and the qqline fits the model so the normality assumption appears to be met and now all of the assumptions are met.

Code:

```
> plot(imod1$fitted.values, imod1$residuals, xlab= 'Fitted Values', ylab = 'Residuals')
> abline(0,0)
> plot(imod2$fitted.values, imod1$residuals, xlab= 'Fitted Values', ylab = 'Residuals')
Error in xy.coords(x, y, xlabel, ylabel, log) :
```

```
  'x' and 'y' lengths differ
> plot(imod2$fitted.values, imod2$residuals, xlab= 'Fitted Values', ylab = 'Residuals')
> abline(0,0)
> plot(pizza$Score, pizza$Calories, xlab= 'Fitted Values', ylab = 'Residuals')
> par(mfrow=c(2,2))
> plot(pizza$Score, pizza$Calories, xlab= 'Calories', ylab = 'Score')
> plot(pizza$Score, pizza$Fat, xlab= 'Fat', ylab = 'Score')
> plot(pizza$Score, pizza$Type, xlab= 'Type', ylab = 'Score')
> hist(imod2$residuals, main = "Histogram of residuals", xlab = 'Residuals')
> qqnorm(imod2$residuals)
> qqline(imod2$residuals)
```

4. Check collinearity for the model fitted in Problem 3. Does there exist any serious collinearity? If does, could you find the reason?

```
The downloaded binary packages are in
    /var/folders/39/lg3ckdv50qs5009m0wl1tnw80000gn/T//RtmpNOmRYC/downloaded_packages
> library(car)
Warning message:
package 'car' was built under R version 3.4.3
> vif(imod2)
Calories     Type      Fat
12.52500  1.48502 12.37517
> vif(imod2.tmp)
Error in vif(imod2.tmp) : object 'imod2.tmp' not found
> cor(pizza2$Calories, pizza2$Fat)
[1] 0.9585401
```

There seems to be high collinearity between Calories and Fat based on the High Vif which is over 10 and this does not occur with Type. We further look at the correlation between the two and found it to be .9585401 which is high. Being someone deeply involved with nutrition I can see why these two are related . 1 gram of fat contains 9 calories which as opposed to carbs and protein (which contain 4 calories per gram) are very high so it is understandable why with a higher fat content we could track the increase in calories.

5. We now use the full dataset pizza. Do we need to consider the interaction between Calories and Type? Explain. Add an interaction term to the model if you think it is necessary, and fit the new model. Interpret the resulting coefficient of interaction term.

```
> imod4 <- lm(Score ~ Calories + Type + Fat + Cost, data = pizza)
> summary(imod4)

Call:
lm(formula = Score ~ Calories + Type + Fat + Cost, data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.795  -6.791   4.066  16.876  25.684 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -149.5069    79.5528  -1.879   0.0724 .
Calories       0.7635     0.3241   2.356   0.0270 *
Type          15.2647     8.4057   1.816   0.0819 .
Fat           -4.0808     2.3206  -1.759   0.0914 .
Cost          -3.3028    13.8879  -0.238   0.8140  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.17 on 24 degrees of freedom
Multiple R-squared:  0.289,	Adjusted R-squared:  0.1705 
F-statistic: 2.439 on 4 and 24 DF,  p-value: 0.07458
```

```
> imod4 <- lm(Score ~ Calories + Type + Fat + Cost + Calories*Type, data = pizza)
> summary(imod4)

Call:
lm(formula = Score ~ Calories + Type + Fat + Cost + Calories * 
    Type, data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.518 -14.485   2.827  11.791  22.799 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -361.3576    93.6488  -3.859 0.000799 ***
Calories         1.4056     0.3378   4.161 0.000377 ***
Type           288.0736    84.2085   3.421 0.002337 ** 
Fat             -6.5009     2.0985  -3.098 0.005074 ** 
Cost            15.6173    13.1054   1.192 0.245543    
Calories:Type   -0.7806     0.2401  -3.251 0.003519 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.06 on 23 degrees of freedom
Multiple R-squared:  0.5129,	Adjusted R-squared:  0.407 
F-statistic: 4.843 on 5 and 23 DF,  p-value: 0.003595
```

When we use the full pizza data set with cost and Add the interaction term between Calories and Type we are able to see an increased multiple R-Squared and an Increased F-Statistics making it even more significant on 5 variables. The P-value of the multiple regression model also became significant to use under a regression model. However Cost has no impact on this model. I think Calories and Type should be considered as an interaction term.

Interpretation:
For a pizza of type cheese the Score of the pizza is expected to decrease by -0.7806 more than the pizzas with peperoni/no cheese pizzas.

```
> imod4 <- lm(Score ~ Calories + Type + Fat + Calories*Fat, data = pizza)
> summary(imod4)

Call:
lm(formula = Score ~ Calories + Type + Fat + Calories * Fat,
    data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max
-38.837  -7.474   2.900  14.514  27.724

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.857e+02  1.130e+02  -1.643    0.113
Calories      8.523e-01  3.926e-01   2.171    0.040 *
Type          1.317e+01  9.840e+00   1.338    0.193
Fat          -5.155e-01  7.690e+00  -0.067    0.947
Calories:Fat -9.309e-03  2.034e-02  -0.458    0.651
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.11 on 24 degrees of freedom
Multiple R-squared:  0.2935,  Adjusted R-squared:  0.1757
F-statistic: 2.492 on 4 and 24 DF,  p-value: 0.06995

> imod4 <- lm(Score ~ Calories + Type + Fat + Fat*Type, data = pizza)
> summary(imod4)

Call:
lm(formula = Score ~ Calories + Type + Fat + Fat * Type, data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max
-38.755 -11.076   5.601  13.891  21.957

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -190.8766    76.0753  -2.509  0.01926 *
Calories       0.8495     0.2928   2.902  0.00783 **
Type          64.9881    25.0223   2.597  0.01580 *
Fat           -3.7218     2.0116  -1.850  0.07664 .
Type:Fat      -3.3648     1.6250  -2.071  0.04931 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.6 on 24 degrees of freedom
Multiple R-squared:  0.3953,  Adjusted R-squared:  0.2946
F-statistic: 3.923 on 4 and 24 DF,  p-value: 0.01372
```

After playing around with a few interaction terms none of them stood out I believe the best would be between Calories and type.