Julie Rachford
Student Project
Regression Analysis
Winter 2012

Baskin Robbins Ice Cream Caloric Values

Introduction

One of my favorite desserts is ice cream. Therefore, I choose to perform a regression analysis that allows me to investigate what ice cream attributes are indicative of the calorie content of these tasty desserts. In my analysis, I employed the following seven (7) explanatory variables: sugar (g), saturated fat (g), total fat (g), cholesterol (g), sodium (mg), carbohydrates (g) and protein (g).

Data

I relied upon the data from the Baskin Robbins website to perform my analysis. The data can be found at the following website:

http://www.baskinrobbins.com/Nutrition/productlist.aspx?category=Ice%20Cream

I choose to utilize data from the classic ice cream flavors for the 2.5 oz serving size. The data I compiled can be found on the attached excel spreadsheet.

Equation and Variables

The equation and the 7 variables for the full model are:

$$Y = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7$$

Where: Y = calories
α = intercept
β_i = least squares coefficients
X_1 = sugar (grams)
X_2 = saturated fat (grams)
X_3 = total fat (grams)
X_4 = cholesterol (milligrams)
X_5 = sodium (milligrams)
X_6 = carbohydrates (grams)
X_7 = protein (grams)

Hypothesis

The null hypothesis is that all least squares coefficients are zero: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Data Analysis

I utilized the Regression data analysis tool in Excel to obtain the following regression statistics and ANOVA tables:

7 Variable Regression Full Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.971049188
R Square	0.942936526
Adjusted R Square	0.909649499
Standard Error	4.498194855
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	7	4012.194917	573.1707024	28.32744822
Residual	12	242.8050834	20.23375695	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	38.12856667	24.9773286	1.526527007	0.152795844
Sugar(g)	-0.820917233	1.683613484	-0.487592456	0.634624958
Saturated Fat (g)	3.487268326	2.053836987	1.697928486	0.11527863
Total Fat (g)	8.304910949	1.565872675	5.30369492	0.000187209
Cholesterol (mg)	-0.076815849	0.360073112	-0.213334034	0.834647878
Sodium (mg)	0.149832585	0.085750836	1.747301744	0.106101727
Carbohydrates (g)	2.750345948	0.962489623	2.857533091	0.01441972
Protein (g)	-3.578552457	3.925335288	-0.911655233	0.37990008

Model Equation:

$$Y = 38.1286 - 0.8209X_1 + 3.4873X_2 + 8.3049X_3 - 0.07682X_4 + 01498X_5 + 2.7503X_6 - 3.5756X_7$$

The R^2 of the full model is 94.2937% demonstrating that the full model a good indicator of determining calories based on the provided nutritional facts. However, given that cholesterol has the highest p-value and one of the lowest t stat values, it appears that cholesterol may not be a great explanatory variable for this model. Due to this, I will remove cholesterol from the model, and rely upon the six (6) remaining variables.

6 Variable Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.970937746
R Square	0.942720106
Adjusted R Square	0.916283231
Standard Error	4.329913396
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	6	4011.27405	668.545675	35.65928768
Residual	13	243.7259503	18.74815002	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	35.20642953	20.10512437	1.751117221	0.103469146
Sugar(g)	-0.818782656	1.62059939	-0.505234459	0.621850721
Saturated Fat (g)	3.309413213	1.806798528	1.831644847	0.090008616
Total Fat (g)	8.250811448	1.48739457	5.547157167	9.42993E-05
Sodium (mg)	0.147644241	0.081950106	1.801635742	0.094829616
Carbohydrates (g)	2.808530865	0.888506082	3.160958513	0.007511479
Protein (g)	-3.250220342	3.475943392	-0.935061356	0.366801682

Model Equation:

$$Y = 35.2064 - 0.8188X_1 + 3.3094X_2 + 8.2508X_3 + 0.1476X_5 + 2.8085X_6 - 3.2502X_7$$

The R^2 of this model is 94.2720%, which is near our initial model R^2 of 94.2937% demonstrating that this model is also a good indicator of determining calories based on the provided nutritional facts. However, this model produced a slightly lower standard error (4.3299 vs. 4.4982) and an increased F-statistic (35.6593 vs. 28.3274) reflecting the fact that the six (6) variable model is a better fit than the full model. Again, given that sugar has the highest p-value and one of the lowest t stat values, it appears that the model may produce a better result if this explanatory variable is eliminated. Therefore, I will remove sugar from the model, and rely upon the five (5) remaining variables.

5 Variable Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97035838
R Square	0.941595385
Adjusted R Square	0.920736594
Standard Error	4.21317354
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	5	4006.488362	801.2976724	45.14141676
Residual	14	248.5116379	17.75083128	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	27.53170464	12.81559238	2.148297467	0.04967426
Saturated Fat (g)	2.901799362	1.573083324	1.844657125	0.086347176
Total Fat (g)	8.512510988	1.356697655	6.274434806	2.04113E-05
Sodium (mg)	0.127463727	0.069627618	1.830648962	0.088518551
Carbohydrates (g)	2.487979298	0.605244405	4.110701853	0.001059756
Protein (g)	-2.900968918	3.314669583	-0.875190979	0.396237027

Model Equation:

$$Y = 27.6317 + 2.902X_2 + 8.5125X_3 + 0.1274X_5 + 2.4880X_6 - 2.9010X_7$$

The R^2 of this model is 94.1595%, which is slightly less than the six variable model R^2 of 94.2720% demonstrating that this model is also a good indicator of determining calories based on the provided nutritional facts. However, this model produced an even lower standard error (4.2132 vs. 4.3299) and larger F-statistic (45.1414 vs. 35.6593) reflecting the fact that the five (5) variable model is a better fit than both the six (6) variable model and the full model. Again, given that protein has the highest p-value and the lowest t stat value of the remaining variables, it appears that the model may produce an even better result if this explanatory variable is eliminated. Therefore, I will remove protein from the model, and rely upon the four (4) remaining variables.

4 Variable Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.968710477
R Square	0.938399988
Adjusted R Square	0.921973318
Standard Error	4.180175844
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	4	3992.891949	998.2229872	57.12661147
Residual	15	262.1080513	17.47387009	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	27.03922832	12.70295745	2.128577414	0.050272135
Saturated Fat (g)	2.996427188	1.557071845	1.924398799	0.07348764
Total Fat (g)	7.580010992	0.833289515	9.096491502	1.71099E-07
Sodium (mg)	0.104163478	0.063832873	1.631815611	0.123529861
Carbohydrates (g)	2.562609691	0.594514185	4.310426489	0.000618694

Model Equation:

$$Y = 27.0392 + 2.9964X_2 + 7.5800X_3 + 0.1041X_5 + 2.5626X_6$$

The R^2 of this model is 93.8400%, which is slightly less than the five variable model R^2 of 94.1595% demonstrating that this model is also a good indicator of determining calories based on the provided nutritional facts. However, this model produced an even lower standard error (4.1802 vs. 4.2132) and larger F-statistic (57.1266 vs. 45.1414) reflecting the fact that the four (4) variable model is a better fit than the three previous run models. Again, given that sodium has the highest p-value and the lowest t stat value of the remaining variables, it appears that the model may produce an even better result if this explanatory variable is eliminated. Therefore, I will remove sodium from the model, and rely upon the three (3) remaining variables.

3 Variable Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963049668
R Square	0.927464663
Adjusted R Square	0.913864287
Standard Error	4.392023026
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	3946.36214	1315.454047	68.19404699
Residual	16	308.6378602	19.28986627	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	18.13518164	12.05268678	1.504658834	0.151895566
Saturated Fat (g)	2.612114339	1.617160762	1.615247167	0.125801706
Total Fat (g)	8.160520862	0.791724858	10.30726871	1.79991E-08
Carbohydrates (g)	3.234497321	0.450587385	7.178401855	2.19587E-06

Model Equation:

$$Y = 18.1352 + 2.6121X_2 + 8.1605X_3 + 3.2345X_6$$

The R^2 of this model is 92.7465%, which is slightly less than the four variable model R^2 of 93.8400% demonstrating that this model is also a good indicator of determining calories based on the provided nutritional facts. Although, this model produced a faintly higher standard error (4.3920 vs. 4.1802), the F-statistic increased (57.1266 vs. 45.1414) reflecting the fact that the three variable model is a better fit than the four previous run models. Again, given that saturated fat has the highest p-value and the lowest t stat value of the remaining variables, it appears that the model may produce an even better result if this explanatory variable is eliminated. Therefore, I will remove sodium from the model, and rely upon the two (2) remaining variables.

2 Variable Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.956889107
R Square	0.915636763
Adjusted R Square	0.905711676
Standard Error	4.595173766
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	3896.034427	1948.017214	92.25478744
Residual	17	358.9655729	21.11562194	
Total	19	4255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	23.20007845	12.17593794	1.905403803	0.073784063
Total Fat (g)	8.838894273	0.702195048	12.58752009	4.82664E-10
Carbohydrates (g)	3.36263787	0.464064814	7.246052208	1.36699E-06

Model Equation:

$$Y = 23.2001 + 8.8389X_3 + 3.3626X_6$$

The R^2 of this model is 91.5637%, which is slightly less than the three variable model R^2 of 92.7465% demonstrating that this model is also a good indicator of determining calories based on the provided nutritional facts. Although, this model produced a faintly higher standard error (4.5952 vs. 4.3920), the F-statistic increased significantly (92.2548 vs. 57.1266) reflecting the fact that the two variable model is a better fit than the five previous run models. The p-values for the remaining explanatory variables are very close to zero, allowing us to reject the null hypothesis.

Conclusion

Beginning my regression analysis with sugar, saturated fat, total fat, cholesterol, sodium, carbohydrates and protein, I eliminated explanatory variables one by one that did not appear to be a good fit with the model. Based on the results of the six regression analysis performed, I determined that the equation with two explanatory variables in which fat is the biggest contributor to calories provides the best fit.

$$Y = 23.2001 + 8.8389X_3 + 3.3626X_6$$

Where: X_3 = total fat (grams)
X_6 = carbohydrates (grams)

The R^2 values for this equation is high (91.5637%), the F- statistic was the highest of all the models, and the p-values of the remaining variables were very close to zero.